UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

DIVISIÓN DE INGENIERÍA MECÁNICA E INDUSTRIAL

Serie de Calidad y Estadística Industrial

Fundamentos de Estadística y Aplicaciones, con R, Minitab y Excel

OCTAVIO ESTRADA CASTILLO

Serie de Calidad y Estadística Industrial

Fundamentos de Estadística y Aplicaciones, con R, Minitab y Excel

Octavio Estrada Castillo

Leonardita
de Sebastián, 1992

ESTRADA CASTILLO, Octavio.
Fundamentos de Estadística
y Aplicaciones, con R, Minitab y Excel
Universidad Nacional Autónoma de México,
Facultad de Ingeniería, 2023, 471 p.

978-607-30-7596-1

Fundamentos de Estadística
y Aplicaciones, con R, Minitab y Excel

Primera edición electrónica
de un ejemplar (14 MB) en formato PDF
Publicado en línea: 8 de mayo de 2023

FACULTAD DE INGENIERÍA
http://www.ingenieria.unam.mx/

978-607-30-7596-1

Hecho en México.

Unidad de Apoyo Editorial
Cuidado de la edición: Patricia Eugenia García Naranjo
Diseño y formación editorial: Nismet Díaz Ferro

Fotografía de portada: Luis Enrique Vite Rangel

Pierre-Simon Laplace (1749-1827) afirmó: «Es notable que una ciencia que comenzó con consideraciones sobre juegos de azar haya llegado a ser el objeto más importante del conocimiento humano». Comprender y estudiar el azar es indispensable, porque la probabilidad es un soporte necesario para tomar decisiones en cualquier ámbito.

Napoleón, refiriéndose a su obra Exposition du système du monde, comentó a Laplace: «Me cuentan que ha escrito usted este gran libro sobre el sistema del universo sin haber mencionado ni una sola vez a su creador», y Laplace contestó: «Sieur, nunca he necesitado esa hipótesis».

Prólogo

Desde 1988 he estado inmerso en la práctica profesional de la Ingeniería, particularmente, en la industria manufacturera y en específico en el área de Calidad y Estadística Industrial. Paralelamente, he hecho una trayectoria académica como profesor en la Facultad de Ingeniería de la UNAM, desde hace más de 35 años. He impartido más de 25 asignaturas diferentes de matemáticas, física, computación, probabilidad, estadística, investigación de operaciones y calidad. Me ha tocado laborar en el sector público federal también. Desde hace más de 30 años que soy auditor certificado de sistemas de calidad, a través de ISO 9001 y me ha tocado evaluar a 134 empresas proveedoras del sector metal mecánico y eléctrico.

Siempre he tenido la inquietud de escribir textos sobre las asignaturas en las que he participado, pero generalmente he tenido cargos académico-administrativos en la administración central de la UNAM o en la propia Facultad, que no me dejaban dedicarme a esta noble y gratificante labor. ahora que me integro completamente como profesor de carrera en el área de calidad, investigación de operaciones y estadística industrial, he llevado a la praxis este deseo. Este es el cuarto volumen de la Serie de Calidad y Estadística Industrial. Esta serie tendrá al menos los siguientes volúmenes:

I. Desarrollo Histórico de la Calidad.
II. Metodología y Herramientas para la Solución de Problemas y para la Mejora Continua.
III. Fundamentos de probabilidad y aplicaciones con R, Minitab y Excel.
IV. Fundamentos de estadística y aplicaciones con R, Minitab y Excel.
V. Muestreo de aceptación y aplicaciones con R, Minitab y Excel.
VI. Control estadístico de procesos y aplicaciones con R, Minitab y Excel.
VII. Normatividad Vigente sobre Sistemas de Calidad.
VIII. Metrología, Certificación de Producto y Certificación de Software.
IX. Teoría del muestreo.
X. Estadística no paramétrica.

XI. Diseño de experimentos.
XII. Regresión y correlación.
XIII. Confiabilidad.
XIV. Estadística multivariable.
XV. Procesos Estocásticos.

El propósito de estos libros es proporcionar la teoría necesaria, la metodología, las herramientas, ejemplos y aplicaciones prácticas de cada uno de los temas, de una manera formal, dinámica, amena y didáctica. Quisiera remarcar que en estos libros hablo de mis conocimientos y experiencia en el apasionante tema de la calidad y que traté de apegarme lo más posible a citar a los autores originales de estas ideas, pero no debe olvidarse que se trata de un texto dirigido a alumnos por lo cual no lleno de citas el texto, para hacerlo más didáctico.

OBJETIVO DE ESTE LIBRO:

El alumno aplicará los conceptos, la metodología, las herramientas y técnicas de la estadística para interpretar algunos fenómenos aleatorios que ocurren en la naturaleza, la sociedad y la industria, así como modelar y resolver problemas sujetos a incertidumbre.

Índice temático

1. Introducción a la Estadística y al tratamiento de datos

1.1. Esbozo histórico de la Estadística

La palabra Estadística procede del vocablo "Estado", pues era función principal de los gobiernos de los estados establecer registros de población, nacimientos, defunciones, impuestos, cosechas, etcétera. La necesidad de poseer datos cifrados sobre la población y sus condiciones materiales de existencia han surgido desde que se establecieron sociedades humanas organizadas. La palabra alemana statistik, introducida primeramente en 1749 por Gottfried Achenwall (1719-1772), originalmente designaba al análisis de datos acerca del estado. El término fue introducido en Inglaterra en 1792 por sir John Sinclair (1754-1835) cuando publicó el primero de los 21 volúmenes titulados *Statistical Account of Scotland.*

Hacia el año 3000 aC los babilonios usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y los géneros vendidos o cambiados mediante el trueque.

Los egipcios ya analizaban los datos de la población y la renta del país a partir del año 3050 aC, mucho antes de construir las pirámides. En los antiguos monumentos egipcios se encontraron interesantes documentos que demuestran la sabia organización y administración de este pueblo; ellos llevaban cuenta de los movimientos poblacionales y continuamente hacían censos, hasta tenían a la diosa Seshat o Safnkit, diosa de los libros y las cuentas.

Figura 1.1. Diosa Seshat
Jeft Dahl, 2008. Recuperado de:
https://es.wikipedia.org/wiki/Seshat

Números, el cuarto libro del Pentateuco del Antiguo Testamento de la Biblia, deriva su nombre de las listas del censo que realizó Moisés después de la salida de Egipto.

En China existían los censos chinos ordenados por el emperador Tao hacia el año 2200 aC.

Hacia el año 500 aC, se realizaron censos en Roma para conocer la población existente en aquel momento. Se erigió la figura del censor, cuya misión consistía en controlar el número de habitantes y su distribución por los distintos territorios.

En la Edad Media, en el año 762, Carlomagno ordenó la creación de un registro de todas sus propiedades, así como de los bienes de la iglesia.

El primer escrito de estadística fue encontrado en un libro del siglo IX titulado Manuscrito sobre el descifrado de mensajes criptográficos, escrito por Al-Kindi (801-873). En su libro, Al-Kindi da una descripción detallada sobre el uso de las estadísticas y análisis de frecuencias en el descifrado de mensajes, este fue el nacimiento tanto de la estadística como del criptoanálisis.

Después de la conquista normanda de Inglaterra en 1066, el rey Guillermo I, el Conquistador, elaboró un catastro que puede considerarse el primero de Europa.

Los Reyes Católicos ordenaron a Alonso de Quintanilla en 1482 el recuento de hogares de las provincias de Castilla.

En 1662 un mercader de lencería londinense, John Graunt (1620-1674) a quien se considera el primer demógrafo, el fundador de la bioestadística y el precursor de la epidemiología, publicó un tratado con las observaciones políticas y naturales, donde pone de manifiesto las cifras brutas de nacimientos y defunciones ocurridas en Londres durante el periodo 1604-1661, así como las influencias que ejercían las causas naturales, sociales y políticas de dichos acontecimientos. Puede considerarse el primer trabajo estadístico serio sobre la población.

Figura 1.2. Capitán John Graunt (1623-1687)

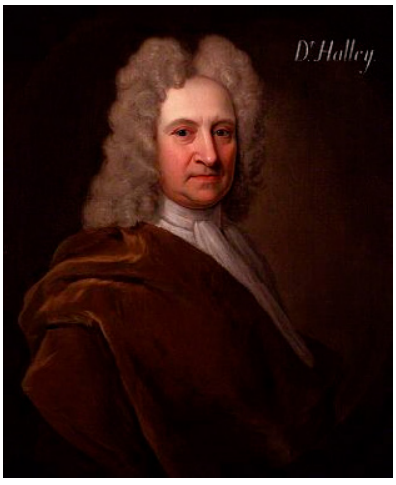
Recuperado de T O´Donell, 1936: https://es.wikipedia.org/wiki/John_Graunt

Curiosamente, Graunt no conocía los trabajos de Blaise Pascal (1623-1662) ni de Christiaan Huygens (1629-1695) sobre estos mismos temas. Un poco más tarde, el astrónomo Edmund Halley (1656-1742) presenta la primera tabla de mortalidad que se puede considerar como base de los estudios contemporáneos. En dicho trabajo se intenta establecer el precio de las anualidades que se pagaban a las compañías de seguros. Es decir, en Londres y en París se estaban construyendo, casi de manera simultánea, las dos disciplinas que actualmente se llaman estadística y probabilidad.

Figura 1.3. Blaise Pascal (1623-1662) y Christiaan Huygens (1629-1695)

Copia de la pintura de François II Quesnel, que fue hecha por Gérard Edelinck en 1691. Recuperado de: https://es.wikipedia.org/wiki/Blaise_Pascal Pintura de Caspar Nestcher de 1671. Recuperado de: https://es.wikipedia.org/wiki/Christiaan_Huygens

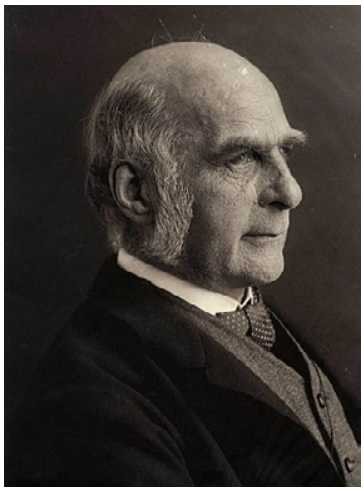
Figura 1.4. Edmund Halley (1656-1741)

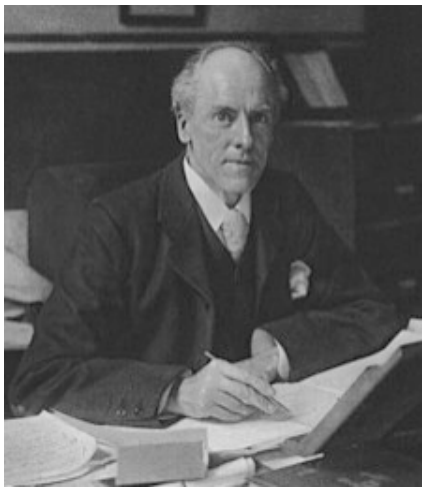


Retrato de Richard Phillips de 1722. Recuperado de: https://es.wikipedia.org/wiki/Edmund_Halley

En el siglo XIX, la estadística entra en una nueva fase de su desarrollo con la generalización del método para estudiar fenómenos de las ciencias naturales y sociales. Francis Galton (1822-1911) y Karl Pearson (1857-1936) se pueden considerar como los padres de la estadística moderna, pues a ellos se debe el paso de la estadística deductiva a la estadística inductiva.

Figura 1.5. Francis Galton y Karl Pearson





Eveleen Myers, 1890, recuperado de: https://es.wikipedia.org/wiki/Francis_Galton
Desconocido, 1910, recuperado de: https://es.wikipedia.org/wiki/Karl_Pearson

Los fundamentos de la estadística actual y muchos de los métodos de inferencia son debidos a Sir Ronald Aylmer Fisher (1890-1962). Se interesó primeramente por la eugenesia, lo que le conduce, siguiendo los pasos de Galton a la investigación estadística; sus trabajos culminan con la publicación de la obra *Statistical methods for researchers.* En él aparece la metodología estadística tal y como hoy se conoce.

Figura 1.6. Sir Ronald Aylmer Fisher (1890-1962)

Desconocido, 1913 recuperado de: https://commons.wikimedia.org/w/index.php?curid=42616717

A partir de mediados del siglo XX comienza lo que se puede denominar la estadística moderna, uno de los factores determinantes es la aparición y popularización de las computadoras. Las aplicaciones en este periodo de la Estadística a la Economía conducen a una disciplina con contenido propio: la Econometría. La investigación estadística en problemas militares durante la segunda guerra mundial y los nuevos métodos de programación matemática dan lugar a la Investigación de Operaciones.

El centro de gravedad de la metodología estadística se empieza a desplazar a técnicas de computación intensiva aplicadas a grandes masas de datos, lo que se conoce actualmente como "Big Data", y se empieza a considerar el método estadístico como un proceso iterativo de búsqueda del modelo ideal. Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día. Sin embargo, no es la cantidad de datos lo que es importante sino lo que las organizaciones hacen con los datos. Big Data se puede aplicar para obtener ideas que conduzcan a mejores decisiones y movimientos de negocios estratégicos.

1.2 Marco Metodológico de la Investigación y necesidad de la Estadística

La investigación es esencialmente una actividad o proceso orientado básicamente a dos fines: la extensión del conocimiento y la solución de un problema. Las características que definen su naturaleza se describen a continuación:

a. Puede tomar una variedad de formas.
b. Debe ser válida:
   - » Validez interna: extensión para la cual los resultados pueden ser exactamente interpretados;
   - » Validez externa: extensión para la cual los resultados pueden ser generalizados a poblaciones y condiciones.

c. Debe ser confiable; la confiabilidad de la investigación tiene que ver con la replicabilidad y la consistencia de los métodos, condiciones y resultados.
d. Debe ser sistemática. Los pasos básicos que se aplican generalmente son los siguientes:
   i. Observar un problema o fenómeno de interés;
   ii. Formular una hipótesis;
   iii. Experimentar para comprobar la hipótesis;
   iv. Emitir conclusiones.

La investigación, según Ezequiel Ander-Egg (1), es un procedimiento reflexivo, sistemático, controlado y crítico, que permite descubrir nuevos hechos o datos, relaciones o leyes, en cualquier campo del conocimiento humano. La investigación social es el proceso que, utilizando la metodología científica, permite obtener nuevos conocimientos en el campo de la realidad social. Dicho autor estableció su propio esquema de los pasos que debe contener la investigación, los cuales se muestran en la figura 1.7. Como se aprecia en dicho esquema, no se puede concebir la investigación si no se aplica o utiliza la Estadística, la cual se define como la rama de la ciencia que estudia las reglas para recolectar, capturar, organizar, presentar, procesar y analizar los datos obtenidos al realizar varios ensayos de un experimento y para inferir conclusiones acerca de este último. Proporciona, además, los métodos para el diseño estadístico de experimentos y para tomar decisiones cuando aparecen situaciones de incertidumbre.

Algunos autores establecen que la estadística no es ciencia, ya que algunas de las reglas que emplea son empíricas, como el hecho de realizar una tabla de frecuencias, lo cual se verá más adelante.

Al partir de esta definición, claramente se puede apreciar en el esquema de investigación de Ander-Egg que la Estadística interviene en las técnicas que se utilizarán, en el trabajo de recolección de datos en campo, en la ordenación y clasificación de datos, y en el procesamiento de estos.

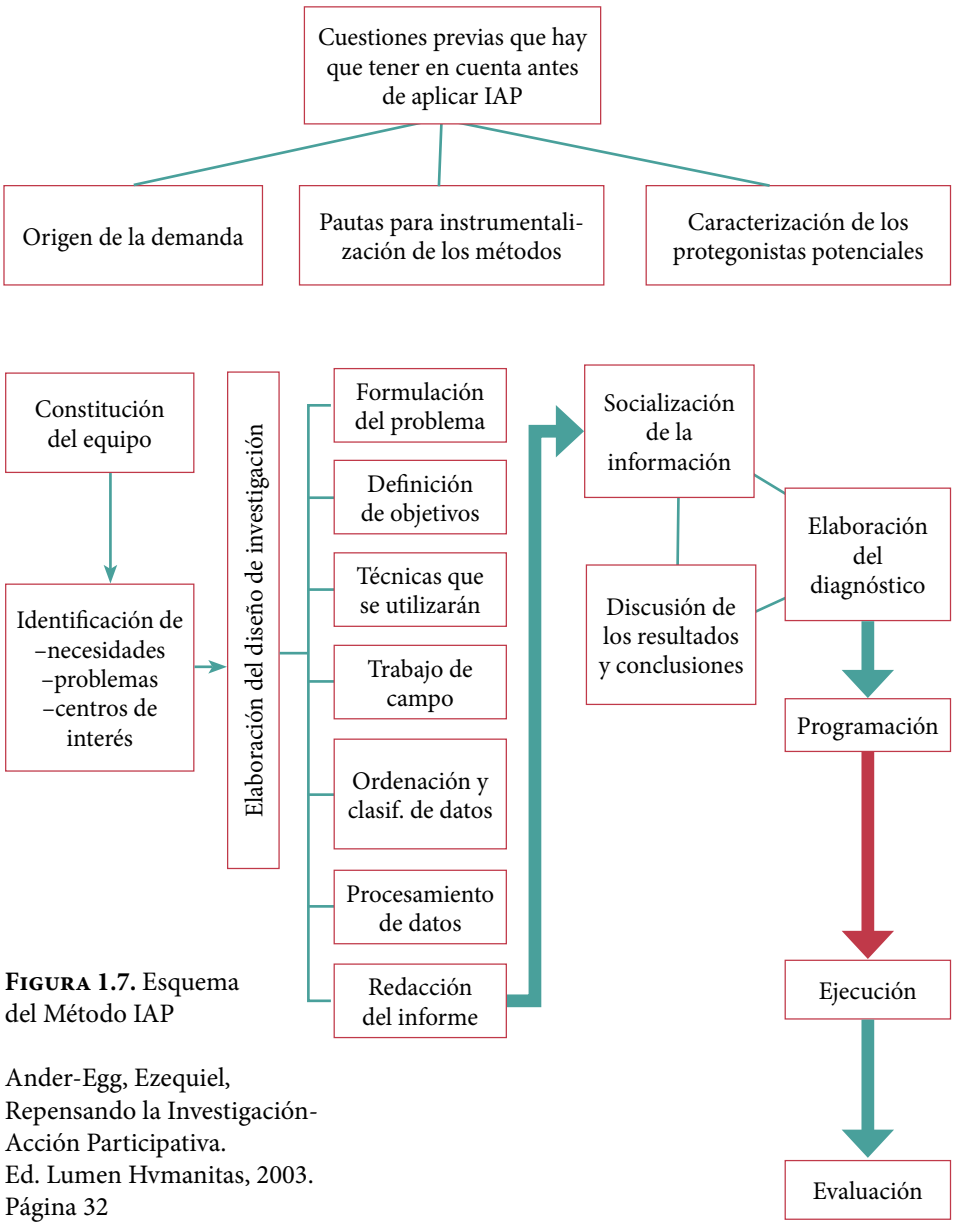


Figura 1.7. Esquema del Método IAP

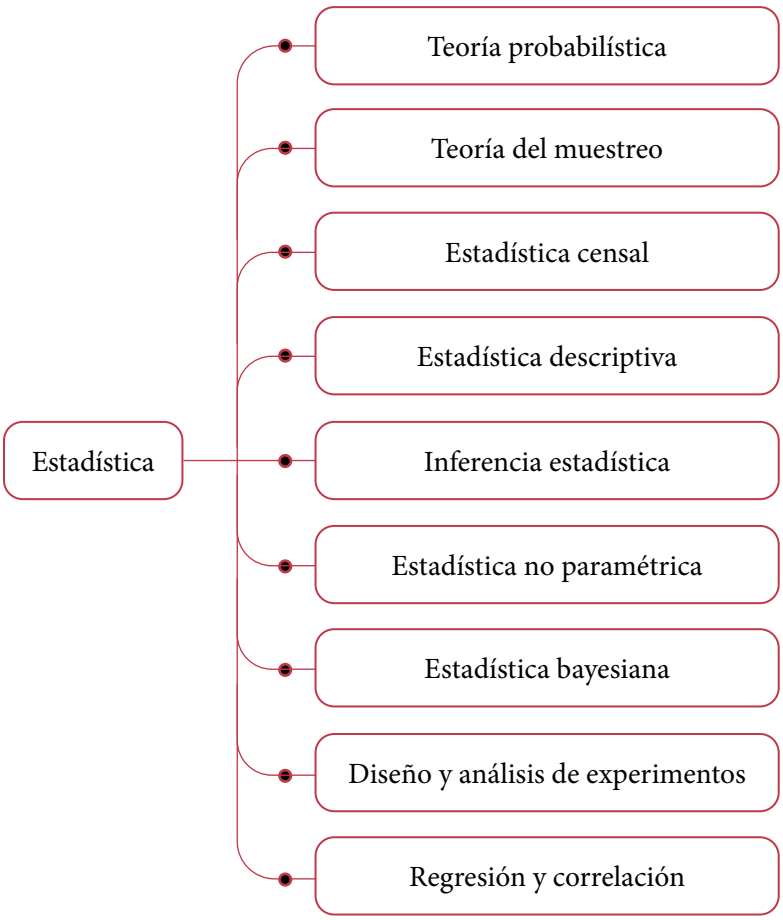
Ander-Egg, Ezequiel, Repensando la Investigación-Acción Participativa. Ed. Lumen Hvmanitas, 2003. Página 32

1.3. Clasificación de la Estadística

Al realizar un análisis estadístico, el estudio puede llevarse a efecto considerando una sola variable, por ejemplo, las ganancias, el rendimiento, los costos, índice de masa corporal, etcétera. A este tipo de Estadística se le conoce como de una variable. Si el análisis comprende dos o más variables, como son peso y estatura; temperatura y presión; temperatura, presión y tiempo; etcétera, se le denomina Estadística de varias variables.

Las partes en las que se descompone la estadística se muestran en la figura 1.7 y se definen a continuación.

Figura 1.7. Clasificación de la Estadística



Teoría Probabilística: rama de la Matemática que proporciona los fundamentos, modelos matemáticos y el lenguaje que se usa en la Estadística.

Teoría del Muestreo: Es la rama de la Estadística que se encarga de definir las reglas para tomar muestras de una población específica, el tamaño de dichas muestras, el método a seguir para tomarlas y los parámetros que indicarán la representatividad de estas.

La Estadística Censal se refiere al estudio estadístico de las poblaciones en su conjunto, su dimensión, estructura, parámetros y su dinámica. La Demografía (del griego Δήμος dēmos 'pueblo' y γραφία grafía 'trazo, descripción' –estudio de la población–) es una rama del conocimiento que estudia las poblaciones humanas, su dimensión, estructura, evolución y características generales. La demografía estudia estadísticamente la estructura y la dinámica de las poblaciones, así como los procesos concretos que determinan su formación, conservación y desaparición. Tales procesos son los de fecundidad, mortalidad y migración (emigración e inmigración).

La Estadística Descriptiva es la rama de la Estadística que se encarga de analizar las reglas para recolectar, presentar y procesar los datos obtenidos al hacer una medición u observación de alguna característica particular de un objeto, con la finalidad de conocer su comportamiento. Si se conocen con certeza los valores que tomará la característica particular en cuestión, previamente al experimento, a dicha característica se le denomina determinística. En este caso se puede conocer su comportamiento sin necesidad de hacer el experimento; si es el caso, el experimento se realizará con la finalidad de comprobar los resultados esperados. Si los valores que tomará la característica no pueden predecirse con certeza, antes del experimento, a dicha variable se le denomina aleatoria. Por otra parte, dentro del estudio de características aleatorias, se puede ver que existen dos tipos:

a. Variables aleatorias discretas son aquellas cuyos resultados pueden ser medidos en forma discreta; por ejemplo: el número de llegadas a una cola, el número de defectos en un lote, el número de ases que se obtienen en un juego de pocker, etcétera.
b. Variables aleatorias continuas, son aquellas que tienen unidades de medida continua; por ejemplo: la cantidad de leche que produce una vaca diariamente, el tiempo de vida de un producto, el tiempo de espera en una cola, etcétera.

La Inferencia Estadística es la rama de la Estadística que proporciona las reglas para estimar ciertos valores de una población, con base en los resultados de una muestra, formular hipótesis sobre la verdad de estas estimaciones y tomar decisiones de acuerdo con estos resultados. La estimación de parámetros poblacionales a partir de muestras se realiza a través de estimadores puntuales, de los cuales hay que determinar qué cualidades deben reunir para ser válidos y representativos, o a través de estimadores por intervalos de confianza, de los cuales hay que determinar la distribución de probabilidad que presentan y el nivel de confianza que se desea tener.

Las hipótesis estadísticas que se formulen deben ser probadas para comprobar su validez y representatividad. Tanto la estimación por intervalos de confianza como las pruebas de hipótesis se pueden clasificar para una población o para dos poblaciones o más. Para una población, generalmente los parámetros que se estiman son la media, varianza, desviación estándar, fracción de éxitos, tamaño poblacional. Para dos poblaciones generalmente lo que se estima es el cociente entre varianzas, la diferencia de medias o la diferencia de proporciones. También dentro de la Estadística Inferencial se ven otro tipo de pruebas estadísticas como lo son las pruebas de bondad de ajuste, que permiten probar la adecuación de un conjunto de datos obtenidos empíricamente a un modelo probabilístico específico, principalmente el modelo probabilístico normal, a lo que se denomina pruebas de normalidad.

La Estadística Paramétrica es la rama de la Estadística Inferencial que comprende los procedimientos estadísticos y de decisión que están basados en las distribuciones de los datos reales, los cuales generalmente se suponen normales. Estos son determinados usando un número finito de parámetros. Para aplicar la estadística paramétrica se requiere conocer la distribución de probabilidad que siguen los datos o en su defecto, con base en el Teorema del Límite Central, tomar muestras "grandes". Cuando se desconoce la distribución de probabilidad que siguen los datos y no se pueden tomar muestras "grandes" por restricciones económicas o de otro tipo, entonces se debe aplicar la Estadística No Paramétrica. Por ejemplo, los datos categorizados en: niños, jóvenes, adultos y ancianos no pueden ser interpretados mediante la estadística paramétrica, ya que no se puede hallar un parámetro numérico (como por ejemplo la media de edad) cuando los datos no son numéricos.

La Estadística Bayesiana es una rama de la Estadística en la que la evidencia sobre el verdadero estado del mundo se expresa en términos de estimaciones extraídas de datos históricos duros, así como de estimaciones basadas en la

opinión de expertos, con ciertos grados de creencia o, más específicamente, las probabilidades bayesianas.

El Diseño y Análisis de Experimentos es una rama de la Estadística que permite plantear si una o más variables conocidas como efectos, dependen a su vez de otras variables llamadas posibles causas o posibles factores. Esta rama de la Estadística permite diseñar, realizar y cuantificar, a través de experimentos, el grado de dependencia que existe entre las causas y los efectos. En un diseño experimental se manejan deliberadamente una o más variables, vinculadas a las causas, para medir el efecto que tienen sobre otras variables de interés llamadas efectos. El diseño experimental establece la serie de pautas a seguir para determinar qué variables hay que considerar, de qué manera, cuántas veces hay que repetir el experimento y en qué orden para poder establecer con un grado de confianza predefinido la necesidad de una presunta relación de causa-efecto. El diseño experimental encuentra aplicaciones muy diversas en la industria, agricultura, mercadotecnia, medicina, ecología, ciencias de la conducta, etcétera, constituyendo una fase esencial en el desarrollo de un estudio experimental.

Una regresión es un ajuste de un conjunto de puntos obtenidos empíricamente a un modelo matemático en particular. Si el ajuste es entre dos variables $y=f(x)$, la regresión puede ser lineal, polinomial, exponencial, logarítmica, trigonométrica, etcétera. Si el ajuste es a un modelo matemático de varias variables, se denomina regresión múltiple.

Una correlación corresponde con la definición de indicadores que permitan determinar qué tan bueno es el ajuste entre el conjunto de puntos dados y el modelo matemático usado.

1.4. Síntesis de la Teoría Probabilística

Como se recordará del volumen previo de Fundamentos de Probabilidad y sus aplicaciones, históricamente existen cuatro escuelas de probabilidad que la definen de la siguiente forma:

Escuela Clásica o de Laplace

Si un evento A contenido en el espacio muestral finito S de un experimento, está formado por $N(A)$ puntos muestrales y el espacio muestral por $N(S)$ puntos muestrales igualmente verosímiles o que tienen la misma posibilidad de ocurrir (equiprobables), se dice que la probabilidad de que el evento A ocurra está dada por la relación

$$p(A) = \frac{N(A)}{N(S)} \tag{1.1}$$

Escuela Frecuentista o de Von Misses

Suponga que se realiza un experimento n veces y se observa que un evento A ocurre $n(A)$veces, el cociente $fA = n(A)/n$ se denomina frecuencia relativa del evento A. En el enfoque de la Estadística se estima la probabilidad de la ocurrencia de un evento A, a partir del concepto de frecuencia relativa. La interpretación frecuencial de probabilidad consiste en asignar como probabilidad del evento A a la relación $n(A)/n$, que es su frecuencia relativa. Es de esperar que, mientras mayor sea el número de veces que se realice el experimento, esta aproximación será mejor y, en el límite, se obtendrá el valor preciso:

$$p(A) = Lim_{n\to\infty} \frac{n(A)}{n} \tag{1.2}$$

Escuela Subjetivista o de Savage

Bajo esta escuela, la probabilidad de un evento A puede ser simplemente una medida del grado de credibilidad que se tiene sobre la ocurrencia del evento A,

expresado en términos numéricos. Se ha desarrollado una serie de teorías subjetivistas, que consideran que la probabilidad de un evento es el grado de creencia que un sujeto tiene sobre la ocurrencia de ese evento, determinada a partir de su intuición, sentimiento, sentido común, experiencia o conocimiento. Otra persona puede tener diferente opinión o información distinta y asignar una probabilidad diferente al mismo resultado. Se requiere de un individuo idealmente racional que enjuicie la fuerza de la evidencia, desde la más absoluta imparcialidad, haciéndola equivaler a un determinado grado de probabilidad.

Escuela Axiomática, Constructivista o de Kolmogorov

Sean S el espacio muestral y A cualquier evento de S. Se llamará función de probabilidad sobre el espacio muestral S a $p(A)$ si satisface los siguientes axiomas:

1. $P(A) \geq 0$ (1.3)
2. $P(S) = 1$ (1.4)
3. Si $A \cap B = \phi \quad \rightarrow \quad p(A \cup B) = p(A) + p(B)$ (1.5)

En el contexto de la ingeniería, cada una es valiosa en sí misma y las cuatro son aplicables en diferentes situaciones.

Probabilidad Condicional, Independencia Estadística y Teorema de Bayes

Sea A un evento previo a un segundo evento B, la probabilidad condicional de B dado A, lo cual se escribe como $p(B \mid A)$, se define como:

$$p(B|A) = \frac{p(A \cap B)}{p(A)} \tag{1.6}$$

De la misma forma se define

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{1.7}$$

Nótese que no existe conmutatividad en las dos definiciones anteriores, son conceptos diferentes.

Una propiedad importante del concepto de probabilidad condicional parte de despejar la probabilidad de la intersección de A y B en las ecuaciones 1.6 y 1.7 e igualar:

$$P(A \cap B) = P(B|A)\,P(A) = P(A|B)\,P(B) \tag{1.8}$$

Suponga que se realiza un evento A y posteriormente un evento B, si al calcular la probabilidad de B dado A sucede que $P(B|A) = P(B)$ o en forma inversa $p(A|B) = p(A)$, se dice que los eventos A y B son estadísticamente independientes.

De otra forma, fácilmente se puede demostrar que dos eventos A y B son estadísticamente independientes si se cumple que:

$$P(A \cap B) = P(A)\,P(B) \tag{1.9}$$

Teorema o Ley de Probabilidad Total

$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + p(A|B_3)p(B_3) + \cdots + p(AB_k)p(B_k) \tag{1.10}$$

Teorema de Bayes

$$p(B_i \mid A) = \frac{p(A \mid B_i)\,p(B_i)}{\sum_{j=1}^{j=k} p(A \mid B_j)\,p(B_j)}$$

$$p(B_i|A) = \frac{p(A|B_i)\,p(B_i)}{p(A \mid B_1)\,p(B_1) + p(A \mid B_2)\,p(B_2) + \cdots + p(A \mid B_k)\,p(B_k)} \tag{1.11}$$

Se denomina variable aleatoria (o estocástica) a una relación funcional del tipo $f\colon S \to R$ que le asigna a cada evento simple de un espacio muestral S un número, cuyo valor puede ser discreto o continuo en los números reales. Estos valores posibles representan los resultados de experimentos que todavía no se han llevado a cabo o cantidades inciertas.

La función de probabilidad de una variable aleatoria x cumple las siguientes propiedades:

i. $0 \le f(x) \forall x \in [a,b]$

ii.
$$\sum_{x=a}^{x=b} f(x) = 1 \qquad x\_ \textit{discreta} \tag{1.12}$$

$$\int_a^b f(x)\,dx = 1 \qquad x\_ \textit{continua}$$

Asimismo, la función de probabilidad acumulada $F(x)$, de una variable aleatoria $x \in [a, b]$, se define como:

$$F(x) = \sum_{t=a}^{t=x} f(t) \qquad x\_ \textit{discreta} \tag{1.13}$$

$$F(x) = \int_a^x f(t)\,dt \qquad x\_ \textit{continua}$$

La cual cumple las siguientes propiedades:

i. $0 \le F(x) \le 1 \qquad \forall x \in [a,b]$

ii. $F(x)$ es no decreciente, es decir, (1.14)

$$Si \quad x_i \le x_2 \quad entonces \quad P(x_1) \le P(x_2)$$

iii. $F(a) = 0$

iv. $F(b) = 1$

Esperanza Matemática

Sea x una variable aleatoria, con función de probabilidad $p(x)$ para el caso discreto (en el caso discreto no toma todos los valores del intervalo, solo los que se definan en el dominio) y $f(x)$ para el caso continuo, en el intervalo $x \in [a, b]$. Sea $g(x)$ una función de x. Se define la Esperanza Matemática de $g(x)$ como:

$$E\{g(x)\} = \begin{Bmatrix} \sum_{\forall x \in [a,b]} g(x)\, p(x) & caso\_discreto \\ \int_{\forall x \in [a,b]} g(x)\, f(x)\, dx & caso\_continuo \end{Bmatrix} \tag{1.15}$$

Propiedades de la Esperanza Matemática:

i. $E\{k\} = k$

ii. $E\{kg(x)\} = kE\{g(x)\}$

iii. $E\{g_1(x) + g_2(x)\} = E\{g_1(x)\} + E\{g_2(x)\}$

iv. $E\{k_1 g_1(x) + k_2 g_2(\mathrm{x})\} = k_1 E\{g_1(x)\} + k_2 E\{g_2(x)\}$

v. $E\{k_1 g_1(x) k_2 g_2(x) + \cdots + k_n g_n(x)\} = k_1 E\{g_1(x)\} + k_2 E\{g_2(x)\} + \cdots + k_n E\{g_n(x)\}$

(1.16)

Sea x una variable aleatoria con función de probabilidad $p(x)$ si esta es discreta y $f(x)$ si es continua, en el intervalo $x \in [a, b]$; se define el momento de orden k con respecto al origen, como:

$$\mu'_k = \begin{Bmatrix} \sum_{\forall x \in [a,b]} x^k p(x) & caso\_discreto \\ \int_{\forall x \in [a,b]} x^k f(x)\, dx & caso\_continuo \end{Bmatrix} \tag{1.17}$$

Al momento de orden uno de la variable aleatoria x, se le conoce como media o valor esperado de x y representa una medida de la tendencia central que presenta la gráfica de la función de probabilidad de la variable aleatoria x:

$$\mu_x = \begin{Bmatrix} \sum_{\forall x \in [a,b]} xp(x) & caso\_discreto \\ \int_{\forall x \in [a,b]} xf(x)dx & caso\_continuo \end{Bmatrix} \tag{1.18}$$

El concepto de media de una variable aleatoria es fundamental para entender la tendencia central de su función de probabilidad; conviene analizar algunas de sus propiedades:

i. La media de una constante es la constante misma:

$$\mu_a = E\{a\} = a \tag{1.19}$$

ii. La media de una constante por una variable aleatoria es la constante por la media de esa variable aleatoria:

$$\mu_{ax} = E\{ax\} = aE\{x\} = a\mu_x \tag{1.20}$$

iii. La media de una suma de dos variables aleatorias es igual a la suma de las medias de cada una de ellas:

$$\mu_{x_1+x_2} = E\{x_1 + x_2\} = E\{x_1\} + E\{x_2\} = \mu_{x_1} + \mu_{x_2} \tag{1.21}$$

iv. La media de una combinación lineal de dos variables aleatorias es igual a la misma combinación lineal de las medias de cada una de las dos variables:

$$\mu_{a_1x_1+a_2x_2} = E\{a_1x_1 + a_2x_2\} = a_1E\{x_1\} + a_2E\{x_2\} = a_1\,\mu_{x_1} + a_2\,\mu_{x_2} \tag{1.22}$$

v. La media de una combinación lineal de *n* variables aleatorias es igual a la misma combinación lineal de las medias de cada una de las *n* variables aleatorias:

$$\mu_{a_1x_1+a_2x_2+\cdots+a_n x_n} = a_1\,\mu_{x_1} + a_2\,\mu_{x_2} + \cdots + a_n\,\mu_{x_n} \tag{1.23}$$

Sea x una variable aleatoria con función de probabilidad $p(x)$ si esta es discreta y $f(x)$ si es continua, en el intervalo $x \in [a, b]$; se define el momento de orden k con respecto a la media, como:

$$\mu_k = \begin{cases} \sum\limits_{\forall x \in [a,b]} (x-\mu_x)^k\, p(x) & caso\_discreto \\ \int\limits_{\forall x \in [a,b]} (x-\mu_x)^k\, f(x)\,dx & caso\_continuo \end{cases} \tag{1.24}$$

Observe que existe una relación directa entre los momentos de orden k con respecto a la media y los momentos de orden k con respecto al origen. Se mostrarán a continuación las relaciones para los cuatro primeros momentos con respecto a la media y se proporcionará una fórmula para cualquier valor de k.

El momento de orden uno con respecto a la media es cero, como se observa a continuación:

$$\mu_1 = E\{(x-\mu_x)\} = E\{x\} - \mu_x = \mu_x - \mu_x = 0 \tag{1.25}$$

El momento de orden dos con respecto a la media:

$$\mu_2 = E\{(x-\mu_x)^2\} = E\{x^2 - 2\mu_x + \mu_x^2\} = \mathrm{E}\{x^2\} - 2\mu_x E\{x\} + \mu_x^2$$

$$\mu_2 = E\{x^2\} - \mu_x^2 = \mu'_2 - \mu'^2_1 \tag{1.26}$$

El momento de orden dos con respecto a la media recibe el nombre de Varianza y se representa por σ^2, la cual es una medida del grado de dispersión de la gráfica de la función de probabilidad de la variable aleatoria x con respecto a su media:

$$var\{x\} = \sigma_x^2 = \mu_2 = \left\{ \begin{array}{ll} \sum\limits_{\forall x \in [a,b]} (x-\mu_x)^2 \; p(x) & caso\_discreto \\ \int\limits_{\forall x \in [a,b]} (x-\mu_x)^2 f(x)\,dx & caso\_continuo \end{array} \right\} \tag{1.27}$$

La varianza también se puede calcular como el momento de segundo orden con respecto al origen menos el momento de primer orden con respecto al origen al cuadrado, es decir:

$$\sigma_x^2 = var(x) = \mu'_2 - \mu'^2_1 = E\{x^2\} - \mu_x^2 \tag{1.28}$$

El concepto de varianza de una variable aleatoria es fundamental para entender la dispersión de la variable y de su función de probabilidad, por lo que conviene analizar algunas de sus propiedades:

i. La varianza de una constante es cero:

$$\sigma_a^2 = var\{a\} = E\{(a-a)^2\} = 0 \tag{1.29}$$

ii. La varianza de una constante por una variable aleatoria es la constante al cuadrado por la varianza de esa variable aleatoria:

$$\sigma^2_{ax} = var\{ax\} = E\{(ax - a\mu_x)^2\} = E\{a^2(x-\mu_x)^2\} = a^2\,\sigma^2_x \tag{1.30}$$

iii. La varianza de una suma de dos variables aleatorias es igual a la suma de las varianzas de cada una de ellas más dos veces la covarianza entre ellas, es decir:

$$\sigma^2_{x_1+x_2} = \text{var}\{x_1\} + \text{var}\{x_2\} + 2\text{cov}\{x_1, x_2\} \tag{1.31}$$

$$\sigma^2_{x_1+x_2} = \sigma^2_{x_1} + \sigma^2_{x_2} + 2\sigma^2_{x_1,\,x_2}$$

La covarianza de dos variables aleatorias x_1 y x_2, con funciones de probabilidad $f_1(x_1)$ y $f_2(x_2)$, se define como:

$$\sigma^2_{x_1,x_2} = \text{cov}\{x_1, x_2\} = E\{(x_1-\mu_1)(x_2-\mu_2)\}$$

$$\sigma^2_{x_1,x_2} = \left\{ \begin{array}{ll} \sum_{\forall x_1}\sum_{\forall x_2}(x_1-\mu_1)(x_2-\mu_2)\,f_1(x_1)\,f_2(x_2) & x_1, x_2 \textit{ discretas} \\ \int_{\forall x_1}\int_{\forall x_1}(x_1-\mu_1)(x_2-\mu_2)\,f_1(x_1)\,f_2(x_2)\,dx_1 dx_2 & x_1, x_2 \textit{ continuas} \end{array} \right\} \tag{1.32}$$

Más adelante se analizarán las propiedades de la covarianza, por lo pronto, cabe remarcar que, si la variable aleatoria x_1 no depende de la variable aleatoria x_2, es decir, si ambas son estadísticamente independientes, la covarianza de ambas es cero, por lo que se puede afirmar que:

$$\textit{Si } x_1 \textit{ es independiente de } x_2 \quad \Rightarrow \quad \sigma^2_{x_1,\,x_2} = \text{cov}\{x_1, x_2\} = 0 \tag{1.33}$$

Lo que implica que:

$$\textit{Si } x_1 \textit{ es independiente de } x_2 \textit{ entonces} \quad \sigma^2_{x_1+x_2} = \sigma^2_{x_1} + \sigma^2_{x_2} \tag{1.34}$$

iv. La varianza de una combinación lineal de dos variables aleatorias es igual a la misma combinación lineal de las varianzas de cada una de las dos variables, más la covarianza entre ellas:

$$\sigma^2_{a_1x_1+a_2x_2} = a_1^2\,\sigma^2_{x_1} + a_2^2\,\sigma^2_{x_2} + 2a_1a_2\sigma^2_{x_1,\,x_2} \tag{1.35}$$

Nuevamente:

Si x_1 *es independiente de* x_2 *entonces* $\sigma^2_{a_1x_1 + a_2x_2} = a_1^2\,\sigma^2_{x_1} + a_2^2\,\sigma^2_{x_2}$ (1.36)

v. La varianza de una combinación lineal de *n* variables aleatorias es igual a la misma combinación lineal de las varianzas de cada una de las *n* variables aleatorias más dos veces la covarianza de dos en dos de cada una de las variables aleatorias definidas:

$$\sigma^2_{a_1x_1 + a_2x_2} = a_1^2\,\sigma^2_{x_1} + a_2^2\,\sigma^2_{x_2} + \cdots + a_n^2\,\sigma^2_{x_n} \tag{1.37}$$

$$+ 2a_1\,a_2\,\sigma^2_{x_1,\,x_2} + 2a_1\,a_3\,\sigma^2_{x_1,\,x_3} + \cdots + 2a_{n-1}\,a_n\,\sigma^2_{x_n-1,\,x_n}$$

De la misma forma:

Si x_i *es independiente de* $x_j\ \forall i \neq j$ *entonces*

$$\sigma^2_{a_1x_1 + a_2x_2 + \cdots + a_nx_n} = a_1^2\,\sigma^2_{x_1} + a_2^2\,\sigma^2_{x_2} + \cdots + a_{x_n}^2\,\sigma^2_{x_n} \tag{1.38}$$

A la raíz cuadrada de la varianza de una variable aleatoria x se le conoce como Desviación Estándar de x, y se acostumbra representarla como σ_x:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\text{var}\{x\}} = \sqrt{E\{(x-\mu_x)^2\}} \tag{1.39}$$

Un concepto que se utiliza como alternativa al concepto de desviación estándar es el denominado desviación media:

$$\textit{Desv Media} = E\{|x-\mu_x|\} \tag{1.40}$$

El valor absoluto en la expresión anterior es necesario, de no hacerlo, se haría cero.

Se le llama coeficiente de variación (CV) al cociente de la desviación estándar entre la media de la variable aleatoria:

$$CV = \frac{\sigma_x}{\mu_x} \tag{1.41}$$

Nótese que el momento de orden tres con respecto a la media puede expresarse en función de los momentos con respecto al origen

$$\mu_3 = E\left\{(x-\mu_x)^3\right\} = \mu'_3 - 3\mu'_1\,\sigma_x^2 - \mu'^3_1 \tag{1.42}$$

Un concepto importante que se define en términos de los momentos de orden tres y de orden dos con respecto a la media es el coeficiente de asimetría, el cual representa la asimetría que presenta la gráfica de la función de probabilidad de la variable aleatoria x con respecto a su media

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{1.43}$$

El momento de orden cuatro con respecto a la media también puede definirse en función de los momentos con respecto al origen:

$$\mu_4 = E\left\{(x-\mu_x)^4\right\} = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'^2_1\,\sigma_x^2 + 3\mu'^4_1 \tag{1.44}$$

Otro concepto importante que se define en términos de los momentos de orden cuatro y dos con respecto a la media es el coeficiente de curtosis, el cual representa el grado de aplanamiento o picudez que presenta la gráfica de la función de probabilidad de la variable aleatoria x:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \tag{1.45}$$

En general, los momentos de orden k con respecto a la media pueden expresarse en términos de los momentos con respecto al origen, utilizando la siguiente expresión:

$$\mu_k = \sum_{j=0}^{j=k} (-1)^j \binom{k}{j} \mu^j\,\mu'_{k-j} \qquad k = 0,1,2,\cdots \tag{1.46}$$

Para el caso de una variable aleatoria discreta, existe una forma de estimar los percentiles a través de interpolación lineal, usando la expresión 3.12 de la ecuación de una recta dados dos de sus puntos. Sea x una variable aleatoria discreta, con función de probabilidad $p(x)$. Suponga que se desea conocer el percentil x_p para el cual, la probabilidad acumulada hasta ese punto es p, es decir, $p = p(x_p)$. Suponga también que $x_1 < x_p < x_2$ y que se sabe que (x_1, p_1) y (x_2, p_2), en donde, $p_1 = p(x_1)$ y $p_2 = p(x_2)$, entonces, el percentil xp estaría dado por la expresión:

$$x_p = \frac{x_2 - x_1}{p_2 - p_1}\left(p - p_1\right) + x_1 \tag{1.47}$$

En la anterior expresión, se debe buscar que los puntos x_1 y x_2 estén lo más cercanos posibles a x_p, para que la interpolación lineal se ajuste adecuadamente.

Función Generatriz de Momentos

Calcular los momentos definidos anteriormente puede llegar a ser muy complicado, por lo cual, existe una forma alterna para determinarlos, a través de lo que se conoce como la Función Generatriz de Momentos:

$$FGM(t) = E\{e^{tx}\} = \begin{cases} \sum_{\forall x \in [a,b]} e^{tx} p(x) & caso\_discreto \\ \int_{\forall x \in [a,b]} e^{tx} f(x)\,dx & caso\_continuo \end{cases} \tag{1.48}$$

Se le denomina Función Generatriz de Momentos a esta expresión porque al derivarla con respecto a t y evaluar en $t = 0$, se obtienen los diversos momentos con respecto al origen:

$$\frac{d^k}{dt^k} FGM(t)\,|_{t=0} = E\{x^k e^{tx}\}_{t=0} = \mu'_k \tag{1.49}$$

Cabe señalar que de existir, la Función Generatriz de Momentos de una variable aleatoria es única de conformidad con las propiedades de unicidad de una sumatoria definida y de una integral definida.

Función Característica de una variable aleatoria

Sea x una variable aleatoria, se le denomina Función Característica de x a la expresión:

$$FCx(t) = E\{e^{itx}\} = \begin{cases} \sum_{\forall x \in [a,b]} e^{itx} p(x) & caso\_discreto \\ \int_{\forall x \in [a,b]} e^{itx} f(x)\,dx & caso\_continuo \end{cases} \tag{1.50}$$

Cabe señalar que de existir, la Función Característica de una variable aleatoria es única de conformidad con las propiedades de unicidad de una sumatoria definida y de una integral definida.

Funciones de probabilidad de variables aleatorias conjuntas continuas

Se dice que $f(x, y)$ es una función de probabilidad bivariada continua, si para cada pareja ordenada (x, y) del dominio de definición de x y y, existe un número $f(x, y)$ que cumple las siguientes propiedades:

$$\text{i.} \quad f(x,y) \geq 0 \qquad \begin{matrix} \forall x \in Dominio\_de\_x \\ \forall y \in Dominio\_de\_y \end{matrix} \tag{1.51}$$

$$\text{ii.} \quad \int_{\forall x} \int_{\forall y} f(x,y)\,dydx = \int_{\forall y} \int_{\forall x} f(x,y)\,dxdy = 1$$

A las funciones:

$$f_x(x) = \int_{\forall y} f(x,y)\,dy \tag{1.52}$$

$$f_y(y) = \int_{\forall x} f(x,y)\,dx$$

Se les denomina funciones de probabilidad marginales continuas de x y de y respectivamente. Cada una de estas distribuciones de probabilidad cumple con las propiedades de una función de probabilidad continua y posee medidas de tendencia central como: media, mediana, moda; medidas de dispersión: varianza, desviación estándar, coeficiente de variación; medidas de forma como coeficiente de asimetría, coeficiente de curtosis, etcétera. Asimismo, cada una de las funciones de probabilidad marginales tiene una función generatriz de momentos que la define.

A las funciones de probabilidad:

$$f_{y|x}(y) = \frac{f(x,y)}{f_x(x)} \; \forall x,y$$

$$f_{y|x}(x) = \frac{f(x,y)}{f_y(y)} \; \forall x,y \tag{1.53}$$

Se les denomina funciones de probabilidad continuas condicionales bivariadas de y dado x, así como de *x* dado y, respectivamente.

A las expresiones:

$$E\{y|x\} = \int_{\forall y} y f_{y|x}(y)\,dy \tag{1.54}$$
$$E\{x|y\} = \int_{\forall x} x f_{x|y}(x)\,dx$$

Se les denomina esperanzas condicionales discretas de *y* dado *x*, así como de *x* dado *y*, respectivamente. A las gráficas de estas esperanzas condicionales se les denomina Regresión de *y* dado *x*, así como de *x* dado *y*, respectivamente.

La variable aleatoria *x* es estadísticamente independiente de la variable aleatoria *y*, si y solo si,

$$f_{xy}(x,y) = f_x(x) f_y(y) \tag{1.55}$$

Este resultado se puede generalizar para el caso de *n* variables estadísticamente independientes entre sí, $x_1, x_2, \cdots, x_n$:

$$F_{x_1 x_2 \cdots x_n}(x_1, x_2, \cdots, x_n) = F_{x_1}(x_1)\; F_{x_2}(x_2) \cdots F_{x_n}(x_n) \tag{1.56}$$

Sean las variables aleatorias *x* y y, estadísticamente independientes entre sí, con funciones generatrices de momentos $FGM_x(t)$ y $FGM_y(t)$, y sea $z = g(x,y)$. La función generatriz de momentos de z se obtiene como el producto de las funciones generatrices de *x* y y, es decir,

$$FGM_z(t) = FGM_x(t)\; FGM_y(t) \tag{1.57}$$

También se puede generalizar esta expresión:

$$FGM_{x_1 x_2 \cdots x_n}(t) = FGM_{x_1}(t)\; FGM_{x_2}(t) \cdots FGM_{x_n}(t) \tag{1.58}$$

Modelos probabilísticos de fenómenos aleatorios

Modelos Probabilísticos Discretos

Distribución	Parámetros	Variable	Valores que toma x	Función de Probabilidad: p(x)	Función de Probabilidad Acumulada: P(x)=p(t £ x)	Media	Varianza	Función Generatriz de Momentos
Bernoulli	0 < p < 1, probabilidad de éxito en cada ensayo	X= éxito en un ensayo	x= 0, fracaso x= 1, éxito	$p(x)=\begin{cases} p^x(1-p)^{(1-x)} & x=0,1 \\ 0 & en\_otro\_caso \end{cases}$	$P(x)=p(t\le x)=\begin{cases} 1-p & x=0 \\ 1 & x=1 \end{cases}$	p	p(1-p)	$fgm(t)=pe^t+(1-p)$
Binomial	n= 1, 2, …, número de ensayos o tamaño de muestra 0 < p < 1, probabilidad constante de éxito en cada ensayo	x= número de éxitos en n ensayos	x= 0, 1, 2, …, n	$p(x)=\begin{cases} \binom{n}{x}p^x(1-p)^{(n-x)} & x=0,1,2,\ldots,n \\ 0 & en\_otro\_caso \end{cases}$	$p(t\le x)=\sum_{t=0}^{t=x}\binom{n}{t}p^t(1-p)^{(n-t)}$	np	np(1-p)	$fgm(t)=\left[(pe^t+(1-p)\right]^n$
Geométrica	0 < p < 1, probabilidad de éxito en cada ensayo	x= número de ensayos hasta tener el primer éxito	x= 0, 1, 2, …	$p(x)=\begin{cases} p(1-p)^{(x-1)} & x=0,1,2,\ldots \\ 0 & en\_otro\_caso \end{cases}$	$p(t\le x)=\sum_{t=0}^{t=x}p(1-p)^{(t-1)}$	1/p	$(1-p)/p^2$	$fgm(t)=pe^t/\left[1-(1-p)e^t\right]$
Pascal (Binomial Negativa)	0 < p < 1, probabilidad de éxito en cada ensayo r=1, 2, … (r>0)	x= número de ensayos hasta tener r éxitos	x= r, r+1, r+2, …	$p(x)=\begin{cases} \binom{x-1}{r-1}p^r(1-p)^{(x-1)} & x=0,1,2,\cdots \\ 0 & en\_otro\_caso \end{cases}$	$p(t\le x)=\sum_{t=0}^{t=x}\binom{t-1}{r-1}p^r(1-p)^{(t-r)}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$fgm(t)=\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$
Hipergeométrica	N=1, 2, …, tamaño del lote o población n= 1, 2, …, N, tamaño de muestra D= 1, 2, …, N, Número de Defectuosos en el Lote o Población	x= número de éxitos en una muestra de tamaño n, para un lote de tamaño N, donde existen D éxitos	x= 0, 1, 2, …, mín(n, D)	$p(x)=\begin{cases} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} & x=0,1,2,\ldots,mín(n,D) \\ 0 & en\_otro\_caso \end{cases}$	$p(t\le x)=\sum_{t=0}^{t=x}\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$	$n\left(\frac{D}{N}\right)$	$n\left(\frac{D}{N}\right)\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right)$	
Poisson	c > 0, número de ocurrencias en promedio por unidad	x= número de ocurrencias, éxitos o llegadas por unidad	x= 0, 1, 2, …	$p(x)=\begin{cases} e^{-c}\frac{c^x}{x!} & x=0,1 \\ 0 & en\_otro\_caso \end{cases}$	$p(t\le x)=\sum_{t=0}^{t=x}e^{-c}\frac{c^x}{x!}$	c	c	$fgm(t)=e^{c\left(e^t-1\right)}$

Figura 1.8. Modelos Probabilísticos Discretos

Modelos Probabilísticos Continuos

Distribución	Parámetros	Valores que toma x	Función Densidad de Probabilidad: f(x)	Función de Distribución o de Probabilidad Acumulada: F(x)=p(t £ x)	Media	Varianza	Función Generatriz de Momentos
Uniforme	a, b, b>a	$a \le x \le b$ $x \in \Re$	$f(x)=\begin{cases}\frac{1}{b-a} & a\le x\le b\\ 0 & en\_otro\_caso\end{cases}$	$F(x)=\begin{cases}0 & si \quad x<a\\ \frac{x-a}{b-a} & si \quad a\le x\le b\\ 1 & si \quad x>b\end{cases}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Triangular	a, b, c c > b > a	$a \le x \le c$ $x \in \Re$	$f(x)=\begin{cases}\frac{2(x-a)}{(c-a)(b-a)} & si \quad a\le x\le b\\ \frac{2(c-x)}{(c-a)(c-b)} & si \quad b<x\le c\\ 0 & en\_cualquier\_otro\_caso\end{cases}$	$F(x)=\begin{cases}0 & si \quad x<a\\ \frac{(x-a)^2}{(c-a)(b-a)} & si \quad a\le x\le b\\ 1-\frac{(c-x)^2}{(c-a)(c-b)} & si \quad b<x\le c\\ 1 & si \quad x>c\end{cases}$	$\frac{c+b+a}{3}$	$\frac{1}{18}\left[a^2+b^2+c^2-ab-ac-bc\right]$	
Exponencial	l > 0	$x > 0$ $x \in \Re^+$	$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & en\_otro\_caso\end{cases}$	$F(x)=\begin{cases}0 & x\le 0\\ 1-e^{-\lambda x} & x>0\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\left(1-\frac{t}{\lambda}\right)^{-1}$
Normal	$\mu \in \Re$ $\sigma \in \Re$	$x \in \Re$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2}}dt$	μ	σ	$e^{\left[t\mu+\frac{\sigma^2t^2}{2}\right]}$
LogNormal	$\mu \in \Re$ $\sigma \in \Re$	$x > 0$ $x \in \Re^+$ $y = Ln(x)$	$f(x)=\frac{1}{x\sigma_y\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(Ln(x)-\mu_y)^2}{\sigma_y^2}}$	$F(x)=\frac{1}{\sigma_y\sqrt{2\pi}}\int_0^x\frac{e^{-\frac{1}{2}\frac{(Ln(t)-\mu_y)^2}{\sigma_y^2}}}{t}dt$	$\mu_x=e^{\left[\mu_y+\frac{\sigma_y^2}{2}\right]}$	$\sigma_x^2=\mu_x^2\left(e^{\sigma_y^2-1}\right)$	
Gamma	$r>0$ $\lambda>0$	$x>0$	$f(x)=\frac{\lambda}{\Gamma(r)}(\lambda x)^{(r-1)}e^{-\lambda x}$	$F(x)=\frac{\lambda}{\Gamma(r)}\int_0^x(\lambda t)^{(r-1)}e^{-\lambda t}dt$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\left(1-\frac{t}{\lambda}\right)^{-r}$
Beta	$r>0$ $\lambda>0$	$0\le x\le 1$	$f(x)=\frac{\Gamma(\lambda+r)}{\Gamma(\lambda)\Gamma(r)}x^{\lambda-1}(1-x)^{r-1}$ $0\le x\le 1 \quad \lambda>0 \quad r>0$	$F(x)=\frac{\Gamma(\lambda+r)}{\Gamma(\lambda)\Gamma(r)}\int_0^x t^{\lambda-1}(1-t)^{r-1}dt$	$\mu_x=\frac{\lambda}{\lambda+r}$	$\sigma_x^2=\frac{\lambda r}{(\lambda+r)^2(\lambda+r+1)}$	
Weibull	$\gamma \in \Re$ $\delta>0$ $\beta>0$	$x\ge\gamma$	$f(x)=\frac{\beta}{\delta}\left(\frac{x-\gamma}{\delta}\right)^{(\beta-1)}e^{-\left(\frac{x-\gamma}{\delta}\right)^{\beta}}$	$F(x)=\frac{\beta}{\delta}\int_\gamma^x\left(\frac{t-\gamma}{\delta}\right)^{(\beta-1)}e^{-\left(\frac{t-\gamma}{\delta}\right)^{\beta}}dt$	$\mu_x=\gamma+\delta\Gamma\left(\frac{1}{\beta}-1\right)$	$\sigma_x^2=\delta^2\left\{\Gamma\left(\frac{2}{\beta}+1\right)-\left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2\right\}$	

Figura 1.9. Modelos Probabilísticos Continuos

Modelos Probabilísticos de Distribuciones Muestrales

Distribución	Parámetros	Valores que toma x	Función Densidad de Probabilidad: f(x)	Función de Distribución o de Probabilidad Acumulada: F(x)=p(t ≤ x)	Media	Varianza
Ji Cuadrada (c^2)	$k \in \aleph$	$u = \chi^2 = z_1^2 + z_2^2 + \cdots + z_k^2$ $u = \chi^2 > 0$	$f(u) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} u^{\left(\frac{k}{2}-1\right)} e^{-\frac{u}{2}}$	$F(u) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \int_0^u v^{\left(\frac{k}{2}-1\right)} e^{-\frac{u}{2}}$	k	$2k$
t de Student	$k \in \aleph$	$t = \frac{z}{\sqrt{\frac{\chi^2}{k}}}$ $t \in \Re$	$f(t) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\sqrt{\pi k}\,\Gamma\left[\frac{k}{2}\right]} \frac{1}{\left[\frac{t^2}{k}+1\right]^{\left(\frac{k+1}{2}\right)}}$	$F(t) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\sqrt{\pi k}\,\Gamma\left[\frac{k}{2}\right]} \int_{-\infty}^{t} \frac{1}{\left[\frac{s^2}{k}+1\right]^{\left(\frac{k+1}{2}\right)}} \frac{1}{\left[\frac{s^2}{k}+1\right]^{\left(\frac{k+1}{2}\right)}} ds$	0	$\frac{k}{(k-2)}$
F de Fisher	$u, v \in \aleph$ $u > 0$ $v > 0$	$f = \frac{\frac{w}{u}}{\frac{y}{v}} > 0$ $w, y\_son\_va\_\chi^2$	$h(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\left(\frac{u}{2}\right)} f^{\left(\frac{u}{2}-1\right)}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\frac{u}{v} f + 1\right]^{\frac{(u+v)}{2}}}$	$H(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\left(\frac{u}{2}\right)}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \int_0^f \frac{t^{\left(\frac{u}{2}-1\right)}}{\left[\frac{u}{v}t+1\right]^{\frac{(u+v)}{2}}} dt$	$\mu_f = \frac{v}{(v-2)}$ $v > 2$	$\sigma_f^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)}$ $v > 4$

Figura 1.10. Modelos Probabilísticos de Distribuciones Muestrales

Teorema de Aditividad o de Reproducibilidad de la Distribución Normal

Sean

$$x_1 \sim N(\mu_1, \sigma_1)$$
$$x_2 \sim N(\mu_2, \sigma_2)$$
$$\vdots$$
$$x_n \sim N(\mu_n, \sigma_n)$$

n variables aleatorias normales, estadísticamente independientes y sea

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

Entonces

$$y \sim N(\mu_y, \sigma_y) \tag{1.59}$$

Donde

$$\mu_y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

$$\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Teorema de De Moivre - Laplace

Sea x una variable aleatoria con distribución binomial, con parámetro p, de media np y varianza np(1–p); su función de probabilidad está dada por

$$p_x(x) = \binom{n}{x} p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

La variable aleatoria x es el resultado de una suma de variables aleatorias con función de probabilidad de Bernoulli:

$$x = x_1 + x_2 + \cdots + x_n$$

donde

$$p(x_i) = \begin{Bmatrix} 1-p & x_i = 0 \\ p & x_i = 1 \end{Bmatrix}$$

$$\forall i = 1, 2, \cdots, n$$

Sea x una variable aleatoria normal con media np y varianza np(1–p), con función de probabilidad dada por la expresión:

$$\phi(x) = \frac{1}{\sqrt{2\pi np(1-p)}} \; e^{-\frac{1}{2}\left(\frac{x-np}{\sqrt{np\,(1-p)}}\right)^2}$$

Con función de probabilidad acumulada

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-np}{\sqrt{np(1-p)}}\right)^2} dt$$

Entonces:

$$p(a \le x \le b) = Lim_{n\to\infty}\left[\sum_{x=a}^{x=b} \binom{n}{x} p^x (1-p)^{n-x}\right] = \Phi(b) - \Phi(a) \qquad (1.60)$$

La expresión anterior implica también que

$$Lim_{n\to\infty}\binom{n}{x} p^x (1-p)^{n-x} = \frac{1}{\sqrt{2\pi np(1-p)}} \; e^{-\frac{1}{2}\left(\frac{x-np}{\sqrt{np\,(1-p)}}\right)^2} \qquad (1.61)$$

Este teorema establece que una variable aleatoria binomial, con media np y varianza $np(1–p)$ tiende a comportarse como una variable aleatoria normal, con la misma media y varianza, en la medida en que el tamaño de muestra tienda a infinito.

Teorema del Límite Central para una suma de variables aleatorias con la misma distribución de probabilidad

Teorema: Sean n variables aleatorias estadísticamente independientes, $x_1, x_2, \cdots, x_n$, con medias $\mu_1, \mu_2, \cdots, \mu_n$ y con desviaciones estándar $\sigma_1, \sigma_2, \cdots, \sigma_n$, respectivamente, con la misma distribución de probabilidad, sin importar de qué distribución de probabilidad se trate. Sea $y = x_1 + x_2 + \cdots + x_n$, entonces,

$$Lim_{n\to\infty}[y] \sim N(\mu_y, \sigma_y) \tag{1.62}$$

donde

$$\mu_y = \mu_1 + \mu_2 + \cdots + \mu_n$$

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$$

Teorema del Límite Central para una combinación lineal de variables aleatorias con cualquier distribución de probabilidad

Teorema: Sean n variables aleatorias estadísticamente independientes, $x_1, x_2, \cdots, x_n$, con medias $\mu_1, \mu_2, \cdots, \mu_n$ y con desviaciones estándar $\sigma_1, \sigma_2, \cdots, \sigma_n$, con cualquier distribución de probabilidad. Sea $y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, entonces,

$$Lim_{n\to\infty}[y] \sim N(\mu_y, \sigma_y) \tag{1.63}$$

donde

$$\mu_y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

$$\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Nótese que la media aritmética de una muestra se calcula como una suma de variables aleatorias idénticamente distribuidas

$$\bar{x} = \frac{1}{n}[x_1 + x_2 + \cdots + x_n]$$

Por lo que aplicando el teorema anterior se puede afirmar que la media muestral también es normal:

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{1.64}$$

Por el teorema de aditividad de la distribución normal, se puede afirmar que si una población presenta distribución normal, entonces la media aritmética de una muestra de dicha población también presenta distribución normal. De la misma forma, por el Teorema del Límite Central, se puede afirmar que si se extrae una muestra de tamaño grande de una población con cualquier distribución, la media muestral tenderá a tener distribución normal en la medida en que n tienda a infinito.

En la práctica profesional de la Ingeniería, el pensar en una muestra de tamaño infinito es ilusorio; en realidad, los lotes presentan tamaño finito y las muestras con mayor razón; surge entonces la pregunta, ¿de qué tamaño debe ser la muestra para suponer que la media muestral de una población con distribución diferente a la normal es normal? La respuesta desde luego depende de la distribución real que presente. Algunas reglas empíricas o "de dedo", establecen experimentalmente lo siguiente:

i. Si la distribución de probabilidad de una población es parecida a una normal, es decir, si la gráfica de la función de probabilidad de una población presenta una forma parecida al perfil de una campana casi simétrica, basta tomar $n \geq 4$.

ii. Si la distribución de probabilidad de una población es parecida a una distribución uniforme, basta tomar $n \geq 12$.

iii. Si la distribución de probabilidad de una población presenta la mayor parte de sus medidas en alguno de sus extremos (cóncava hacia arriba y no simétrica), por ejemplo, la distribución exponencial negativa, que se podría denominar comportamiento antinormal, $n \geq 100$.

Distribución de Probabilidad Ji Cuadrada

Sean $z_1, z_2, \cdots, z_k$, variables aleatorias normales estándar, estadísticamente independientes, es decir,

$$z_i \sim N(0,1) \qquad i = 1, 2, \cdots, k$$

Sea

$$\chi^2 = z_1^2 + z_2^2 + \cdots + z_k^2 \tag{1.65}$$

Cabe señalar que en la expresión anterior χ^2 (ji cuadrada) representa a una variable, no es que esté elevada al cuadrado, es solo notación, se indica así solo para recordar que se trata de una suma de cuadrados.

La función de probabilidad de esta variable aleatoria ji cuadrada es:

$$f(\chi^2) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}(\chi^2)^{\frac{k}{2}-1} e^{-\frac{\chi^2}{2}} \qquad \chi^2 > 0 \tag{1.66}$$

Su función generadora de momentos se expresa como:

$$FGM_{\chi 2}(t) = (1-2t)^{\frac{-k}{2}} \qquad t < 1/2 \tag{1.67}$$

Su función característica es:

$$FC_{\chi 2}(t) = (1-2it)^{\frac{-k}{2}} \qquad t < 1/2 \tag{1.68}$$

La media y la varianza de la función de probabilidad ji cuadrada son:

$$\mu_{\chi 2} = k \tag{1.69}$$

$$\sigma^2_{\chi 2} = 2k \tag{1.70}$$

Su coeficiente de asimetría y su coeficiente de curtosis son:

$$\gamma_{1\chi^2} = \sqrt{\frac{8}{k}} \tag{1.71}$$

$$\gamma_{2\chi^2} = \frac{12}{k} \tag{1.72}$$

Teorema de Aditividad de la Distribución Ji Cuadrada

Sean

$$\chi^2_{1,k_1}, \chi^2_{2,k_2}, \cdots, \chi^2_{p,k_p}$$

p variables aleatorias con distribución ji cuadrada, estadísticamente independientes con $k_1, k_2, \cdots, k_p$ grados de libertad respectivamente. Entonces, la suma

$$y = \chi^2_{1,k_1} + \chi^2_{2,k_2} + \cdots + \chi^2_{p,k_p} \tag{1.73}$$

También presenta distribución ji cuadrada con $k = k_1 + k_2 + \cdots + k_p$ grados de libertad.

$$y \sim \chi^2_k$$

$$k = k_1 + k_2 + \cdots + k_p$$

La variable aleatoria que se muestra a continuación presenta distribución ji cuadrada con $n-1$ grados de libertad.

$$\frac{(n-1)S^2_{n-1}}{\sigma^2_x} \sim \chi^2_{n-1} \tag{1.74}$$

Distribución de Probabilidad t de Student

Sean $z \sim N(0,1)$, una variable aleatoria normal estándar, y $v \sim \chi^2_k$ una variable aleatoria ji cuadrada con k grados de libertad, ambas estadísticamente independientes. Sea la variable aleatoria

$$t = \frac{Z}{\sqrt{\frac{V}{k}}} \tag{1.75}$$

A la variable aleatoria "t" se le denomina t de Student, por el pseudónimo con que firmaba su autor William Sealy Gosset (1876-1937). La función de densidad de probabilidad de la variable aleatoria t de Student o de Gosset está dada por la expresión:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \quad \frac{1}{\left[\frac{t^2}{k}+1\right]^{\frac{(k+1)}{2}}} \quad t \in \Re \tag{1.76}$$

La media y la varianza de la función de probabilidad t de Student son:

$$\mu_t = 0 \qquad k > 1 \tag{1.77}$$

$$\sigma_t^2 = \frac{k}{k-2} \qquad k > 2 \tag{1.78}$$

Su coeficiente de asimetría y su coeficiente de curtosis son:

$$\gamma_{1t} = 0 \qquad k > 3 \tag{1.79}$$

$$\gamma_{2t} = \frac{6}{k-4} \qquad k > 4 \tag{1.80}$$

La variable aleatoria

$$\frac{\bar{x} - \mu_x}{\frac{S_{n-1}}{\sqrt{n}}} \approx t_{n-1} \tag{1.81}$$

Presenta distribución t de Student con $(n-1)$ grados de libertad, y para valores de $n > 30$ su función de probabilidad es aproximadamente normal, con media cero y varianza $n/(n-2)$.

Distribución de Probabilidad F de Fisher - Snedecor

Sean $u \sim \chi^2_{k_1}$ y $v \sim \chi^2_{k_1}$ dos variables aleatorias ji cuadrada con k_1 y k_2 grados de libertad respectivamente, ambas estadísticamente independientes. Sea la variable aleatoria

$$F = \frac{u/k_1}{v/k_2} \tag{1.82}$$

A la variable aleatoria "F" se le denomina F de Fisher-Snedecor, por los dos principales científicos que se dedicaron a analizarla: Sir Ronald Aylmer Fisher (1890-1962), ya citado en la página 11 de este volumen y George Waddel Snedecor (1881-1974).

La función de densidad de probabilidad de la variable aleatoria F de Fisher-Snedecor está dada por la expresión:

$$f(F) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \frac{F^{\frac{k_1}{2}-1}}{\left[\frac{k_1}{k_2}F+1\right]^{\frac{(k_1+k_2)}{2}}} \tag{1.83}$$

La cual sigue la distribución F con k_1 grados de libertad en el numerador y k_2 grados de libertad en el denominador, lo que suele abreviarse como F_{k_1, k_2}.

La función de probabilidad acumulada F de Fisher Snedecor está dada por la expresión:

$$F(f) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \int_0^f \frac{u^{\frac{k_1}{2}-1}}{\left[\frac{k_1}{k_2}u+1\right]^{\frac{(k_1+k_2)}{2}}} du \tag{1.84}$$

La media y la varianza de la función de probabilidad F de Fisher Snedecor están dadas por las expresiones:

$$\mu_F = \frac{k_2}{k_2 - 2} \qquad k_2 > 2 \tag{1.85}$$

$$\sigma_F^2 = \frac{2k_2^2\left(k_1 + k_2 - 2\right)}{k_1\left(k_2 - 2\right)^2\left(k_2 - 4\right)} \qquad k_2 > 4 \tag{1.86}$$

Suponga que se tienen dos variables aleatorias normales, estadísticamente independientes, $x_1 \sim N(\mu_1, \sigma_1)$ y $x_2 \sim N(\mu_1, \sigma_1)$. Si de cada una de ellas se obtiene una muestra de tamaño n_1 y n_2, respectivamente y se calcula su desviación estándar muestral para cada una de ellas, entonces

$$\frac{\left(n_1 - 1\right) S_{n_1 - 1}^2}{\sigma_{x_1}^2} \sim \chi_{n_1 - 1}^2$$

$$\frac{\left(n_2 - 1\right) S_{n_2 - 1}^2}{\sigma_{x_2}^2} \sim \chi_{n_2 - 1}^2$$

Lo que implica que el cociente de estas dos variables aleatorias entre sus grados de libertad es una variable aleatoria F con n_1–1 grados de libertad en el numerador y n_2–1 grados de libertad en el denominador, es decir

$$\frac{\dfrac{S_{n_1 - 1}^2}{\sigma_1^2}}{\dfrac{S_{n_2 - 1}^2}{\sigma_2^2}} \sim F_{n_1 - 1,\ n_2 - 1} \tag{1.87}$$

1.5. Generación de números aleatorios con cierta distribución de probabilidad

Suponga que se tiene una variable x con distribución uniforme continua en el intervalo $[a, b]$. Su función de probabilidad acumulada está dada por la expresión:

$$F(x) = \frac{x-a}{b-a} \quad \forall x \in [a,b]$$

Donde, por definición $0 \leq F(x) \leq 1$, suponga que se hace $u = F(x)$ y se despeja la variable aleatoria x, es decir

$$x = a + u(b-a) \tag{1.88}$$

Si a través de una tabla de números aleatorios o utilizando alguno de los paquetes estadísticos se genera un número aleatorio u, entre cero y uno, y se aplica el algoritmo de la expresión (1.88), se obtiene un número aleatorio con distribución uniforme entre a y b.

Ejemplo 1.1

Para ilustrar la forma en que se generan números aleatorios con distribución continua uniforme, suponga que una persona llega a clases siguiendo una distribución uniforme entre las 6:50 y las 7:20 horas. Esto implica, que esta persona nunca falta a clases y siempre llega en ese intervalo de tiempo. Simule diez llegadas de esta persona al salón de clases.

Cabe señalar que las 6:50 es 6 + 50/60 = 6.833333 horas y 7:20 es 7 + 20/60 = 7.333333 horas. El algoritmo en Excel sería

x = 6.833333 + (7.333333–6.833333) * aleatorio()

Al correr este algoritmo en Excel se obtienen los siguientes valores:

No.	x (horas)	x(h:min:seg)
1	6.85937365	06:51:33
2	6.91491385	06:54:53
3	7.24435022	07:14:39
4	7.25526591	07:15:18
5	7.2002653	07:12:00
6	6.89620187	06:53:46
7	7.22232478	07:13:20
8	7.12237353	07:07:20
9	7.12943694	07:07:45
10	7.07990582	07:04:47

El amable lector que corra este algoritmo usando Excel no obtendrá los mismos valores, precisamente por la pseudoaleatoriedad en la generación de u. Se dice pseudoaleatoridad porque la generación no es totalmente aleatoria, se utiliza un algoritmo matemático en Excel para generar u = aleatorio(), ¿podría el amable lector investigar en Excel qué algoritmo se utiliza? Más adelante en este mismo volumen se verá cómo se puede probar estadísticamente que estos datos presentan distribución uniforme continua entre 6:50 y 7:20 horas.

Otra forma de generar números aleatorios con distribución uniforme entre $a = 6.833333$ y $b = 7.333333$ es usar Minitab, para ello en el menú principal se da click en Calc, luego en Random Data y finalmente en Uniform, apareciendo la ventana mostrada en la figura 1.11, en la cual se teclean los valores de n, número de números aleatorios a generar; a, límite inferior del intervalo de uniformidad; y b, límite superior del intervalo de uniformidad y se oprime Ok, obteniéndose los siguientes resultados:

x-uniforme: 6.98834, 6.86239, 6.88003, 7.22875, 6.85801, 6.94142, 6.87952, 7.18332, 6.94536, 6.97879

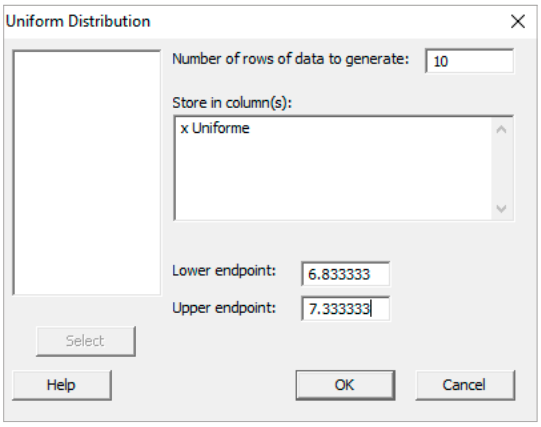


Figura 1.11

Otra forma de generar números aleatorios con distribución uniforme entre a = 6.833333 y b = 7.333333 es, usando el software R, con el siguiente comando:

x = runif(n, min = a, max = b)

En este caso

x = runif(10, min = 6.833333, max = 7.333333)

Al correr este comando se obtiene lo siguiente: 7.126656 6.992034 7.138805 7.107215 7.305826 7.275338 7.206063 7.017805 7.194726 7.215625

Ahora se generarán números aleatorios con distribución exponencial negativa. La función de probabilidad acumulada de la exponencial negativa está dada por la expresión:

$$F(x) = 1 - \exp(-\lambda x)$$

Por lo que si se hace nuevamente $F(x) = u$ y se despeja x:

$$x = \frac{1}{\lambda} Ln\left(\frac{1}{1-u}\right) \tag{1.89}$$

Ejemplo 1.2

Suponga que el tiempo de vida de una lámpara de neón se comporta con distribución exponencial negativa con una media de 2000 horas de uso. Simule el tiempo de vida de diez lámparas de neón.

La media de la distribución exponencial negativa es $1/\lambda = 2000$, por lo que corriendo el algoritmo en Excel se obtienen los siguientes resultados:

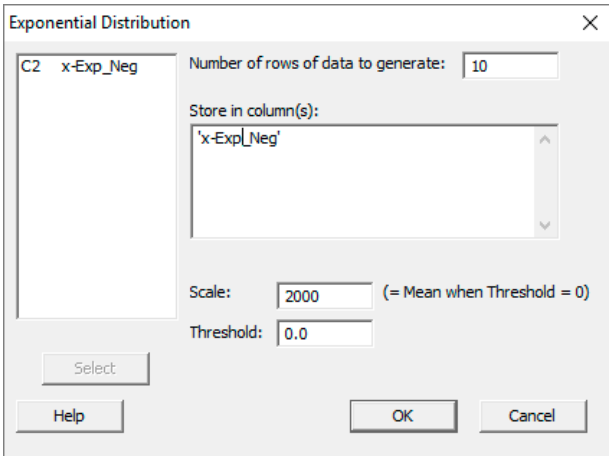
No.	x
1	422.471
2	3704.77
3	255.424
4	3196.09
5	1457.89
6	60.8494
7	4065.14
8	950.178
9	1390.43
10	1953.54

En la anterior simulación nótese que la lámpara 6 solo duró un poco menos de 61 horas de uso, en cambio la lámpara 7 duró un poco más de 4065 horas; más del doble de la media.

Otra forma de generar números aleatorios con distribución exponencial con una media de $1/\lambda = 2000$ es usar Minitab; para ello, en el menú principal se da click en Calc, luego en Random Data y finalmente en Exponential, apareciendo la ventana mostrada en la figura 1.12, en la cual se teclean los valores de n = 10, número de números aleatorios a generar; $1/\lambda = 2000$ y se oprime Ok, obteniéndose los siguientes resultados:

x-Exp_Neg: 2611.72, 4495.67, 42.08, 460.44, 0.45, 1149.06, 722.82, 2235.93, 291.65, 2848.27

Figura 1.12



Otra forma de generar números aleatorios con distribución exponencial negativa con una media de $1/\lambda = 2000$ es, usando el software R, con el siguiente comando:

x = rexp(n, rate = tasa promedio de ocurrencias)

En este caso

x = rexp(10, rate = 0.0005)

Al correr este comando se obtiene lo siguiente: rexp(10, rate = 0.0005): 2165.2933, 1288.2904, 2553.5096, 488.2835, 5615.4145, 1656.1038 651.1113, 711.9925, 2749.3669, 1899.1402

Ejemplo 1.3

La tasa de interés de una sucursal bancaria se comporta con distribución triangular de parámetros optimista a = 0.20 o 20%, el más probable b = 0.35 o 35% y pesimista c = 0.60 o 60%. Genere 20 números aleatorios con dicha distribución triangular.

La función de probabilidad acumulada de la distribución triangular se muestra en el cuadro resumen de la figura 1.9, su expresión matemática es

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{(x-a)^2}{(c-a)(b-a)} & a \le x \le b \\ 1 - \dfrac{(c-x)^2}{(c-a)(c-b)} & b < x \le c \\ 1 & x > c \end{cases}$$

En donde el punto que separa a las dos ramas de la distribución es $x = b$. En este punto, la probabilidad está dada por:

$$F\left(x = b\right) = \frac{(b-a)^2}{(c-a)(b-a)} = \frac{(b-a)}{(c-a)}$$

A partir de lo cual, si se denomina u = F(x), se puede establecer que si

$$u \le (b-a)/(c-a)$$

Se usa la primera rama y si ocurre lo contrario se usa la segunda rama de la función, de tal manera que si se despeja x de ambas ramas de la función de probabilidad acumulada, se obtiene el siguiente algoritmo:

$$x = \begin{cases} a + \sqrt{(b-a)(c-a)u} & u \leq \dfrac{(b-a)}{(c-a)} \\ c - \sqrt{(c-b)(c-a)(1-u)} & u > \dfrac{(b-a)}{(c-a)} \end{cases} \tag{1.90}$$

Al sustituir los valores de a, b y c, y generar números aleatorios uniformes entre cero y uno, el algoritmo para generar números aleatorios con distribución triangular con Excel sería:

$$x = \begin{cases} 0.2 + \sqrt{0.006 * aleatorio\ ()} & x \leq 0.375 \\ 0.6 - \sqrt{0.01 * aleatorio\ ()} & x > 0.375 \end{cases}$$

Al aplicar este algoritmo se generan los siguientes números aleatorios:

u	x
0.972186	0.583323
0.553475	0.533178
0.778643	0.552951
0.171581	0.232086
0.913018	0.570507
0.103112	0.224873
0.183973	0.233224
0.706892	0.545861
0.891987	0.567135
0.043555	0.216166

Otra forma de generar números aleatorios con distribución triangular con valores $a = 20\%$, $b = 35\%$ y $c = 60\%$, es usar Minitab; para ello, en el menú principal se da click en Calc, luego en Random Data y finalmente en Triangular, apareciendo la ventana mostrada en la figura 1.13, en la cual se teclean los valores de $n = 10$, valor mínimo a, valor más probable b y valor máximo c, y se oprime Ok, obteniéndose los siguientes resultados:

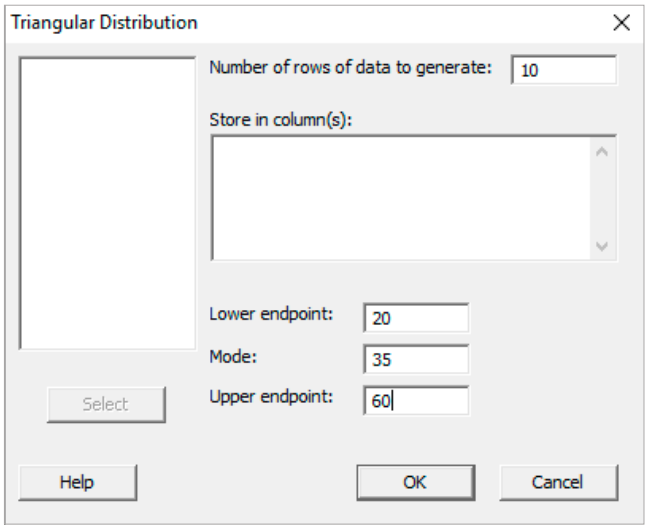


Figura 1.13

x-Triangular: 34.0136, 41.9168, 38.9889, 41.0404, 29.3552, 44.9505, 49.2064, 35.1221, 36.4790, 27.2066

En la mayoría de las distribuciones de probabilidad acumulada es prácticamente imposible despejar el valor de x a partir del número u; los algoritmos que se usan son diferentes dependiendo de la función de probabilidad que se pretenda generar. Por ejemplo, la función de probabilidad más común de todas es la normal, el tratar de obtener el valor de x a partir de un valor de probabilidad $F(x) = u$, equivaldría a despejar el valor de x de la siguiente ecuación integral:

$$\frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu_x}{\sigma_x}\right)^2} dt = u$$

Intentar despejar analíticamente a la variable x en la expresión anterior es prácticamente imposible.

Una forma de generar números aleatorios con distribución normal es aplicando el teorema del límite central, el cual establece que si se suman un número suficiente de variables aleatorias con cualquier distribución, dicha suma tendrá distribución aproximadamente normal si n es grande. Un caso particular sería sumar variables aleatorias con distribución uniforme, experimentalmente se puede comprobar que basta sumar $n >= 12$ variables aleatorias uniformes para obtener una distribución normal.

Recuérdese, de conformidad con el cuadro resumen de la figura 1.9, que si x es uniforme en el intervalo $[0, 1]$ su media es 0.5 y su varianza 1/12, esto implica que si se generan 12 números aleatorios con distribución uniforme entre cero y uno, y se suman estos, se obtiene una variable aleatoria aproximadamente normal con media seis y varianza uno, la cual puede ser estandarizada de la siguiente forma

$$z = \frac{[x_1 + x_2 + \cdots + x_{12}] - 6}{1}$$

y convertida a otra distribución normal con media μ_x y desviación estándar σ_x, a través del proceso contrario a la estandarización

$$x = \mu_x + z\sigma_x$$

en donde z es una variable aleatoria normal estándar.

Ejemplo 1.4

Genere una muestra aleatoria de tamaño $n = 64$, de números aleatorios con distribución normal de media cinco y desviación estándar tres, utilizando el Teorema del Límite Central.

Lo primero que se hará, será generar $m = 12$ muestras de tamaño $n = 64$ de variables aleatorias uniformes, luego se obtendrá la suma de estas 12 variables aleatorias uniformes. De conformidad con el Teorema del Límite Central, la suma será aproximadamente normal con media 6 y varianza 1, después se estandarizará esta última y por último se calculará una variable aleatoria normal con media cinco y desviación estándar tres. Se obtiene la siguiente tabla usando Excel:

x_1 Uniform	x_2 Uniform	x_3 Uniform	x_4 Uniform	x_5 Uniform	x_6 Uniform	x_7 Uniform	x_8 Uniform	x_9 Uniform	x_{10} Uniform	x_{11} Uniform	x_{12} Uniform	y (suma)	z (estándar)	x (Normal)
0.223710	0.747615	0.521265	0.513036	0.093331	0.714209	0.007002	0.681426	0.674932	0.535092	0.916968	0.047063	5.675649	-0.324351	4.02694729
0.577146	0.492744	0.162713	0.428703	0.719717	0.843791	0.880261	0.887069	0.288033	0.453611	0.745099	0.265253	6.744139	0.744139	7.23241724
0.511237	0.334609	0.464492	0.440322	0.899965	0.951841	0.655498	0.346495	0.368991	0.498950	0.780321	0.075181	6.327901	0.327901	5.98370439
0.794634	0.031698	0.826644	0.856202	0.218845	0.094821	0.051726	0.653762	0.867285	0.299704	0.969599	0.023293	5.688212	-0.311788	4.06463521
0.685423	0.459354	0.740213	0.768305	0.910893	0.075730	0.535767	0.054055	0.685451	0.692249	0.368372	0.762506	6.738318	0.738318	7.21495454
0.855749	0.975965	0.104126	0.312212	0.434144	0.984712	0.800268	0.700577	0.062350	0.665327	0.623503	0.518954	7.037887	1.037887	8.11366137
0.260006	0.684121	0.089779	0.535715	0.037441	0.318876	0.891632	0.978476	0.631502	0.037448	0.635383	0.719650	5.820029	-0.179971	4.46008672
0.483489	0.384344	0.810750	0.101469	0.872961	0.706363	0.907932	0.661445	0.399693	0.482334	0.635235	0.344640	6.790656	0.790656	7.37196745
0.140728	0.798351	0.980142	0.001595	0.803106	0.266012	0.514065	0.173704	0.004334	0.010908	0.121388	0.192973	4.007305	-1.992695	-0.97808495
0.807150	0.639387	0.146503	0.904797	0.503667	0.949319	0.376278	0.114943	0.093700	0.617384	0.685549	0.826305	6.664981	0.664981	6.99494287
0.171113	0.390133	0.426485	0.559015	0.049128	0.839002	0.699214	0.204503	0.888407	0.273564	0.757828	0.098711	5.357104	-0.642896	3.07131077
0.259105	0.480644	0.779997	0.767064	0.977177	0.308411	0.142126	0.327437	0.248438	0.197566	0.491093	0.667595	5.646653	-0.353347	3.93995821
0.854785	0.349147	0.746376	0.504825	0.555539	0.246437	0.449415	0.590265	0.579423	0.901949	0.377144	0.480578	6.635883	0.635883	6.90764775
0.066142	0.630965	0.231596	0.813287	0.266961	0.152371	0.124155	0.561336	0.746299	0.644377	0.069121	0.804504	5.111113	-0.888887	2.33334012
0.032942	0.022027	0.121524	0.045810	0.880587	0.262862	0.521950	0.130350	0.698087	0.260288	0.556153	0.943673	4.476253	-1.523747	0.42875974
0.249801	0.517715	0.057766	0.492886	0.921572	0.604837	0.234976	0.995251	0.111371	0.744254	0.845868	0.508411	6.284708	0.284708	5.85412275
0.008874	0.007762	0.123526	0.785295	0.665990	0.450206	0.970446	0.012914	0.843763	0.845842	0.689446	0.617725	6.021790	0.021790	5.06536881
0.942390	0.485214	0.404149	0.565020	0.903565	0.190730	0.482545	0.737304	0.138280	0.414040	0.190061	0.841858	6.295158	0.295158	5.88547534
0.753381	0.013273	0.120503	0.397747	0.790502	0.773334	0.085931	0.971620	0.388507	0.515659	0.981768	0.970368	6.762594	0.762594	7.28778101
0.660630	0.833383	0.396719	0.900312	0.435439	0.367585	0.786341	0.908077	0.187465	0.700570	0.665059	0.809549	7.651130	1.651130	9.95338892
0.573740	0.802954	0.602465	0.490824	0.254383	0.410352	0.602506	0.143136	0.245818	0.362145	0.415875	0.098228	5.002426	-0.997574	2.00727944
0.193264	0.061427	0.186110	0.087837	0.189828	0.354543	0.998043	0.040446	0.645085	0.696934	0.009266	0.596017	4.058802	-1.941198	-0.82359513
0.700205	0.771318	0.493291	0.351368	0.397859	0.361757	0.927139	0.104994	0.317903	0.536511	0.773024	0.801744	6.537111	0.537111	6.61133418
0.128516	0.947170	0.944535	0.693575	0.784715	0.854318	0.655413	0.603536	0.126990	0.736543	0.644109	0.809816	7.929236	1.929236	10.7877082
0.823877	0.186004	0.710210	0.290826	0.716804	0.836288	0.705530	0.040449	0.867309	0.202281	0.791587	0.931634	7.102798	1.102798	8.30839294
0.398489	0.003951	0.362049	0.971855	0.062648	0.641223	0.701854	0.930848	0.045037	0.981329	0.864863	0.879954	6.844101	0.844101	7.53230214
0.777406	0.311576	0.056333	0.003544	0.130622	0.999054	0.803133	0.290252	0.433499	0.694951	0.319468	0.312756	5.132593	-0.867407	2.39777956
0.512104	0.779617	0.254196	0.810981	0.761160	0.258559	0.700704	0.077380	0.385643	0.588307	0.621917	0.306663	6.057232	0.057232	5.17169466
0.209956	0.498021	0.947310	0.564942	0.883599	0.813608	0.324958	0.372241	0.267735	0.111887	0.108794	0.708851	5.811902	-0.188098	4.43570561
0.550449	0.428776	0.755604	0.770357	0.403243	0.673120	0.442595	0.182171	0.411051	0.952399	0.472786	0.578726	6.621277	0.621277	6.86383035
0.481666	0.109504	0.893746	0.010813	0.002890	0.578273	0.723918	0.027166	0.140993	0.016388	0.847505	0.340965	4.173826	-1.826174	-0.47852201
0.633119	0.772040	0.760880	0.037201	0.566615	0.582631	0.762279	0.341466	0.036926	0.949348	0.955815	0.985040	7.383359	1.383359	9.1500768
0.503454	0.779018	0.992477	0.501086	0.191616	0.923519	0.389604	0.036143	0.962069	0.225423	0.946448	0.761962	7.212819	1.212819	8.6384574
0.429931	0.135809	0.390555	0.205951	0.260047	0.373249	0.300031	0.055795	0.763094	0.622502	0.284628	0.579047	4.400640	-1.599360	0.20191893
0.094088	0.855793	0.538357	0.144241	0.210460	0.638058	0.706950	0.490150	0.832762	0.761246	0.475928	0.763844	6.511877	0.511877	6.53563027
0.190939	0.172655	0.572204	0.523780	0.750432	0.409212	0.156098	0.689448	0.592450	0.837199	0.300449	0.155569	5.350433	-0.649567	3.0512994
0.924158	0.297281	0.577100	0.371196	0.769945	0.632876	0.893440	0.358217	0.340387	0.421468	0.753900	0.406255	6.746222	0.746222	7.23866646
0.737450	0.431367	0.976626	0.183790	0.811559	0.489769	0.826091	0.512078	0.224286	0.346293	0.215829	0.462973	6.218112	0.218112	5.6543347
0.806681	0.241683	0.057632	0.625897	0.956128	0.111827	0.677092	0.123610	0.391303	0.667633	0.188035	0.180990	5.028510	-0.971490	2.08553098
0.598073	0.849777	0.909092	0.160383	0.616221	0.992068	0.126003	0.650993	0.472117	0.731694	0.138591	0.584645	6.829657	0.829657	7.48897216
0.067083	0.630648	0.458185	0.333288	0.428609	0.178578	0.937319	0.805272	0.890453	0.570814	0.558330	0.216379	6.074958	0.074958	5.22487292
0.705912	0.370722	0.802044	0.483844	0.409884	0.746428	0.251676	0.468221	0.635561	0.772076	0.175796	0.983844	6.806008	0.806008	7.41802367
0.202538	0.107180	0.135534	0.204977	0.739982	0.247277	0.868274	0.554983	0.590387	0.973168	0.544794	0.085799	5.254891	-0.745109	2.76467162
0.590605	0.567538	0.178435	0.592908	0.671709	0.507059	0.544012	0.414411	0.377044	0.145716	0.729264	0.119822	5.438524	-0.561476	3.31557108
0.783441	0.527062	0.764983	0.020886	0.467012	0.357882	0.888209	0.796753	0.668266	0.099187	0.563452	0.521160	6.458294	0.458294	6.37488222
0.249899	0.940316	0.558816	0.949590	0.944331	0.421034	0.411957	0.238941	0.706629	0.950452	0.739659	0.751523	7.863147	1.863147	10.5894416
0.073565	0.805508	0.069704	0.125106	0.295828	0.210722	0.431898	0.230473	0.287834	0.151982	0.861813	0.466019	4.010451	-1.989549	-0.96864585
0.415724	0.370007	0.145240	0.241385	0.470119	0.942667	0.608932	0.378343	0.755655	0.941284	0.109305	0.942986	6.321647	0.321647	5.96494025
0.583488	0.772052	0.339847	0.414988	0.943225	0.617097	0.232938	0.210159	0.130343	0.576876	0.091793	0.577572	5.490379	-0.509621	3.47113739
0.618456	0.014223	0.146294	0.050177	0.760133	0.844383	0.085701	0.598478	0.265018	0.680898	0.478338	0.364051	4.906149	-1.093851	1.71844725
0.806203	0.110237	0.257586	0.205121	0.599036	0.139137	0.490021	0.449903	0.339868	0.517457	0.039829	0.069876	4.024274	-1.975726	-0.92717913
0.991363	0.456829	0.348783	0.663254	0.540674	0.831688	0.846025	0.015326	0.631448	0.277068	0.635587	0.870845	7.108890	1.108890	8.32667137
0.394007	0.846460	0.672382	0.273652	0.243992	0.013050	0.993165	0.813812	0.486040	0.944978	0.947141	0.852529	7.481207	1.481207	9.44362061
0.697044	0.581650	0.927699	0.668980	0.652267	0.511124	0.230572	0.024737	0.259788	0.875759	0.894087	0.185130	6.508836	0.508836	6.5265077
0.256332	0.376248	0.589386	0.885736	0.175396	0.705293	0.101435	0.073343	0.084249	0.909912	0.760359	0.011117	4.928806	-1.071194	1.78641718
0.478169	0.498704	0.336626	0.864212	0.286614	0.522191	0.635543	0.375237	0.636942	0.980329	0.974030	0.768722	7.357318	1.357318	9.07195389
0.128041	0.809986	0.581262	0.709321	0.407178	0.390075	0.210878	0.359751	0.523084	0.805319	0.756257	0.156976	5.838128	-0.161872	4.51438479
0.535113	0.521288	0.145584	0.864721	0.129095	0.861547	0.909275	0.335058	0.810755	0.555925	0.590469	0.504077	6.762906	0.762906	7.28871871
0.780138	0.985307	0.782815	0.936661	0.833818	0.828046	0.625572	0.123112	0.194278	0.157063	0.236184	0.471827	6.954823	0.954823	7.86446864
0.577518	0.989645	0.279199	0.912987	0.904117	0.138107	0.409857	0.485348	0.017225	0.085207	0.749705	0.768391	6.317307	0.317307	5.95191964
0.201548	0.538569	0.886074	0.072915	0.221266	0.476616	0.564805	0.156642	0.616483	0.823202	0.974636	0.061180	5.593935	-0.406065	3.78180633
0.580466	0.175199	0.849106	0.594267	0.274544	0.953577	0.582265	0.397104	0.565534	0.872176	0.954605	0.741134	7.539978	1.539978	9.61993498
0.598682	0.662085	0.912917	0.889018	0.066331	0.247077	0.366285	0.182882	0.361542	0.767887	0.551663	0.111389	5.717758	-0.282242	4.15327251
0.845874	0.934876	0.447972	0.711639	0.879622	0.429450	0.509343	0.547545	0.523871	0.819140	0.761466	0.684064	8.094862	2.094862	11.2845874

Posteriormente se realizará un ejercicio para demostrar estadísticamente que los datos de la última columna tienen distribución normal de media cinco y desviación estándar tres.

Otra forma de generar una muestra aleatoria de tamaño $n = 64$ con distribución normal de media cinco y desviación estándar tres es usar los comandos de Excel, Minitab y R.

Con Excel se emplea el siguiente comando:

x = INV.NORM(u, mx, sx) = INV.NORM(aleatorio(), 5, 3)

Con Minitab:

En el menú principal se dá un clic en Calc, luego en Random Data y posteriormente en Normal, aparece la siguiente pantalla, en la cual se indica el tamaño de muestra $n = 64$, la media 5 y la desviación estándar 3:

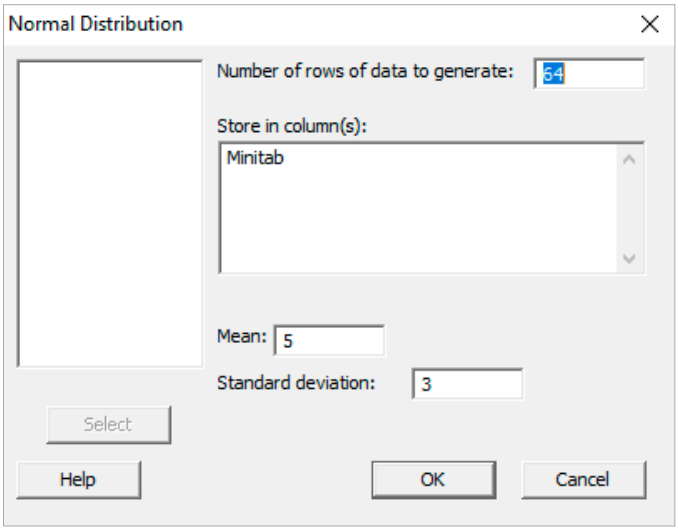


Figura 1.14.

Con R se emplea el siguiente comando:

x = rnorm(n, mean = mx, sd = sx) = rnorm(64, mean = 5, sd = 3)

A continuación, se muestra la tabla de datos que se genera con cada uno de los paquetes. Posteriormente se verá la forma de comprobar que obedecen a una distribución normal de media cinco y desviación estándar tres.

x-Excel	x-Minitab	x-R
10.172120	4.012171	5.6567778
8.454944	5.872629	5.8453375
6.279027	1.582275	9.7942041
9.782763	4.586038	6.9808161
5.654476	-0.301310	6.7734169
8.392653	4.759686	0.3234426
-0.708656	5.827817	0.9815752
-1.189657	4.111725	4.3495158
-0.140899	8.361171	6.9325076
8.860379	3.556573	2.4664819
1.863580	4.051598	7.4323854
14.992950	7.957254	1.250032
11.434928	6.777039	3.8227475
3.930352	5.364007	4.208093
12.897582	4.840448	-0.5420816
3.598927	4.683302	3.1984086
2.404101	5.361169	5.5435517
8.111854	2.345334	6.3863245
4.488167	6.990460	3.5472626
4.908491	3.305609	5.0295526
5.098757	2.503790	6.5700833
5.380160	0.297526	-1.5514694
1.428804	7.968892	3.9578856
2.992182	3.343949	7.3958967
2.286739	5.171342	8.6468156
3.616052	5.169072	3.7857618
-2.179881	0.484478	8.8454949
2.048835	11.408431	2.659504
5.141748	4.855752	3.4676871
8.182314	4.078096	2.9682971
9.268036	5.369194	4.0478835
2.698810	4.264500	7.8072589

x-Excel	x-Minitab	x-R
-1.693858	8.793831	5.2037445
9.452436	1.856745	6.6312612
4.205412	6.044418	4.9718765
9.446556	7.241720	6.56161
5.397275	7.901804	5.0769104
4.164943	4.820668	1.9485869
9.810215	1.809942	7.5207815
3.713610	3.946630	5.373617
6.982175	6.740257	8.76338
2.350786	7.757918	1.4301685
0.122113	5.650301	3.783257
5.047235	2.379217	6.7268403
7.388470	7.019787	9.8172646
6.308737	4.405815	7.0004911
1.415753	4.578532	6.664393
8.384113	5.809924	3.6623784
1.397148	1.039350	6.9051605
6.023888	4.058391	-2.2921416
5.720005	7.104488	5.1326131
1.296556	2.846918	1.8740969
2.930482	5.474448	5.4579116
5.686940	9.922340	-0.294832
6.703411	9.815990	7.8997273
7.079819	2.986096	8.9630046
6.955715	4.210276	8.2576757
0.864595	7.325535	7.4246878
9.455873	9.445169	2.3167671
6.061402	4.726795	5.5177646
5.340963	1.712726	8.1321718
5.796126	7.056700	3.3666113
5.555349	5.599436	5.2725592
2.403236	1.475326	4.8356254

En la figura 1.15 se muestra un cuadro resumen de cómo con diversos paquetes de cómputo se han logrado diseñar e implantar algoritmos computacionales, para generar números aleatorios de variables aleatorias con una función de probabilidad conocida.

Figura 1.15. Algoritmos para generar números aleatorios con cierta distribución de probabilidad

Algoritmos para generar números aleatorios con cierta distribución

Distribución	Excel	Minitab	R
Binomial	INV.BINOM(n, p, aleatorio())	Calc ⇒ Random Data ⇒ Binomial Indicar m, n y p	rbinom(m, n, p)
Geométrica		Calc ⇒ Random Data ⇒ Geometric Indicar m y p	rgeom(m, p)
Binomial negativa		Calc ⇒ Random Data ⇒ Negative Binomial; Indicar m, p y r	rnbinom(m, r, p, mu)
Hipergeométrica		Calc ⇒ Random Data ⇒ Hypergeometric Indicar m, N, D y n	rhyper(m, D, N-D, n)
Poisson		Calc ⇒ Random Data ⇒ Poisson Indicar m y λ	rpois(m, λ)
Uniforme	aleatorio()*(b-a)+a	Calc ⇒ Random Data ⇒ Uniform Indicar m, a y b	runif(m, a, b)
Triangular		Calc ⇒ Random Data ⇒ Triangular Indicar m, a, b y c	
Exponencial	-LN(1-ALEATORIO())/λ	Calc ⇒ Random Data ⇒ Exponencial Indicar m y 1/λ	rexp(m, λ)
Normal	INV.NORM(aleatorio(), μ, σ)	Calc ⇒ Random Data ⇒ Normal Indicar m, μ y σ	rnorm(m, μ, σ)
Lognormal	INV.LOGNORM(aleatorio(), μ, σ)	Calc ⇒ Random Data ⇒ Lognormal Indicar m, μ, σ y trunc	rlnorm(m, μ, σ)
Gamma	INV.GAMMA(aleatorio(), α, β)	Calc ⇒ Random Data ⇒ Gamma Indicar m, r, 1/λ, trunc	rgamma(m, r, λ, 1/λ)
Beta	INV.BETA.N(u, α, β, a, b)	Calc ⇒ Random Data ⇒ Beta	rbeta(n, α, symbol, ncp)
Weibull		Calc ⇒ Random Data ⇒ Weibull Indicar m, r, 1/λ, trunc	rweibull(m, r, λ)
Ji Cuadrada (χ^2)	INV.CHICUAD(aleatorio(), k)	Calc ⇒ Random Data ⇒ Chi Square Indicar m y k	rchisq(m, k, trunc)
t de Student	INV.T(aleatorio(), k)	Calc ⇒ Random Data ⇒ t Indicar m y k	rt(m, k, trunc)
F de Fisher	INV.F(aleatorio(), k1, k2)	Calc ⇒ Random Data ⇒ F Indicar m, k1 y k2	rf(m, k1, k2, trunc)

Ejercicios propuestos del Capítulo 1

1. Defina los conceptos de probabilidad y estadística, establezca las diferencias y similitudes y explique cómo se relacionan.

2. Clasifique cómo se descompone la Estadística y defina cada una de sus componentes.

3. Esquematice algunos de los métodos de investigación científica existentes.

4. ¿En qué etapas del método científico se aplica la Estadística?

5. Indague qué investigador firmaba con el pseudónimo de Student y por qué lo hacía.

6. Proporcione un ensayo sobre la biografía de Ronald Fisher.

7. Establezca las diferencias y similitudes entre las cuatro escuelas de probabilidad citadas en el capítulo I y explique cómo se relacionan.

8. Una compañía que realiza proyectos de ingeniería en cierta zona del país, clasifica las formaciones geológicas de la zona en tres tipos: I, II y III, con una probabilidad de 0.35, 0.40 y 0.25 para los tres tipos de formaciones respectivamente. Se sabe que hay depósitos de agua (almacenamiento freático en el subsuelo) con una probabilidad del 40% en las formaciones del tipo I, del 20% en las formaciones del tipo II y del 30% en las formaciones del tipo III.
   a. ¿Cuál es la probabilidad de que la compañía encuentre agua en esa zona?
   b. Si la compañía no encuentra agua en esa zona, determinar la probabilidad de que exista una formación del tipo II.

9. Considérese una variable aleatoria continua x, cuya función de densidad de probabilidad está dada por la expresión

$$f(x) = k\cos 2x \qquad 0 < x < p/4$$

   a. Obtener el valor de k para que $f(x)$ sea efectivamente una función de probabilidad.
   b. Determinar la función de probabilidad acumulada $F(x)$.
   c. Calcular la media y la desviación estándar de x.
   d. Obtener el coeficiente de asimetría y el coeficiente de curtosis.
   e. Determinar la función generatriz de momentos de x.
   f. Calcular la mediana y la moda.
   g. Obtener las probabilidades de que $x < p/8$, $p/8 < x < p/4$, $x < 1$.

10. Sea u la temperatura ambiente (en °C) y v el tiempo (en minutos), requeridos para que el motor diesel de una camioneta esté listo para ponerla en movimiento. Supóngase que la densidad conjunta de u y v está dada por $f(u,v) = (4u + 2v + 1)/6640$, $0 \le u \le 40$, $0 \le v \le 2$
   a. Obtener la función de probabilidad conjunta acumulada.
   b. Determinar las funciones de probabilidad marginales de u y de v.
   c. Calcular las medias y varianzas de u y de v.
   d. Obtener las funciones de probabilidad condicionales $f(u|v)$ y $f(v|u)$.
   e. Determinar la covarianza de u y v.
   f. ¿Son independientes u y v? Justifique la respuesta sobre bases matemáticas.
   g. Calcular la probabilidad de que en un día seleccionado aleatoriamente la temperatura ambiente sea mayor a 20°C y se requiera por lo menos un minuto para poner en movimiento la camioneta.

11. Enumere los principales modelos probabilísticos discretos, continuos y de muestreo que existen, explique qué significa la variable aleatoria que representan y realice un trazo aproximado de cada uno de ellos, como ejemplo.

12. ¿Qué modelo probabilístico representa al número de artículos defectuosos que se obtienen de una muestra de tamaño n obtenida de un lote con n artículos, de los cuales se sabe que D son defectuosos?, ¿qué modelo probabilístico se usa para aproximar a este modelo y bajo qué condiciones?

13. Un explorador de petróleo perforará una serie de pozos en cierta área para encontrar un pozo productivo. La probabilidad de que tenga éxito en una prueba es 0.2
    a. ¿Cuál es la probabilidad de que el primer pozo productivo sea el tercer pozo perforado?
    b. ¿Cuál es la probabilidad de que el explorador no vaya a encontrar un pozo productivo si solamente puede perforar 10 pozos?
    c. ¿Cuál es la probabilidad de que el tercer encuentro de petróleo ocurra en el quinto pozo que se perfora?

14. Un call center tiene un servicio de consulta por teléfono para la solución de los problemas de sus usuarios. El servicio está disponible de 9:00 a 17:00 horas en días laborables. La experiencia muestra que la variable aleatoria x, el número de llamadas recibidas por día, tiene una distribución de Poisson con una media de 50 llamadas al día, calcular la probabilidad de que en un día dado, la primera llamada del día se reciba:
    a. Antes de las 9:15 horas.
    b. Después de las 10:00 horas, dado que no se recibió llamada antes de las 9:30 horas.

15. ¿Qué modelo probabilístico representa al número de defectos que se obtienen al inspeccionar diez coches y en qué condiciones?

16. ¿Qué modelo probabilístico representa al tiempo de vida de un dispositivo electrónico hasta antes de su falla?

17. Si se suman siete variables aleatorias con distribución exponencial negativa con una media de 20 minutos cada una, ¿qué distribución de probabilidad se obtiene?, ¿cuál es su media y su desviación estándar?

18. Si se suman 40 variables aleatorias t de Student con varianza 3, ¿qué distribución de probabilidad se obtiene?, ¿cuál su media y su desviación estándar?

19. Si se tienen tres variables aleatorias con funciones de probabilidad normales con medias 5, 7 y 9 y desviaciones estándar 2, 4 y 6 respectivamente, y estas se suman, ¿de qué tipo es la función de probabilidad resultante?, ¿cuál es su media y cuál es su desviación estándar?

20. Se debe colocar un cordón de soldadura a 40 piezas. Basado en su experiencia, un soldador sabe que el tiempo promedio requerido para colocar el cordón de soldadura en una de las piezas es de 5 minutos y su desviación estándar es de 3 minutos. El soldador comienza a aplicar los cordones a las 6:00 p.m. y sabe que el valor medio de aplicación de los 40 cordones debe ser de máximo 4.5 minutos, si desea colocar todos los cordones de soldadura antes de las 9:00 p.m. (hora de término de su turno de trabajo). ¿Cuál es la probabilidad de que el soldador termine de colocar los 40 cordones antes de que termine su turno de trabajo?

21. Se estima que el tiempo transcurrido hasta la falla de una pantalla plana LCD, se distribuye exponencialmente con media igual a tres años. Una compañía ofrece hacer válida la garantía en la tienda por el primer año de uso, ¿qué porcentaje de pantallas hará uso efectivo de la garantía?

22. Un ingeniero viaja diariamente desde su domicilio en Cd. Satélite, hasta su oficina en el centro de la Ciudad de México. En promedio el viaje en un sentido le lleva 24 minutos, con una desviación estándar de 3.8 minutos. Suponga que la distribución de los tiempos de viaje es normal.
    a. Si sale de su domicilio a las 8:35 y se sirve café en la oficina de 8:50 a 9:00 ¿Cuál es la probabilidad de que se pierda el café?
    b. Determine la probabilidad de que dos de los tres viajes siguientes tome por lo menos ½ hora.

23. Genere 20 números aleatorios con distribución triangular con $a = 5$, $b = 7$ y $c = 12$.

24. Genere diez números aleatorios con distribución binomial para $n = 20$ y $p = 0.05$.

25. Genere 25 números aleatorios con distribución de Poisson con una media de tres por cada mil metros, que simulen el número de roturas que presenta un cable de 25000 metros de longitud.

26. Genere 20 números aleatorios con distribución normal con una media de 25 y una desviación estándar de tres.

27. Genere 50 números aleatorios con distribución beta.

28. Genere 30 números aleatorios con distribución gamma.

2. Teoría del Muestreo

2.1. Tipos de Muestreo

Tal como ya se dijo previamente, el Muestreo es la rama de la Estadística que se encarga de definir las reglas para tomar muestras de una población específica, establecer el tamaño de dichas muestras y determinar los parámetros que indicarán la representatividad de estas.

Ya se mencionó que al análisis estadístico de todas las unidades muestrales de una población bajo estudio se le denomina censo; sin embargo, en la gran mayoría de las ocasiones es mejor realizar un muestreo en vez de un censo.

¿Cuál es el propósito del muestreo?

Conocer algún o algunos parámetros de la población a partir de una o varias muestras y tomar decisiones con respecto a ella. Se puede pensar que si lo que se requiere es conocer a una población, lo más adecuado sería realizar un censo, pero como se explicará a continuación, un muestreo nos ahorra tiempo, dinero y esfuerzo.

El muestreo es muy útil en las situaciones siguientes:

1. Cuando la prueba es destructiva.
2. Cuando es muy alto el costo de un censo.
3. Cuando un censo no es tecnológicamente factible, o cuando se necesitaría tanto tiempo que la planeación se vería afectada seriamente.
4. Cuando hay que seleccionar muchas unidades muestrales y la tasa de errores de medición es suficientemente alta, para que un censo pudiera dejar pasar un mayor porcentaje de errores que en el caso de un plan de muestreo.
5. Cuando la población bajo estudio es homogénea y basta con tomar una muestra para conocerla.

Ventajas y desventajas del muestreo sobre el censo

Cuando se compara el muestreo con un censo, el primero tiene las ventajas siguientes:

1. Por lo general es menos costoso, pues requiere menos medición.
2. Hay un menor manejo de los elementos de la población y por tanto se reducen sus posibles daños.
3. Es el único que puede aplicarse en el caso de pruebas destructivas.
4. Hay menos personal implicado en las actividades de medición.
5. A menudo reduce notablemente la cantidad de errores de medición.

El muestreo, sin embargo, tiene también varias desventajas; entre ellas están las siguientes:

1. Si el objetivo del muestreo es tomar una decisión sobre uno o varios parámetros de una población, dado que no se cuenta con el análisis completo de la población, se pueden cometer uno de dos posibles errores:

    a. Estar rechazando una alternativa cuando en realidad debiera aceptarse. A este tipo de errores se le conoce como error tipo I y a la probabilidad de cometerlo se le conoce como a, es decir, p(error tipo I) = a.
    b. Estar aceptando una alternativa cuando en realidad debiera rechazarse. A este tipo de errores se le conoce como error tipo II y a la probabilidad de cometerlo se le conoce como b, es decir p(error tipo II) = b.

2. Se genera normalmente menos información sobre la población.
3. El muestreo de aceptación necesita planeación y documentación del procedimiento de muestreo, mientras que una inspección al 100% no lo requiere.

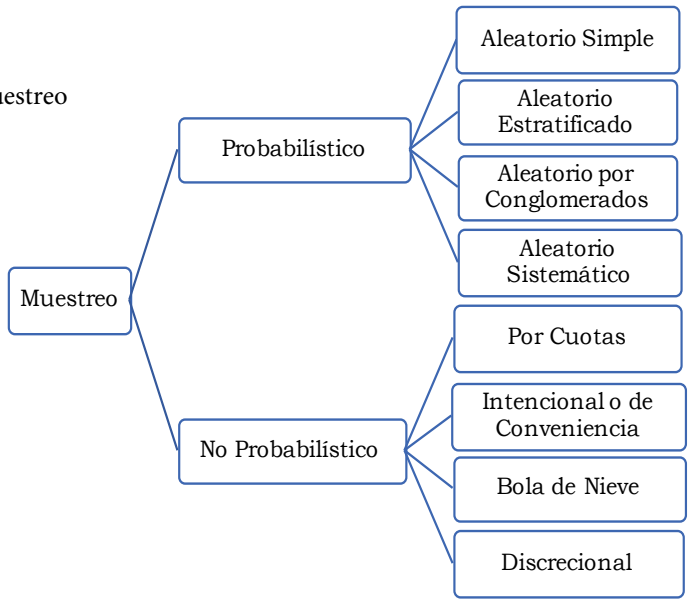
Existen diferentes criterios de clasificación de los diversos tipos de muestreo, aunque en general pueden dividirse en dos grandes grupos: métodos de muestreo probabilísticos y métodos de muestreo no probabilísticos, como se muestra en la figura 2.1

El muestreo probabilístico es una técnica de muestreo, ya que las muestras son recogidas a través de un proceso que brinda a todos los individuos de la población, las mismas oportunidades de ser seleccionados. En esta técnica de

muestreo, el responsable del estudio debe garantizar que cada individuo tenga las mismas oportunidades de ser seleccionado, lo cual se logra si el investigador utiliza la aleatorización.

La ventaja de utilizar una muestra aleatoria es la ausencia de sesgos de muestreo y sistemáticos. Si la selección aleatoria se hace correctamente, la muestra será representativa de toda la población. El efecto de esto es un sesgo sistemático ausente o mínimo que es la diferencia entre los resultados de la muestra y los resultados de la población. El sesgo de muestreo también se elimina ya que los sujetos son elegidos al azar.

Figura 2.1.
Tipos de Muestreo



La primera finalidad del muestreo es obtener muestras representativas de la población bajo estudio. Una muestra es representativa si es obtenida aleatoriamente. Se dice que el Muestreo es Aleatorio si cumple las siguientes cualidades:

» Todos los posibles resultados del experimento deben tener la misma posibilidad de ocurrir.
» Los resultados deben ser independientes entre sí.

El experimentador controla la cantidad de información contenida en la muestra por medio del número *n* de unidades muestrales que se incluyen en esta y por el método usado para seleccionar los datos.

¿Cómo se puede determinar cuál procedimiento usar y el número de elementos a elegir de la muestra? La respuesta depende de dos factores: ¿qué tanta representatividad se desea? y ¿qué tan seguro se requiere estar de esta representatividad?, es decir

1. Si u es la variable de interés y $\hat{u}$ es un estimador de u, entonces se debe especificar un límite superior para el error de estimación, esto es

$$|u-\hat{u}| < \varepsilon \tag{2.1}$$

La variable de interés u a estudiar puede ser la media de la población μ, el total poblacional τ, la fracción de interés p o cualquier otro parámetro poblacional de interés.

La variable $\hat{u}$ representa un estimador puntual del parámetro poblacional u. Para el caso de la media poblacional, un estimador puntual podría ser la media, la mediana m_e, o la moda m_o muestrales, según cuál sea más representativa de la tendencia central de los datos.

Nótese que la desigualdad, 2.1 también puede ser escrita como:

$$u-\varepsilon < \hat{u} < u+\varepsilon \tag{2.2}$$

2. Se debe fijar la probabilidad de que efectivamente el error de estimación sea menor de ε, esto es, la fracción de las veces en que el muestreo tiene como error de estimación un valor menor a ε

$$p\left[\text{Error de Estimación} < \varepsilon\right] = 1-\alpha \tag{2.3}$$

Nótese que la probabilidad $p\left[\text{Error de Estimación} < \varepsilon\right] = 1-\alpha$ también puede ser escrita como:

$$p\left[\,|u-\hat{u}| < \varepsilon\right] = 1-\alpha$$

$$p\left(u-\varepsilon < \hat{u} < u+\varepsilon\right) = 1-\alpha \tag{2.4}$$

Generalmente el criterio que se adopta para fijar un valor al límite superior del error de estimación es definirlo como un múltiplo de la desviación estándar del estimador $\hat{u}$, es decir,

$$\varepsilon = k\sigma_{\hat{u}} \tag{2.5}$$

En donde k depende de la función de probabilidad que tenga el estimador $\hat{u}$ y del nivel de confianza $(1 - \alpha)$ que se desee tener. Si la función de probabilidad de $\hat{u}$ es normal $k = z_{\alpha/2}$, en donde z representa a una variable aleatoria normal.

Algunas de las bases para la toma de una muestra son:

a. Cada lote debe representar la producción durante un intervalo de tiempo, tal que todas las partes o productos en el lote se hayan elaborado esencialmente bajo las mismas condiciones (partes de orígenes diferentes o en condiciones diferentes no deben mezclarse en el mismo lote); esto se recomienda para que todas las unidades muestrales tengan la misma posibilidad o probabilidad de ser elegidos.

b. Son preferibles lotes grandes en vez de pequeños. Esto se recomienda para poder aplicar el teorema del límite central y con ello suponer normalidad de los datos, como más adelante se verá.

Por las dos causas anteriores se requieren definir algunos tipos de muestreo probabilístico para que la muestra obtenida sea representativa de la población bajo estudio:

a. Muestreo Aleatorio Simple.
b. Muestreo Aleatorio Estratificado.
c. Muestreo Aleatorio por Conglomerados.
d. Muestreo Aleatorio Sistemático.

2.2. Muestreo Aleatorio Simple

Se selecciona un grupo de n unidades muestrales de tal manera que cada muestra de tamaño n tenga la misma probabilidad de ser seleccionada. Este tipo de muestreo es utilizado cuando se tiene acceso a la población completa, la cual se encuentra "a granel" o en un solo contenedor y no hay distinción entre cada elemento de la población.

Para elegir una muestra aleatoria simple, el procedimiento empleado es el siguiente:

i. Se asigna un número a cada individuo de la población y

ii. A través de algún medio mecánico, como por ejemplo, un dado homogéneo con tantas caras como números se vayan a elegir, bolas dentro de una urna, tablas de números aleatorios, números aleatorios generados con una calculadora u ordenador, etcétera, se eligen tantos sujetos como sea necesario para completar el tamaño de muestra requerido. Al efectuar el muestreo se debe definir previamente si este va a ser realizado con repetición o sin repetición; por ejemplo, al lanzar un dado homogéneo de 20 caras, puede caer el número tres, si se vuelve a lanzar el dado puede volver a caer el número tres; se requiere definir si es válido que se vuelva a tomar el número tres o se va a eliminar una vez tomado. Este procedimiento, atractivo por su sencillez, no se requiere cuando la población es muy grande o infinita.

Suponga que se tiene un lote con N artículos, de los cuales una parte reducida de ellos, D, es defectuosa; suponga que se obtiene una muestra aleatoria simple de tamaño n; al tomar la muestra y determinar cuáles de ellos son defectuosos, suponga que x representa al número de artículos defectuosos en la muestra, entonces la función de probabilidad del número de artículos defectuosos en la muestra presenta distribución hipergeométrica:

$$p(x; n, D, N) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, 2, \cdots, \textit{mínimo}(D, n)$$

La media de esta distribución y su varianza están dadas de la siguiente forma:

$$\mu_x = n\left(\frac{D}{N}\right)$$

$$\sigma_x^2 = n\left(\frac{D}{N}\right)\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Sea $p = D/N$ la fracción defectuosa en la población bajo estudio y sea

$$\hat{p} = \frac{x}{n}$$

la fracción defectuosa en la muestra, entonces

$$p\left(p-\varepsilon < \frac{x}{n} < p+\varepsilon\right) = 1-\alpha$$

Recuerde que $\text{var}\{ax\} = a^2\text{var}\{x\}$ por lo que $\text{var}\{x/n\} = \text{var}\{x\}/n^2$

Lo que implica que la varianza del estimador de p está dada por:

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

p es el parámetro poblacional que representa la fracción defectuosa en el lote, el cual obviamente es desconocido, por lo que se requiere estimarlo con un cierto grado de aproximación. Una forma de hacerlo es conociendo los resultados de una premuestra en la cual se tienen d defectuosos en una premuestra de tamaño m, con lo que se podría usar como estimador de p el cociente $p\hat{} = d/m$. Una aproximación mayor sería si se conocen varias muestras donde se tienen las fracciones defectuosas de cada una de ellas $p_1, p_2, \cdots, p_q$; en este caso se podría usar como estimador de p la media de las p_j anteriores.

La varianza del estimador de p quedaría expresada como:

$$\sigma_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right) \qquad (2.6)$$

El límite superior del error de estimación estaría dado por:

$$\varepsilon = k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}$$

Recuerde que si $np > 5$ para $p < 1/2$, la distribución hipergeométrica puede ser aproximada por una distribución normal, por lo que se puede considerar que $k = z_{\alpha/2}$.

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)} \tag{2.7}$$

Suponga que se desea diseñar un plan de muestreo aleatorio simple para estimar la fracción defectuosa p en una población de artículos de tamaño finito N. En este caso, para calcular el tamaño de la muestra que se debe obtener, se despejaría n de la expresión 2.7.

Despejando n de la expresión 2.7:

$$n \geq \frac{\hat{p}(1-\hat{p})N}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 (N-1) + \hat{p}(1-\hat{p})} \tag{2.8}$$

Las expresiones 2.6, 2.7 y 2.8 son válidas para una población finita de tamaño N, pero, ¿qué pasaría si el tamaño de la población fuera muy grande?, de tal forma que se pudiera considerar como infinito:

Nótese que

$$Lim_{N\to\infty}\left(\frac{N-n}{N-1}\right) = 1 \tag{2.9}$$

Lo que implica que la varianza del estimador de p para un muestreo aleatorio simple de una población infinita está dada por:

$$\sigma^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n} \tag{2.10}$$

El límite superior del error de estimación para un muestreo aleatorio simple de una población infinita está dada por:

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.11}$$

Suponga que se desea diseñar un plan de muestreo aleatorio simple para estimar la fracción defectuosa p en una población infinita. En este caso:

$$n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \hat{p}(1-\hat{p}) \tag{2.12}$$

Las expresiones de la 2.6 a la 2.12 corresponden a la estimación del parámetro poblacional p, el cual representa a la fracción defectuosa de una población de artículos. Pero también puede representar cualquier fracción de interés, por ejemplo, de los estudiantes de la Facultad de Ingeniería, qué fracción de ellos son mujeres, qué fracción de pobladores de una ciudad tiene diabetes, qué fracción de los diputados tiene estudios de posgrado, etcétera.

Ahora, se deducirán fórmulas para la media de una población a partir de los estadísticos de tendencia central de una muestra.

En el libro de *Fundamentos de Probabilidad y Aplicaciones*, específicamente en el subtema 7.5. *Distribución de Probabilidad Normal de la Media Aritmética* (promedio) de una muestra, se estableció que por el Teorema de Aditividad de la Distribución Normal, se puede afirmar que si una población presenta distribución normal, entonces la media aritmética de una muestra aleatoria de dicha población también presenta distribución normal. De la misma forma, por el Teorema del Límite Central, se puede afirmar que si se extrae una muestra aleatoria de tamaño grande de una población con cualquier distribución, la media muestral tenderá a tener distribución normal en la medida en que n tienda a infinito.

Esto se puede resumir de la siguiente forma:

Sea

$$\bar{x} = \frac{1}{n}\left[x_1 + x_2 + \cdots + x_n\right]$$

Entonces

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{2.13}$$

Bajo uno de los dos supuestos siguientes: x es normal o n tiende a ser muy grande (tiende a infinito).

En la práctica profesional de la Ingeniería, el pensar en una muestra de tamaño infinito es ilusorio; en realidad, los lotes presentan tamaño finito y las muestras con mayor razón; surge entonces la pregunta, ¿de qué tamaño debe ser la muestra para suponer que la media muestral de una población con distribución diferente a la normal es normal? La respuesta desde luego depende de la distribución real que presente. Algunas reglas empíricas o "de dedo", establecen experimentalmente lo siguiente:

i. Si la distribución de probabilidad de una población es parecida a una normal, es decir, si la gráfica de la función de probabilidad de una población presenta una forma parecida al perfil de una campana casi simétrica, basta tomar $n \geq 4$.

ii. Si la distribución de probabilidad de una población es parecida a una distribución uniforme, basta tomar $n \geq 12$.

iii. Si la distribución de probabilidad de una población presenta la mayor parte de sus medidas en alguno de sus extremos (cóncava hacia arriba y no simétrica), por ejemplo, la distribución exponencial negativa, a lo cual se le podría denominar comportamiento antinormal, $n \geq 100$.

En la deducción de las fórmulas aplicables al parámetro p, se vio que cuando la población es finita, aparece el factor de corrección mostrado en la expresión 2.9. Este mismo factor se considerará para la estimación de la media.

Si se asume que la media muestral es normal, el parámetro poblacional a estimar es μ_x.

El estimador de la media poblacional a usar será el estadístico denominado media muestral:

$$\hat{\mu}_x = \bar{x} = \frac{1}{n}\left[x_1 + x_2 + \cdots + x_n\right] \tag{2.14}$$

De la expresión 2.4:

$$p\left(\mu_x - \varepsilon < \bar{x} < \mu_x + \varepsilon\right) = 1 - \alpha$$

Donde

$$\varepsilon = z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \left(\frac{N-n}{N-1}\right)$$

Como no se conoce la desviación estándar poblacional, se utilizará un estimador de la misma, suponiendo que de muestras históricas o de una premuestra se puede calcular la desviación estándar muestral S_x. De esta forma, la variancia del estimador, denominado media muestral para un muestreo aleatorio simple para una población finita de tamaño N, está dado por:

$$\sigma^2_{\bar{x}} = \frac{S^2_x}{n}\left(\frac{N-n}{N-1}\right) \tag{2.15}$$

El límite superior del error de estimación para la media, con un muestreo aleatorio simple para una población finita de tamaño N, está dado por:

$$\varepsilon = z_{\alpha/2}\, \sigma_{\bar{x}} = z_{\alpha/2} \frac{S_x}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)} \tag{2.16}$$

Suponga que se desea diseñar un plan de muestreo aleatorio simple para estimar la media de una población finita de tamaño N. En este caso, para calcular el tamaño de la muestra que se debe obtener, se despeja n de la expresión 2.16:

$$n \geq \frac{NS^2_x}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + S^2_x} \tag{2.17}$$

Si la población fuera infinita las fórmulas anteriores estarían dadas por las siguientes expresiones:

$$\sigma^2_{\bar{x}} = \frac{S^2_x}{n} \tag{2.18}$$

El límite superior del error de estimación para la media, con un muestreo aleatorio simple para una población infinita, está dado por:

$$\varepsilon = k\sigma_{\bar{x}} = z_{\alpha/2}\frac{S_x}{\sqrt{n}} \tag{2.19}$$

Suponga que se desea diseñar un plan de muestreo aleatorio simple para estimar la media de una población infinita, el tamaño de muestra adecuado se calcula con la siguiente expresión:

$$n \geq \left(\frac{z_{\alpha/2}\, S_x}{\varepsilon}\right)^2 \tag{2.20}$$

Con el objeto de analizar el estimador del parámetro total poblacional τ, si una población es finita, entonces la media de la población está dada por $\mu_x = \tau/N$, lo que implica que $\tau = N\mu_x$. Con este razonamiento, el estimador de τ es

$$\hat{\tau} = N\,\bar{x} \tag{2.21}$$

De tal forma que las fórmulas para el total poblacional quedan de la siguiente manera:

La varianza del estimador τ para una población finita es:

$$\sigma_{\tau}^2 = \frac{(NS_x)^2}{n}\left(\frac{N-n}{N-1}\right) \tag{2.22}$$

El límite superior del error de estimación para el total poblacional está dado por:

$$\varepsilon = z_{\alpha/2}\,\sigma_{\hat{\tau}} = z_{\alpha/2}\frac{NS_x}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)} \tag{2.23}$$

Suponga que se desea diseñar un plan de muestreo aleatorio simple para estimar el total poblacional para una población de tamaño N. En este caso, para calcular el tamaño de la muestra que se debe obtener, se despeja n de la expresión 2.23.

Despejando *n* de la expresión 2.23:

$$n \geq \frac{N^3 S_x^2}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + N^2 S_x^2} \tag{2.24}$$

Las fórmulas a emplear para el muestreo aleatorio simple para una población finita, son las que se muestran en la figura 2.4 a continuación:

Figura 2.4. Estimadores para el Muestreo Aleatorio Simple

MUESTREO ALEATORIO SIMPLE				
Parámetro Poblacional	Estimador Puntual	Varianza del Estimador	Límite Error Estimación	Tamaño de Muestra
Media μ_x	$\hat{\mu}_x = \bar{x} = \frac{1}{n}[x_1 + x_2 + \cdots + x_n]$	$\sigma^2_{\bar{x}} = \frac{S_x^2}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\bar{x}} = z_{\alpha/2}\frac{S_x}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{NS_x^2}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + S_x^2}$
Tamaño Poblacional τ	$\hat{\tau} = N\bar{x}$	$\sigma^2_{\hat{\tau}} = \frac{(NS_x)^2}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{\tau}} = z_{\alpha/2}\frac{NS_x}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N^3 S_x^2}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + N^2 S_x^2}$
Fracción p de la Población	$\hat{p} = \frac{x}{n}$	$\sigma^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{\hat{p}(1-\hat{p})N}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 (N-1) + \hat{p}(1-\hat{p})}$

2.3. Muestreo Aleatorio Estratificado

En Geología se llama estrato a cada una de las capas en que se presentan divididos los sedimentos, las rocas sedimentarias, las rocas piroclásticas y las rocas metamórficas, cuando esas capas se deben al proceso de sedimentación. La rama de la geología que estudia los estratos recibe el nombre de estratigrafía. En la figura 2.5 se muestra un ejemplo de terreno estratificado.

Figura 2.5. Terreno estratificado

Tomasz Kuran (2005). Flysch (turbidita) de los Cárpatos en Komańcza (Polonia). Recuperado de https://upload.wikimedia.org/wikipedia/commons/0/0b/Carpathian_flysch_cm04.jpg

En cada capa o estrato, las partículas que lo componen presentan las mismas propiedades físicas, de allí que se puede determinar a qué estrato pertenece cada partícula.

En el caso de la Estadística, cuando una población se encuentra claramente dividida en contenedores diferentes y en cada contenedor los artículos presentan las mismas características entre ellos, se dice que dicha población se encuentra

estratificada. Por ejemplo, en un salón de clases los alumnos se pueden estratificar claramente en hombres o en mujeres. También pueden ser estratificados por carrera, por generación, por edad, por estatura, por tipo de sangre, por peso, etcétera. En una fábrica, los obreros pueden ser clasificados por turno, por línea de producción, por área, por función, etcétera.
Las razones principales para usar el muestreo aleatorio estratificado en lugar del muestreo aleatorio simple se pueden resumir en tres principios básicos:

1. La estratificación produce errores de estimación con límite superior menor que el que surgiría aplicando el muestreo aleatorio simple y se vuelve mejor en la medida en que los estratos son más homogéneos.

2. El costo por observación se reduce al estratificar la población en grupos convenientes.

3. Se pueden estimar parámetros poblacionales para cada estrato.

Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra. El procedimiento para seleccionar una muestra aplicando muestreo aleatorio estratificado requiere determinar si se muestrearan todos y cada uno de los estratos, o se tomará una muestra aleatoria de estratos; posteriormente, se tiene que determinar si:

a. Para cada estrato elegido se tomará igual número de unidades muestrales

b. Para cada estrato elegido la distribución se hará de acuerdo con el peso específico (tamaño proporcional) del mismo.

c. Para cada estrato elegido se toma en cuenta la proporción y la dispersión de los datos, la cual tiene poca aplicación generalmente porque no se suele conocer la desviación de cada estrato.

Las fórmulas a emplear para el muestreo aleatorio estratificado para una población finita son las que se muestran en la figura 2.6 a continuación.

Figura 2.6. Estimadores para el Muestreo Aleatorio Estratificado

MUESTREO ALEATORIO ESTRATIFICADO					
Parámetro Poblacional	Estimador Puntual	Varianza del Estimador	Límite Error Estimación	Tamaño de Muestra	Condiciones
Media μ_x	$\bar{\mu}_x = \frac{1}{N}\frac{\sum_{i=1}^{L} N_i x_i}{n}$	$\sigma^2_{\bar{\mu}_x} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{S_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\bar{\mu}_x} = \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{L} N_i^2\left(\frac{S_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)}$	$n \geq \frac{L\sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{(N_i-1)}}{N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{(N_i-1)}}$	Para tamaño de muestra n constante en cada estrato $n_i = n/L$
				$n \geq \frac{\sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{\omega_i(N_i-1)}}{N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{\omega_i(N_i-1)}}$	Para tamaño de muestra proporcional al tamaño de cada estrato $n_i = nw_i$
				$n \geq \frac{\left(\sum_{i=1}^{h+1} \frac{N_i\sigma_i}{(N_i-1)\sqrt{k_i}}\right)\left(\sum_{i=1}^{i+1} N_i\sigma_i\sqrt{k_i}\right)}{\left[N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} \frac{N_i^2\sigma_i^2}{(N_i-1)}\right]}$	Para costos fijos de muestreo en cada estrato y minimizando la varianza
Tamaño Poblacional τ	$\hat{\tau} = N\bar{\mu}_x$	$\sigma^2_{\hat{\tau}} = \sum_{i=1}^{L} N_i^2\left(\frac{S_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)$	$\varepsilon = z_{\alpha/2}\sqrt{\sum_{i=1}^{L} N_i^2\left(\frac{S_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)}$	$n \geq \frac{L\sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{\omega_i(N_i-1)}}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{(N_i-1)}}$	Para tamaño de muestra constante en cada estrato $n_i = n/L$
				$n \geq \frac{\sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{\omega_i(N_i-1)}}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{L} \frac{N_i^2\sigma_i^2}{(N_i-1)}}$	Para tamaño de muestra proporcional al tamaño de cada estrato $n_i = nw_i$
				$n \geq \frac{\left(\sum_{i=1}^{h+1} \frac{N_i\sigma_i}{(N_i-1)\sqrt{k_i}}\right)\left(\sum_{i=1}^{i+1} N_i\sigma_i\sqrt{k_i}\right)}{\left[\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} \frac{N_i^2\sigma_i^2}{(N_i-1)}\right]}$	Para costos fijos de muestreo en cada estrato y minimizando la varianza
Fracción p de la Población	$\bar{p} = \frac{1}{N}\frac{\sum_{i=1}^{L} N_i p_i}{n}$	$\sigma^2_{\bar{p}} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\frac{\bar{p}_i(1-\bar{p}_i)}{n_i - 1}\left(\frac{N_i - n_i}{N_i - 1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\bar{p}} = \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{L} N_i^2\frac{\bar{p}_i(1-\bar{p}_i)}{n_i - 1}\left(\frac{N_i - n_i}{N_i - 1}\right)}$	$n \geq \frac{L\left[N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} N_i^2 p_i(1-p_i)/(N_i-1)\right]}{\left[N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} \frac{N_i^2 p_i(1-p_i)}{(N_i-1)}\right]}$	Para tamaño de muestra n constante en cada estrato $n_i = n/L$
				$n \geq \frac{\sum_{i=1}^{i+1} N_i^2 p_i(1-p_i)/(N_i-1)\omega}{\left[N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} N_i^2 p_i(1-p_i)/(N_i-1)\right]}$	Para tamaño de muestra proporcional al tamaño de cada estrato $n_i = nw_i$
				$n \geq \frac{\left(\sum_{i=1}^{i+1} N_i\sqrt{p_i(1-p_i)}/(N_i-1)\sqrt{k_i}\right)\left(\sum_{i=1}^{i+1} N_i\sqrt{p_i(1-p_i)}\sqrt{k_i}/(N_i-1)\right)}{\left[N^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sum_{i=1}^{i+1} N_i^2 p_i(1-p_i)/(N_i-1)\right]}$	Para costos fijos de muestreo en cada estrato y minimizando la varianza

Notación en el cuadro anterior:

L: Número de estratos
N_i: Número de unidades en el estrato i
n_i: Número de unidades en la muestra del estrato i
S_i: Desviación estándar en la muestra del estrato i
σ_i: Desviación estándar del estrato i
p_i: Fracción de éxitos en el estrato i
N: Número de unidades en la población $N = N_1 + N_2 + \cdots + N_L$
ω_i: Ponderación o peso específico del estrato i, $\omega_i = N_i / N \approx n_i/n$
$z_{\alpha/2}$: valor de z para un nivel de confianza $1-\alpha$
c_i: Costo de muestreo para el estrato i

2.4. Muestreo Aleatorio por Conglomerados

En Geología, un conglomerado o rudita es una roca sedimentaria de tipo detrítico (es decir, que se obtiene como resultado de la descomposición de una masa sólida en partículas), formada mayoritariamente por clastos redondeados (así se llaman las partículas que forman la roca sedimentaria clástica) tamaño grava o mayor (>2 mm). Dichos clastos pueden corresponder a cualquier tipo de roca. Las características del conglomerado son:

a. Es heterogéneo.
b. Los granos que lo componen son trozos de otras rocas constituyentes, de tamaños y formas distintos.
c. No está dispuesto en láminas como la pizarra.

En la figura 2.7 se muestra un conglomerado cuarcítico.

Figura 2.7. Conglomerado cuarcítico

Jesús Muñoz P. (2016). Recuperado de https://www.biodiversidadvirtual.org/geologia/data/media/114/Conglomerado-cuarcitico-7761.jpg

En Estadística el muestreo por conglomerados es una técnica que debe aplicarse cuando la población bajo estudio presenta agrupamientos "naturales" no homogéneos o relativamente homogéneos en una población estadística. A menudo se utiliza en la investigación de mercados.

Para ilustrar la necesidad del muestreo por conglomerados, piense en una población como la de la Ciudad de México, la cual se encuentra agrupada por colonias. Se desea medir de alguna forma el nivel socioeconómico de sus pobladores. Se podría pensar que una colonia como Bosques de las Lomas representa un nivel socioeconómico alto, por lo que podría pensarse en un estrato alto ya que es homogéneo, pero esto no se puede generalizar; en la mayoría de las más de 1800 colonias que constituyen a la Ciudad de México las colonias contienen elementos de niveles socioeconómicos altos junto con elementos de niveles socioeconómicos medios o bajos, por lo que las colonias no se comportan como estratos sino como conglomerados.

Se justifica el empleo del muestreo aleatorio por conglomerados cuando se pretende hacerlo a costo mínimo bajo las siguientes condiciones:

1. No se cuenta con el conocimiento del total poblacional porque es costoso, pero se conoce que está constituido por conglomerados.
2. El costo por obtener unidades muestrales se incrementa con la distancia que separa a los elementos.

En esta técnica de muestreo aleatorio por conglomerados, los pasos a seguir son los siguientes:

a. La población total se divide en grupos (o clusters).
b. Se elige una muestra aleatoria simple de estos grupos.
c. Se toma una muestra aleatoria simple de cada uno de los grupos elegidos.
d. Se recopila la información requerida de los elementos dentro de cada grupo seleccionado.

Suponiendo un tamaño de muestra fijo, la técnica ofrece resultados más precisos cuando la mayoría de la variación en la población se encuentra dentro de los grupos y no entre ellos.

Las fórmulas a emplear para el muestreo aleatorio por conglomerados, para una población finita, son las que se muestran en la figura 2.8.

Figura 2.8. Estimadores para el Muestreo Aleatorio por Conglomerados

MUESTREO ALEATORIO POR CONGLOMERADOS				
Parámetro Poblacional	Estimador Puntual	Varianza del Estimador	Límite Error Estimación	Tamaño de Muestra
Media μ_x	$\hat{\mu}_x = \frac{\sum_{i=1}^{i=n} y_i}{\sum_{i=1}^{i=n} m_i}$	$\sigma_{\bar{y}}^2 = \frac{1}{n(n-1)\bar{M}^2}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y}m_i)^2$	$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n(n-1)\bar{M}^2}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y}m_i)^2}$	$n \geq \frac{N\sigma_e^2}{(N-1)\bar{M}^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2+\sigma_e^2}$
Tamaño Poblacional τ	$\hat{\tau} = M\hat{\mu}_x = M\frac{\sum_{i=1}^{i=n} y_i}{\sum_{i=1}^{i=n} m_i}$	$\sigma_{\hat{\tau}}^2 = \frac{N^2}{n(n-1)}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y}m_i)^2$	$\varepsilon = z_{\alpha/2}N\sqrt{\frac{1}{n(n-1)}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y}m_i)^2}$	$n \geq \frac{N\sigma_e^2}{(N-1)\bar{M}^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2+\sigma_e^2}$
	Si no se conoce M:	Si no se conoce M:	Si no se conoce M:	Si no se conoce M:
	$\hat{\tau} = N\hat{\mu}_x = \frac{N}{n}\sum_{j=1}^{j=n} y_j$	$\sigma_{\hat{\tau}}^2 = \frac{N^2}{n(n-1)}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y})^2$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{\tau}} = z_{\alpha/2}N\sqrt{\frac{1}{n(n-1)}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(y_i-\bar{y})^2}$	$n \geq \frac{N\sigma_e^2}{\frac{(N-1)}{N^2}\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2+\sigma_e^2}$
Fracción p de la Población	$\hat{p} = \frac{\sum_{i=1}^{i=n} a_i}{\sum_{i=1}^{i=n} m_i}$	$\sigma_{\hat{p}}^2 = \frac{1}{n(n-1)\bar{M}^2}\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{i=n}(a_i-\hat{p}m_i)^2$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{p}} = \frac{z_{\alpha/2}}{\bar{M}}\sqrt{\frac{1}{n(n-1)}\left(\frac{N-n}{(N-1)}\right)\sum_{i=1}^{i=n}(a_i-\hat{p}m_i)^2}$	$n \geq \frac{N\sigma_e^2}{(N-1)\bar{M}^2\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2+\sigma_e^2}$

Dónde

$$\hat{\sigma}_c^2 = S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2$$

Notación en el cuadro anterior

N: Número de conglomerados en la población
n: Número de conglomerados seleccionados en una muestra aleatoria simple
m_i: Número de elementos en el conglomerado i

$\overline{m} = \frac{1}{n}\sum_{i=1}^{i=n} m_i$: Tamaño promedio de los conglomerados en la muestra

$M = \sum_{i=1}^{i=N} m_i$: Número de elementos en la población

$\overline{M} = M / N$: Tamaño promedio del conglomerado en la población. $\overline{M} \approx \overline{m}$

y_i : Número de observaciones en el i-ésimo conglomerado.

2.5 Muestreo Aleatorio Sistemático

Se utiliza cuando la población está ordenada de alguna forma, por ejemplo, un grupo de estudiantes ordenados de mayor a menor estatura o viceversa; otro ejemplo, en un proceso a lo largo del tiempo, en una línea de producción, los productos van surgiendo a cierta velocidad y se pretende obtener una muestra representativa del proceso a lo largo del tiempo.

Una muestra que se obtiene al seleccionar aleatoriamente un elemento de los primeros k elementos que surgen de un proceso y después se eligen los demás elementos de la muestra cada k-ésimo elemento se denomina muestreo sistemático de 1 en k.

Los pasos para aplicar el muestreo aleatorio sistemático son los siguientes:

a. Primero hay que identificar las unidades, elementos o artículos y ordenarlas bajo un criterio o relacionarlas con el tiempo (cuando proceda).

b. Luego hay que calcular una constante, denominada coeficiente de elevación: $K = N/n$, donde n es el tamaño de la población y n el tamaño de la muestra.

c. Para determinar en qué tiempo se hará la primera extracción de la línea de producción se elige al azar un número entre 1 y K.

d. De allí en adelante se toma una unidad muestral cada K unidades a intervalos regulares. Ocasionalmente, es conveniente tener en cuenta la periodicidad del fenómeno. Cuando la línea de producción se liga con el tiempo, entonces el proceso extractivo se hace no cada K unidades, sino cada cierto tiempo, el cual va a depender de la velocidad de proceso.

Las fórmulas que se utilizan para el muestreo aleatorio sistemático, para una población finita, son las que se muestran en la figura 2.9.

Figura 2.9. Estimadores para el Muestreo Aleatorio Sistemático

MUESTREO ALEATORIO SISTEMÁTICO				
Parámetro Poblacional	Estimador Puntual	Varianza del Estimador	Límite Error Estimación	Tamaño de Muestra
Media μ_x	$\hat{\mu}_x = \frac{\sum_{i=1}^{i=n} x_i}{n}$	$\sigma^2_{\hat{\mu}_x} = \frac{S_x^2}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{\mu}_x} = z_{\alpha/2}\frac{S_x}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N\sigma_x^2}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + \sigma_x^2}$
Tamaño Poblacional τ	$\hat{\tau} = N\hat{\mu}_x$	$\sigma^2_{\tau} = \frac{N^2 S_x^2}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{\tau}} = z_{\alpha/2}\frac{NS_x}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N^3\sigma_x^2}{(N-1)\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 + N^2\sigma_x^2}$
Fracción p de la Población	$\hat{p} = \frac{x}{n}$	$\sigma^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\alpha/2}\sigma_{\hat{p}} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N\hat{p}(1-\hat{p}) + \left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 (N-1)}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2 (N-1) + \hat{p}(1-\hat{p})}$

2.6. Métodos de Muestreo No Probabilísticos

En ocasiones el muestreo probabilístico, aunque deseable para darle certeza al proceso, resulta excesivamente costoso y se acude a métodos no probabilísticos por facilidad y rapidez. Entre los métodos de muestreo no probabilístico más utilizados en la práctica profesional se encuentran:

1. Muestreo por cuotas. Se basa en el nivel de conocimientos de los estratos de la población y/o de los individuos más "representativos" o "adecuados" para los fines del estudio. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquél. Este método se utiliza mucho en las encuestas de opinión.

2. Muestreo intencional o de conveniencia. Este tipo de muestreo se distingue por tratar de obtener muestras "representativas" de la población bajo estudio, mediante la inclusión en la muestra de grupos supuestamente típicos. Es muy frecuente su utilización en sondeos preelectorales de zonas que en anteriores votaciones han marcado tendencias de voto. También puede ser que el planeador seleccione directa e intencionadamente a los individuos de la población.

3. Bola de nieve. Se localiza a algunos individuos, los cuales conducen a otros, y estos a su vez a otros, y así hasta conseguir una muestra adecuada. Este tipo de muestreo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas, determinados tipos de enfermos, etcétera.

4. Muestreo Discrecional. A criterio del investigador los elementos son elegidos sobre lo que él cree que pueden aportar al estudio.

Ejercicios del Capítulo 2

1. Explique qué es el muestreo, cuál es su objetivo principal y cuáles son sus ventajas y desventajas al compararse con realizar un censo.

2. Ilustre con casos prácticos cada uno de los tipos de muestreo probabilístico que existen.

3. En el siguiente cuadro se ilustran estadísticas de los partidos de futbol en Europa. Suponga que este cuadro representa un universo de partidos jugados.

No.	Equipo	Campeonato	Goles	Tiros pp	Disciplina	Posesion%	AciertoPase%	Aéreos	Rating
1	Bayern Munich	Bundesliga	55	18.9	352	62.9	87.5	13.8	7.14
2	Paris Saint-Germain	Ligue 1	52	15.9	350	61.1	89.9	8.3	7.1
3	Manchester City	Premier League	65	19.9	442	61	88.7	14.6	7.05
4	Liverpool	Premier League	56	15.5	241	58.3	84	18	7.02
5	Real Madrid	La Liga	39	16.2	434	56.6	86.6	16.9	7.02
6	RasenBallsport Leipzig	Bundesliga	51	16.9	261	54.6	82.5	16.3	7
7	Borussia Dortmund	Bundesliga	51	13.9	181	58.5	86.3	11.9	6.98
8	Leicester	Premier League	52	14.1	230	55.1	82.7	17.3	6.97
9	Atalanta	Serie A	57	20.3	462	55	83.3	16.3	6.97
10	Barcelona	La Liga	50	12.6	484	62.2	88.5	12.2	6.94
11	Lazio	Serie A	47	15.9	542	50.3	84.5	12	6.94
12	Borussia M.Gladbach	Bundesliga	36	13.9	411	52.5	80.2	18.1	6.94
13	Inter	Serie A	42	16.6	543	52	84.9	15.2	6.9
14	Chelsea	Premier League	41	16.7	430	57.5	84.6	19	6.89
15	Juventus	Serie A	40	17.1	462	56.7	87.7	11.8	6.89
16	Lyon	Ligue 1	34	13.7	323	55.4	86.1	16	6.87
17	Bayer Leverkusen	Bundesliga	30	16.2	275	59.9	84.8	15.9	6.85
18	Eintracht Frankfurt	Bundesliga	31	16.4	373	48.6	74.5	20.8	6.82
19	Roma	Serie A	38	17.1	564	53.2	84.2	16.4	6.82
20	Marseille	Ligue 1	30	14.1	503	52.9	83	18.2	6.82

*Se muestran solo jugadores de la Premier League inglesa, Ligue 1 francesa, Bundesliga alemana, Serie A italiana y Liga española. Estadísticas del Futbol Europeo (Feb 2020). Fuente: https://es.whoscored.com/Statistics. © WhoScored.

a. Obtenga el promedio de goles y la desviación estándar de goles de los 20 equipos, el total de goles de todos los equipos, y la fracción de equipos que obtuvo más de 55 goles.
b. Obtenga una muestra aleatoria simple de tamaño $n = 5$ y estime puntualmente, a partir de esta muestra, el promedio de goles y la desviación estándar de goles de los 20 equipos, el total de goles de todos los equipos, y la fracción de equipos que obtuvo más de 55 goles.
c. Obtenga una muestra aleatoria sistemática de tamaño $n = 5$ y estime puntualmente, a partir de esta muestra, el promedio de goles y la desviación estándar de goles de los 20 equipos, el total de goles de todos los equipos, y la fracción de equipos que obtuvo más de 55 goles.
d. Suponiendo que la tercera columna intitulada Campeonato representa el estrato, obtenga una muestra aleatoria estratificada de tamaño $n = 5$ y estime puntualmente, a partir de esta muestra, el promedio de goles y la desviación estándar de goles de los 20 equipos, el total de goles de todos los equipos, y la fracción de equipos que obtuvo más de 55 goles.
e. Compare los resultados estimados de cada tipo de muestreo probabilístico aplicado en los incisos *b*, *c* y *d*, con los resultados reales del inciso a y establezca qué tipo de muestreo es el que estima mejor a los resultados reales.
f. Para el caso del muestreo aleatorio simple, en el caso de querer estimar la media con un error de estimación máximo de cinco unidades, ¿de qué tamaño tendría que ser la muestra para lograrlo?

4. (Problema 5.10 de la página 115 del libro de Scheaffer, Mendenhall y Ott, "Elementos de Muestreo", Grupo Editorial Iberoamérica, 1986). Un guardabosques desea estimar el número total de hectáreas plantadas de árboles en los diversos parques de un estado. El número de hectáreas con árboles plantados varía considerablemente con respecto al tamaño de cada parque, por lo cual es conveniente aplicar un muestreo aleatorio estratificado. Los 240 parques en el estado se clasifican en cuatro estratos de acuerdo con el tamaño. El guardabosques tomó una muestra de 40 parques, aplicando asignación proporcional de acuerdo con el tamaño y los resultados se muestran en la siguiente tabla.

a. Estime puntualmente el número total de hectáreas plantadas de árboles en los 240 parques del estado, la varianza del estimador, el límite máximo para el error de estimación y el intervalo de confianza al 95% de nivel de confianza.

b. De qué tamaño tendría que ser la muestra para obtener un error de estimación máximo de 5000 hectáreas, al 95% de nivel de confianza.

Estrato I		Estrato II		Estrato III		Estrato IV	
0-100 hectáreas		100-200 hectáreas		201-300 hectáreas		más de 300 hectáreas	
N1=	86	N2=	72	N3=	52	N3=	30
n1=	14	n2=	12	n3=	9	n3=	5
Lecturas		Lecturas		Lecturas		Lecturas	
97	67	125	155	142	256	167	
42	125	67	96	310	440	220	
25	92	256	47	495	510	780	
105	86	310	236	320	396	655	
27	43	220	352	196		540	
45	59	142	190				
53	21						

5. Una Organización No Gubernamental desea estimar la proporción de votantes que apoyan a cierto candidato del Partido Morena para Gobernador del Estado de Veracruz, el cual cuenta con 212 municipios. Debido a que la selección de una muestra aleatoria simple de votantes registrados es muy costosa, se utiliza un plan de muestreo por conglomerados, considerando a cada municipio del estado como tal. Para ello, elige un tamaño de muestra de 25 municipios y manda a un representante de la ONG a cada uno de los municipios muestreados el día de la elección, para, mediante consulta directa a los votantes, determinar la fracción de votantes a favor del candidato de Morena, antes del cierre de las elecciones. Los resultados se muestran en la tabla siguiente.

   a. Estime puntualmente la fracción de votantes a favor del candidato de Morena, la varianza del estimador, el error de estimación que se comete y el intervalo de confianza al nivel de confianza del 95%.
   b. De qué tamaño tendría que ser la muestra para considerar un error de estimación máximo de 0.02, con un nivel de confianza del 95%.

Clave del municipio	Municipio	Población total	A favor de Morena
3	Acayucan	87,267	46,001
11	Alvarado	52,927	28,544
19	Astacinga	6,534	3,694
27	Benito Juárez	17,618	10,168
35	Citlaltépetl	12,109	4,138
44	Córdoba	218,153	119,896
45	Cosamaloapan de Carpio	57,147	29,870
53	Cuitláhuac	27,940	16,277
54	Chacaltianguis	12,494	6,952
61	Las Choapas	81,827	44,945
69	Gutiérrez Zamora	24,791	19,542
77	Isla	43,349	33,426
85	Ixtaczoquitlán	68,823	30,922
93	Jilotepec	16,682	7,776
101	Mariano Escobedo	37,285	22,699
109	Misantla	64,249	36,150
118	Orizaba	126,005	67,502
126	Paso de Ovejas	33,392	20,162
134	Puente Nacional	22,454	13,724
142	San Juan Evangelista	33,929	18,938
150	Tamalín	11,750	10,002
158	Tecolutla	23,865	18,018
166	Tepetlán	9,668	6,416
174	Tierra Blanca	106,277	66,601
182	Tlalnelhuayocan	18,715	10,817
		1,215,250	693,180

6. La fábrica de café de Nestlé Toluca, produce cerca de $N = 4000$ pomos de Nescafé de 250 gramos por turno de ocho horas y requiere diseñar un plan de muestreo sistemático para estimar la cantidad promedio de llenado de los pomos de nescafé que salen de la línea de producción. Por datos históricos ha estimado la desviación estándar poblacional del peso del contenido de cada pomo en 3 gramos, utilizando un nivel de confianza del 95% y suponiendo que el máximo error de estimación que se pretende lograr es de 3 gramos. Determine de qué tamaño debe ser la muestra total a recolectar y calcular cada cuándo deberá tomarse un pomo de la línea de producción para completar la muestra.

7. (Problema 10.3 de la página 273 del libro de Scheaffer, Mendenhall y Ott, "Elementos de Muestreo", Grupo Editorial Iberoamérica, 1986). Estimar el número de palabras de tres letras que hay en este libro, obteniendo inicialmente la densidad de palabras de tres letras por página en promedio. Establezca un límite para el error de estimación. Aplique tres técnicas de muestreo diferentes para efectuar esta estimación. ¿En su opinión qué técnica es mejor?, ¿qué suposiciones son necesarias para que estas técnicas sean adecuadas?

Capítulo 3. Estadística descriptiva

3.1. Diagrama de puntos, diagrama de dispersión bidimensional y tridimensional

El origen de la tan conocida frase de "una imagen vale más que mil palabras", proviene de un antiguo proverbio chino, atribuido a Confucio (孔子; 551 - 479 aC). Es más rápido entender el comportamiento de un conjunto de datos a través de un diagrama que a través de parámetros descriptivos de dichos datos. Para ilustrar lo anterior, se abordará el siguiente ejemplo.

Ejercicio 3.1

Se reciben lotes consecutivos de tamaño $N = 800$ de piezas de sustrato de cerámica a las que se les ha aplicado un revestimiento metálico mediante un proceso de deposición por vapor. La calidad del revestimiento depende de su grosor en milésimas de pulgada. Para conocer el comportamiento probabilístico del grosor se decidió tomar una muestra aleatoria de tamaño $n = 90$ y los resultados se muestran a continuación.

94.1	86.6	94.3	94.1	93.1	85.1	84.6	97.3	85.1
93.2	91.2	93.2	92.1	94.6	84.0	83.6	96.8	90.5
90.6	86.1	86.7	96.4	96.3	93.7	85.4	94.4	95.6
91.4	90.4	83.0	88.2	94.7	87.7	89.7	96.1	88.3
88.2	89.1	95.3	86.4	91.1	90.6	87.6	98.0	84.1
86.1	87.3	94.1	85.0	92.4	89.4	85.1	85.4	83.7
95.1	84.1	97.8	84.9	90.6	88.6	89.6	86.6	82.9
90.0	90.1	93.1	87.3	89.1	84.1	90.0	91.7	87.3
92.4	95.2	86.4	89.6	88.8	82.6	90.1	87.5	86.4
87.3	86.1	87.6	90.3	86.4	83.1	94.3	84.2	84.5

Antes de proceder a calcular parámetros estadísticos de la muestra conviene presentarlos de una forma gráfica, utilizando lo que se conoce como Gráfico de Puntos, el cual es útil para mostrar datos cuantitativos de una forma organizada.

Se usan puntos para mostrar datos a lo largo de un eje ordinal. Un gráfico de puntos es similar a un gráfico de líneas, pero sin las líneas. Solamente se muestran los puntos de datos.

Los pasos para representar una muestra a través de un Gráfico de Puntos se enumeran a continuación:

1. Ordene los puntos de menor a mayor.
2. Trace una escala horizontal en donde el menor valor quede a la izquierda del inicio del gráfico y el mayor valor, a la derecha. Cada dato represéntelo con un punto exactamente arriba del valor de cada dato en la escala ordinal.
3. Empiece a colocar los puntos de izquierda a derecha, desde el menor valor. Los valores de los datos que se repitan represéntelos con puntos uno encima del otro tantas veces como repeticiones haya.

Para proceder a mostrar el gráfico de puntos se utilizará Minitab, para lo cual se requiere aplicar los siguientes pasos:

a. Capturar los datos de la muestra en una sola columna de Minitab, por ejemplo, en la columna C1, para el ejemplo 3.1 el encabezado de los datos se llamará espesor, estos datos quedarán capturados en las celdas de la 1 a la 90 y la primera celda de esta columna será el encabezado o rótulo.
b. En el menú principal de Minitab, se da un click en Graph y posteriormente en el submenú DotPlot, como se muestra en la figura 3.1

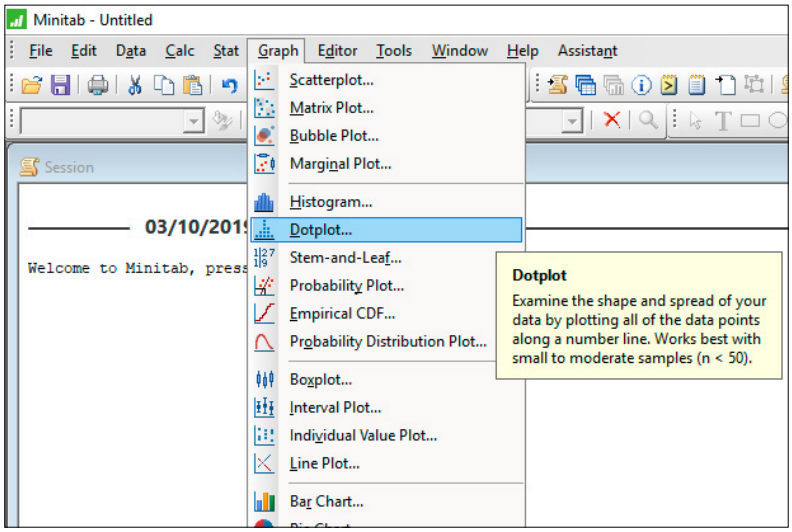


Figura 3.1.

c. Aparece la pantalla 3.2, en la cual se elige la primera opción que aparece señalada en negro, dando un click en el botón Ok.

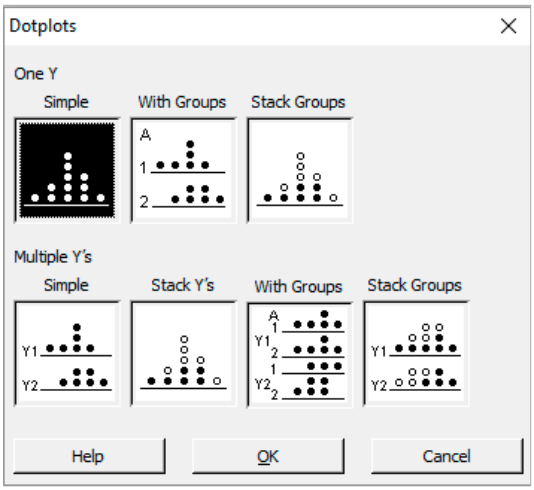


Figura 3.2.

d. Aparece la pantalla 3.3, en la cual se da un click en donde dice C1 Espesor y luego en Select mostrando la palabra Espesor en la ventana que dice Graph variables (puede dar clicks en las opciones Scale, Labels, Multiple Graphs y Data Options para ver a qué se refieren), posteriormente dar un click en el botón Ok.

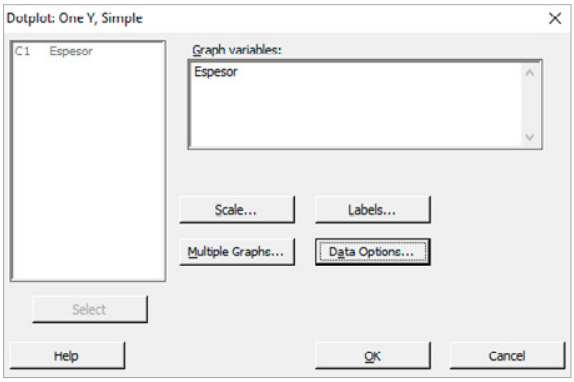
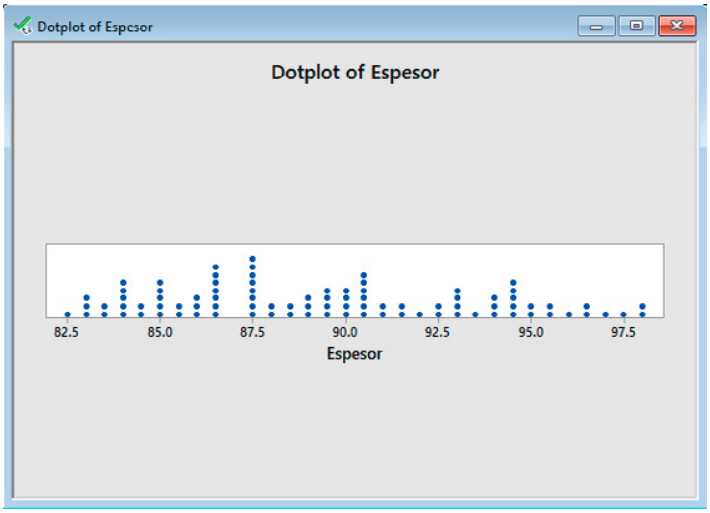


Figura 3.3.

e. Se obtiene el Gráfico de puntos de la Figura 3.4:

Figura 3.4. Gráfico de Puntos del Ejercicio 3.1



En este gráfico se puede ver el intervalo de variación de los datos (de 82.5 hasta 98.0) y qué datos se repiten. Claramente se aprecia que el dato más recurrente está ubicado en 87.5

El gráfico de puntos para el caso de una muestra aleatoria con dos variables, muestra bivariada o bidimensional, se denomina Gráfico de Dispersión, como se aprecia en el siguiente ejemplo.

Ejercicio 3.2

Se tomó una muestra de 64 alumnas de la Facultad de Ingeniería de la UNAM, a las cuales se les preguntó su peso y su estatura. La muestra se ilustra en la figura 3.5.

Para Graficar el Diagrama de Dispersión se aplican los siguientes pasos:

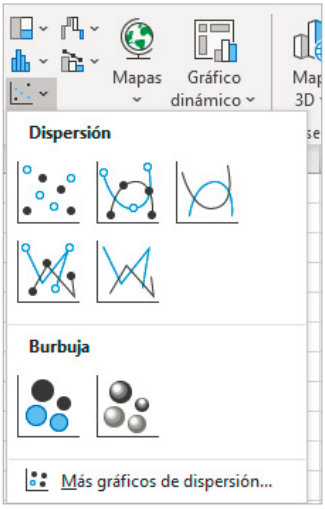
1. Se capturan los datos en una hoja de Excel. Por ejemplo, en las celdas de la A1 hasta la C65 (por los encabezados).

2. En la línea de menú principal en la parte superior de la pantalla principal de Excel se selecciona el menú Insertar, posteriormente, en el submenú que aparece en la parte central, arriba de donde dice Gráficos, aparece una imagen de un primer cuadrante, se selecciona la pestaña que aparece en la figura 3.6.

Figura 3.5. Pesos y Estaturas de una muestra de 64 alumnas de la Facultad de Ingeniería de la UNAM

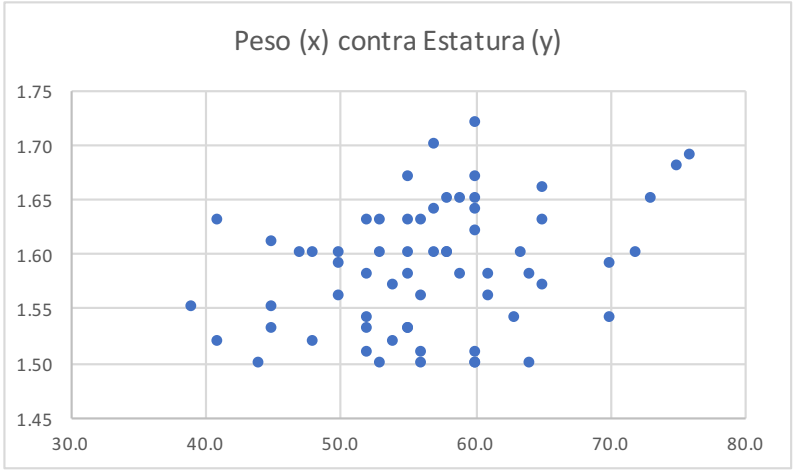
No.	Peso (Kg)	Estatura (m)	No.	Peso (Kg)	Estatura (m)
1	76.0	1.69	33	58.0	1.60
2	56.0	1.51	34	47.0	1.60
3	56.0	1.50	35	54.0	1.57
4	73.0	1.65	36	60.0	1.62
5	60.0	1.51	37	48.0	1.60
6	50.0	1.60	38	52.0	1.51
7	61.0	1.58	39	64.0	1.58
8	61.0	1.56	40	59.0	1.58
9	52.0	1.53	41	60.0	1.67
10	60.0	1.72	42	57.0	1.64
11	70.0	1.54	43	60.0	1.65
12	65.0	1.63	44	65.0	1.57
13	63.0	1.54	45	39.0	1.55
14	57.0	1.70	46	45.0	1.53
15	41.0	1.52	47	56.0	1.63
16	65.0	1.66	48	45.0	1.61
17	44.0	1.50	49	53.0	1.50
18	41.0	1.63	50	48.0	1.52
19	53.0	1.60	51	56.0	1.56
20	70.0	1.59	52	60.0	1.64
21	55.0	1.63	53	60.0	1.50
22	57.0	1.60	54	75.0	1.68
23	52.0	1.58	55	52.0	1.63
24	55.0	1.53	56	54.0	1.52
25	64.0	1.50	57	50.0	1.59
26	58.0	1.65	58	52.0	1.54
27	63.5	1.60	59	50.0	1.56
28	53.0	1.63	60	60.0	1.50
29	55.0	1.53	61	45.0	1.55
30	59.0	1.65	62	55.0	1.60
31	72.0	1.60	63	55.0	1.67
32	58.0	1.60	64	55.0	1.58

Figura 3.6.



3. Se selecciona la primera opción ya que se trata de un Diagrama de Puntos, apareciendo el Diagrama de Dispersión que se muestra en la figura 3.7.

Figura 3.7. Diagrama de Dispersión del peso contra la estatura de 64 alumnas de la Facultad de Ingeniería de la UNAM



También, el Diagrama de Dispersión puede graficarse en tres dimensiones para el caso de una muestra con tres variables. Por ejemplo, si en el anterior ejercicio se agrega una columna con el Índice de Masa Corporal (IMC = Peso/Estatura2) de cada alumna.

Para ello, se usará el software Minitab:

1. Lo primero es capturar los datos de peso en C1, estatura en χ^2 e IMC en C3.

2. En el menú principal se selecciona Graph y posteriormente se da un click en el submenú 3DScatterplot, en la ventana que aparece se selecciona la opción Simple y después Ok, con lo cual aparece la pantalla de la figura 3.8. En esta ventana se da un click en C3 y se selecciona Z, luego un click en χ^2 y se selecciona Y y un click en C1 y se selecciona X.

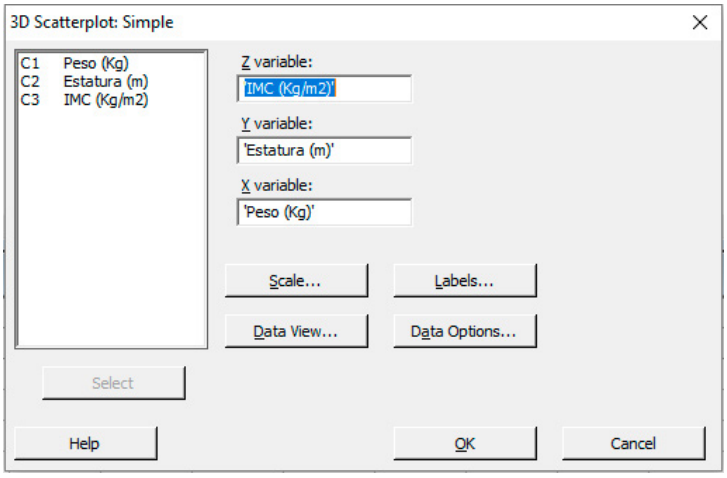


Figura 3.8.

3. Al oprimir Ok se obtiene el Diagrama de Dispersión Múltiple de la figura 3.9.

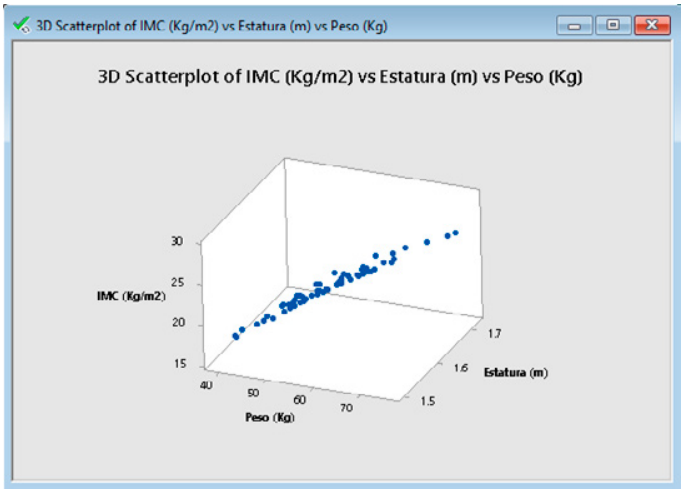


Figura 3.9. Diagrama de Dispersión con tres variables

3.2. Estadísticos muestrales

Uno de los propósitos primordiales de un muestreo probabilístico es tratar de aproximar valores de parámetros poblacionales a partir de valores de parámetros muestrales denominados estadísticos muestrales.

3.2.1. Estadístico muestral

Un estadístico muestral es una medida cuantitativa, obtenida a partir de un conjunto de datos, observaciones o mediciones de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico. Más formalmente, un estadístico muestral es una función medible f: $R^n \rightarrow R$ en la que, a una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$, le corresponde un número real $f(x_1, x_2, x_3, \cdots, x_n)$, que permite estimar un determinado parámetro de la población de la que procede la muestra.

Para ilustrar lo anterior, suponga que se requiere escoger un estadístico $\hat{\theta}$ del parámetro poblacional θ. Suponga a su vez, que existen varios estadísticos que pudieran usarse, ¿qué criterios existen para seleccionar al estadístico más adecuado?

Como más adelante se verá, los criterios que se utilizan son: eficiencia, consistencia, robustez, suficiencia e invariancia.

a. Insesgabilidad. Se dice que un estimador $\hat{\theta}$ es insesgado si la esperanza matemática del estimador $\hat{\theta}_2$ es igual al parámetro poblacional q, es decir,

$$\mu_{\hat{\theta}} = E\{\hat{\theta}\} = \theta \tag{3.1}$$

b. Eficiencia. Se dice que un estimador $\hat{\theta}_1$ es más eficiente o más preciso que otro estimador $\hat{\theta}_2$, si la varianza del primero es menor que la del segundo, es decir,

$$\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2} \tag{3.2}$$

Un estimador es más eficiente o más preciso mientras menor sea su varianza. El estimador más eficiente será aquel que presenta la mínima variancia, al cual se le conoce como estimador óptimo.

c. Consistencia. Un estimador $\hat{\theta}_1$ es consistente en la medida en que el valor de dicho estimador tiende a aproximarse cada vez más al valor del parámetro poblacional θ cuando el tamaño de muestra tiende a crecer, esto implica,

$$Lim_{n\to\infty} \; E\{\hat{\theta}\} = \theta \tag{3.3}$$

$$Lim_{n\to\infty} \; \{\sigma^2_{\hat{\theta}}\} = 0 \tag{3.4}$$

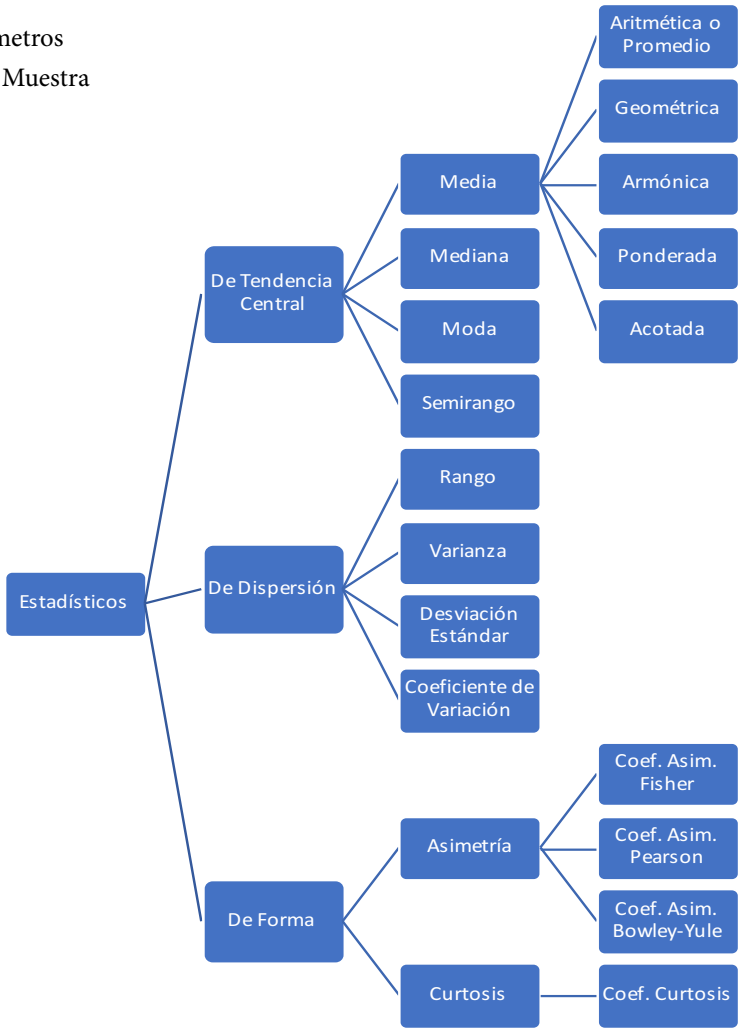
d. Robustez. El estimador $\hat{\theta}_1$ será un estimador robusto del parámetro θ si el no cumplimiento de algunas hipótesis previas en las que se basa la estimación (por ejemplo, suponer que la distribución de probabilidad de los datos es normal, o que la muestra fue obtenida aleatoriamente), no altera de manera significativa los resultados que este proporciona.

e. Suficiencia. Se dice que un estimador es suficiente cuando resume toda la información relevante contenida en la muestra, de forma que ningún otro estimador pueda proporcionar información adicional sobre el parámetro desconocido de la población.

f. Invariancia. Se dice que un estimador es invariante cuando el estimador de la función del parámetro coincide con la función del estimador del parámetro.

En este subtema se definirán los estadísticos muestrales, que permiten estimar a su vez parámetros poblacionales de interés. Dichos estadísticos se pueden clasificar de la forma en que se muestra en la figura 3.10.

Figura 3.10. Parámetros Estadísticos de una Muestra



3.2.2. Medidas de tendencia central

Media aritmética o promedio

Sea una muestra aleatoria con n unidades muestrales, representada de la siguiente forma:

$$Sn = \{x_1, x_2, x_3, \cdots, x_n\}$$

Se define el estadístico de tendencia central media aritmética, también conocida como promedio, a la siguiente función matemática:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i = \frac{1}{n}\left[x_1 + x_2 + x_3 + \cdots + x_n\right] \tag{3.5}$$

Para obtener la media aritmética de un conjunto de datos, utilizando Excel, basta con emplear el comando PROMEDIO.

Nótese que ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Media Aritmética = PROMEDIO(2,-2,5,1) = 1.5

De igual forma con la muestra $\{2,0,5,1\}$:

Media Aritmética = PROMEDIO(2,0,5,1) = 2

De la misma forma, si se toman los datos del ejercicio 3.1 y se capturan en una sola columna de Excel, por ejemplo, en las celdas desde A2 (se excluye A1 porque tiene el encabezado) hasta A91, al aplicar el comando siguiente, en cualquier celda libre de excel, se obtiene el resultado:

= PROMEDIO($A$2:$A$91) = 89.475556

Por el Teorema de Aditividad de la Distribución Normal y por el Teorema del Límite Central, se sabe que la distribución de probabilidad de la media aritmética es normal, es decir,

$$\bar{x} \approx N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{3.6}$$

La media aritmética es el estimador de la media poblacional por excelencia, de las expresiones 3.5 y 3.6 anteriores se desprende que la media aritmética es un estimador insesgado, el más eficiente, o sea, de mínima variancia y es consistente, como se puede comprobar:

$\mu_{\bar{x}} = E\{\bar{x}\} = \mu_x$ *Insesgabilidad*

$\sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}$ *Eficiencia*

$$Lim_{n\to\infty}\left(\bar{x}\right)=\mu_x$$

$$Lim_{n\to\infty}\left(\frac{\sigma_x^2}{n}\right)=0 \qquad \textit{Consistencia}$$

Propiedades de la media aritmética

i. La suma de las desviaciones con respecto a la media aritmética es cero.

Nótese que para $Sn=\{x_1, x_2, x_3, \cdots, x_n\}$, se definen las desviaciones con respecto a la media aritmética como las diferencias entre los valores x_i y su media aritmética, es decir:

$$Desviaciones=\{x_1-\bar{x}, x_2-\bar{x}, x_3-\bar{x}, \cdots, x_n-\bar{x}\}$$

$$\sum_{i=1}^{i=n}\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}x_i-n\bar{x}=0$$

ii. La media aritmética de los cuadrados de las desviaciones de los valores de la variable, con respecto a una constante cualquiera, se hace mínima cuando dicha constante coincide con la media aritmética.

$$s=\left\{\sum_{i=1}^{i=n}\left(x_i-u\right)^2\right\}$$

$$\frac{ds}{du}=-2\sum_{i=1}^{i=n}\left(x_i-u\right)=0$$

$$\sum_{i=1}^{i=n}x_i-nu=0$$

$$u=\frac{1}{n}\sum_{i=1}^{i=n}x_i=\bar{x}$$

De la misma forma se demuestran todas las demás.

iii. Si a todos los valores de la variable se le suma una misma cantidad, la media aritmética queda aumentada en dicha cantidad.
iv. Si todos los valores de la variable se multiplican por una misma constante, la media aritmética queda multiplicada por dicha constante.
v. La media aritmética de un conjunto de números positivos siempre es igual o superior a la media geométrica.
vi. La media aritmética está comprendida entre el valor máximo y el valor mínimo del conjunto de datos.
vii. La media es el centro de gravedad de la distribución de la variable.
viii. La media del producto de una constante k por una variable x es igual al producto de la constante k por la media de la variable dada.
ix. La media de la suma de una constante entera k con una variable x es igual a la suma de la constante k con la media de la variable dada.
x. La media está influenciada por los valores de cada uno de los datos y principalmente se afecta por los valores extremos.
xi. La media no tiene por qué ser igual a uno de los valores de los datos, ni siquiera de su misma naturaleza: datos enteros pueden tener una media decimal.

Media geométrica

Sea una muestra aleatoria con n unidades muestrales representada de la siguiente forma:

$$Sn = \{x_1, x_2, x_3, \cdots, x_n\}$$

Se define el estadístico de tendencia central media geométrica a la siguiente función matemática:

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{i=n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} \tag{3.7}$$

La media aritmética es el valor de tendencia central por excelencia para estimar la media poblacional. Sin embargo, es aplicable solo cuando todos los datos

tienen la misma posibilidad de ser elegidos (equiprobables) y su crecimiento a lo largo del tiempo es lineal. En el caso de que los datos tengan otro tipo de tendencia, la media aritmética no es lo más adecuado. Por ejemplo, si los datos se refieren a número de pobladores en cierto tiempo, se sabe que el crecimiento poblacional es geométrico, no lineal. En este caso, lo más adecuado para calcular la tendencia central es utilizar el concepto de media geométrica.

Para ilustrar lo anterior, según la publicación "Población total por entidad federativa según sexo, 2000, 2005 y 2010" del Instituto Nacional de Estadística y Geografía, edición 2010, en el 2000 había 8,605,239 habitantes en la ciudad de México; para el 2010 había 8,851,08.

Pregunta: ¿cuántos había en el 2005?

Para responder a la pregunta puede ocurrirse calcular un promedio: 8,728,159.5 habitantes. Con el concepto de media geométrica el resultado es: 8,727,293.9 habitantes. El valor que proporciona INEGI es 8,720,916; como se puede apreciar, en el caso de crecimiento poblacional el concepto de media geométrica se acerca más al valor real que el concepto de media aritmética.

El principal problema, del uso de esta definición, desde el punto de vista matemático, es que si n es par y si existen valores de x negativos, al hacer el cálculo, si el argumento es negativo, la raíz par de un número negativo no existiría en los reales. También, el concepto de media geométrica vuelve a fallar, si alguno o algunos de los valores son cero, automáticamente el resultado sería cero.

Nótese qué ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Media Geométrica = MEDIA.GEOM(2,-2,5,1) = #¡NUM!

De igual forma con la muestra $\{2,0,5,1\}$:

Media Geométrica = MEDIA.GEOM(2,-2,0,5,1) = #¡NUM!

En Excel, para calcular la media geométrica, tomando los datos del ejercicio 3.1 y con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando:

MEDIA.GEOM($A$2:$A$91) = 89.380617

Media armónica

Sea una muestra aleatoria con n unidades muestrales representada de la siguiente forma:

$$Sn = \{x_1, x_2, x_3, \cdots, x_n\}$$

Se define el estadístico de tendencia central media armónica, a la siguiente función matemática:

$$x_{armo} = \frac{n}{\sum_{i=1}^{i=n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} \tag{3.8}$$

De la misma forma que para el caso de la media geométrica, hay ciertos valores que por sus unidades de medida se puede saber que la media aritmética no es lo más recomendable para estimar la media de la población. En todos aquellos casos donde el conjunto de datos de la muestra tiene unidades de rapidez de cambio, el concepto de tendencia central más adecuado es el de media armónica. Al decir unidades de rapidez de cambio se refiere a, por ejemplo, cambios en la posición de un cuerpo con respecto al tiempo, o sea velocidad, como metros sobre segundo; cambios en la velocidad con respecto al tiempo, o sea, aceleración, como metros sobre segundo al cuadrado; cambios en el área o superficie de un cuerpo por unidad de tiempo, como metros cuadrados sobre segundo; cambios en el volumen por unidad de tiempo, como metros cúbicos sobre segundo; etcétera.

Para ilustrar lo anterior, considérese el caso hipotético de un tren japonés que tiene que cruzar por tres ciudades situadas a 100 Km cada una de ellas y regresar a la ciudad de inicio, haciendo un circuito cuadrado sin parar en ninguna de las ciudades. Como se muestra en la figura 3.2, el primer recorrido, de la ciudad *A* a la *B*, lo hace a 100 Km/hora; el segundo recorrido, de la ciudad *B* a la *C*, lo hace a 200 Km/hora; el tercer recorrido, de la ciudad *C* a la *D*, lo hace a 300 Km/hora y el cuarto recorrido de la ciudad *D* a *A* lo efectúa a 400 Km/hora. La pregunta sería ¿a qué velocidad promedio recorrió los 400 Km totales?

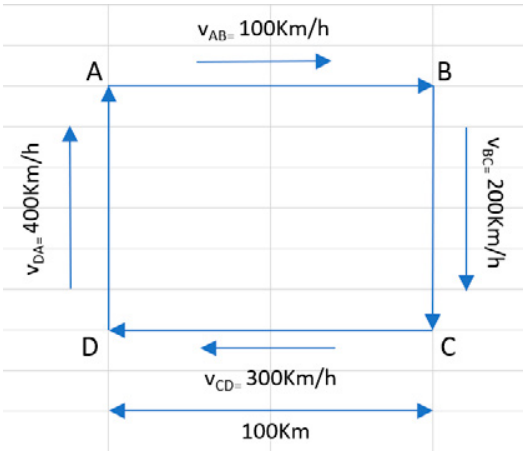


Figura 3.11.

Si se utiliza el concepto de media aritmética o promedio, el recorrido total lo haría en *vprom* = (100 + 200 + 300 + 400)/4 = 250 Km/hora. Usando el concepto de media geométrica el recorrido lo haría a una velocidad de *vgeom* = (100*200 *300*400)^(1/4) = 221.3364 Km/hora. Si se aplica el concepto de media armónica, el recorrido lo haría a *varmo* = 4/(1/100 + 1/200 + 1/300 + 1/400) = 192 Km/hora. Nótese que este concepto nace de la definición de rapidez media como distancia recorrida entre tiempo total de viaje, es decir, velocidad media = 400/(1 + 1/2 + 1/3 + 1/4) = 192 Km/hora.

Como se puede apreciar, los dos primeros conceptos de media aritmética y de media geométrica no reflejan una tendencia central de los datos.

El concepto de media armónica pierde utilidad si alguno de los valores de la muestra es cero, sería imposible calcularlo.

Nótese qué ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Media Armónica = MEDIA.ARMO(2,-2,5,1) = #¡NUM!

De igual forma con la muestra $\{2,0,5,1\}$:

Media Armónica = MEDIA.ARMO(2,-2,0,5,1) = #¡NUM!

De la misma forma, para calcular la media armónica, tomando los datos del ejercicio 3.1 y con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando:

MEDIA.ARMO($A$2:$A$91) = 89.286323

Una aplicación muy común del concepto de media armónica es para el cálculo de la media de un conjunto de fracciones $p_1, p_2, p_3, \cdots, p_n$; donde $0 < p_i < 1$, para cualquier valor de $i = 1, 2, 3, \cdots, n$. El valor de tendencia central más adecuado en este caso es el concepto de media armónica:

$$p_{\text{media}} = n\left(1/p_1 + 1/p_2 + 1/p_3 + \cdots + 1/p_n\right) \tag{3.9}$$

Por ejemplo, suponga que durante una jornada de trabajo de ocho horas, cada hora se toma una muestra de piezas moldeadas a las cuales se les mide su resistencia mecánica, utilizando una prueba de impacto Izod (ver referencia https://es.wikipedia.org/wiki/Ensayo_de_Izod).A las piezas que no soportan la resistencia al impacto se les clasifica como defectuosas y en cada muestra se divide el número de piezas defectuosas entre el total de piezas probadas, obteniéndose los siguientes valores de fracciones defectuosas: $\{0.0255, 0.0242, 0.0265, 0.0277, 0.0227, 0.0209, 0.0285, 0.02345\}$. Obtenga la media de fracciones defectuosas a lo largo del proceso.

Dado que se trata de fracciones se utiliza el concepto de media armónica:

Fracción defectuosa promedio = MEDIA.ARMO(0.0255, 0.0242, 0.0265, 0.0277, 0.0227, 0.0209, 0.0285, 0.02345) = 0.0247

Media ponderada

Si los datos de la muestra se encuentran agrupados en una tabla de frecuencias, como la que se muestra más adelante en la figura 3.2, se define el estadístico de tendencia central media ponderada a la siguiente función matemática:

$$\bar{x}_{ponde} = \frac{1}{n}\sum_{j=1}^{j=m} f_j t_j = \sum_{j=1}^{j=m} f_j^* t_j = \left[f_1^* t_1 + f_2^* + f_3^* + \cdots + f_m^*\right] \tag{3.10}$$

Donde

$f_j^* = \dfrac{f_j}{n}$ *Frecuencia relativa*

t_j *Lectura o marca de clase*

Ejercicio 3.3

Para ilustrar la aplicación de este concepto suponga que se cuenta con la siguiente tabla de frecuencias, obtenida de Las Estadísticas del Personal Académico de la UNAM 2015, página 19, de la Dirección General de Asuntos del Personal Académico de la UNAM (Fuente: http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2015.pdf)

Personal académico en la UNAM por edad												
Universo	Edad (años cumplidos)											Total
	Hasta 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 o más	
UNAM	896	2,356	3,539	4,457	4,924	4,847	5,288	5,106	3,712	2,401	1,822	39,348

Para calcular el promedio de edad del personal académico de la UNAM, debe utilizarse el concepto de media ponderada, multiplicando la marca de clase de cada subintervalo por la frecuencia de cada subintervalo. En este caso, se toma como marca de clase el punto medio de cada subintervalo, es decir, $\{22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72\}$, las frecuencias serían $\{896, 2356, 3539, 4457, 4924, 4847, 5288, 5106, 3712, 2401, 1822\}$. Nótese que para los extremos se toma como marca de clase siguiendo la secuencia del punto medio cada cinco unidades.

De esta forma, el promedio de edad del personal académico de la UNAM en el 2015 está dado por:

$$\text{Media Ponderada} = (22*896 + 27*2356 + 32*3539 + \cdots + 72*1822)/39348 = \\ = 47.8884 \text{ años de edad}$$

Los cálculos se muestran en la siguiente tabla de Excel:

t (Marca de Clase)	f (frecuencia)	$t*f$
22	896	19712
27	2356	63612
32	3539	113248
37	4457	164909
42	4924	206808
47	4847	227809
52	5288	274976
57	5106	291042

t (Marca de Clase)	*f* (frecuencia)	*t* * *f*
62	3712	230144
67	2401	160867
72	1822	131184
Suma =	39348	1884311
		Promedio = 47.8884

Mediana

Es una medida de tendencia central cuyo valor parte por la mitad al conjunto de datos ordenados $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$, el 50% de los datos cae a la izquierda de la mediana y el otro 50% de los datos cae a su derecha. Para obtener la mediana de un conjunto de datos, estos se ordenan de menor a mayor o de mayor a menor y se toma el centro de estos datos ordenados. En el caso de que en $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$ n sea impar, se toma el valor que cae en medio después de ser ordenados, este valor pertenece a la muestra. En el caso de que n sea par, se toma el promedio de los dos valores que caen al centro después de ser ordenados y este promedio no pertenece a la muestra.

Para ilustrar lo anterior, suponga que se generan dos muestras aleatorias, cuyos valores están entre 25 y 50; para producir estas dos muestras se generan 15 valores utilizando el comando de Excel

= aleatorio.entre(25, 50)

Obteniéndose las siguientes dos muestras:

$S_1 = \{39, 46, 36, 25, 47, 35, 25\}$
$S_2 = \{38, 30, 50, 49, 43, 34, 50, 36\}$

Para calcular la mediana de ambas muestras se ordenan sus unidades muestrales de menor a mayor o de mayor a menor

$S_1 = \{\cancel{25}, \cancel{25}, \cancel{35}, 36, \cancel{39}, \cancel{46}, \cancel{47}\}$
$S_2 = \{\cancel{50}, \cancel{50}, \cancel{49}, 43, 38, \cancel{36}, \cancel{34}, \cancel{30}\}$

Tachando sucesivamente los extremos, se puede apreciar que la mediana de cada una de las muestras es:

$M_{eS1} = 36$

$M_{eS2} = (43 + 38)/2 = 40.5$

Nótese que cuando *n* es impar, la mediana es un valor de la muestra, en cambio, cuando *n* es par, la mediana no pertenece a la muestra. Pertenezca o no a la muestra, lo cierto es que el valor de la mediana parte por la mitad a los datos.

La razón de que exista la mediana de una muestra como estimador de la media poblacional se debe a que la media aritmética o promedio se afecta fuertemente por los valores extremos de la muestra, en cambio, la mediana no. Para ejemplificar lo anterior, suponga que se obtiene una muestra de cinco datos del gasto que se realiza en promedio al visitar un tianguis, por ejemplo, $\{200, 250, 275, 300, 300\}$, el promedio en este caso es $(200 + 250 + 275 + 300 + 300)/5 = 265$ y la mediana es 275. Ahora suponga que se sustituye el gasto mayor por una cantidad más grande $\{200, 250, 275, 300, 10000\}$, nótese que la media aritmética es $(200 + 250 + 275 + 300 + 10000)/5 = 2205$, la cual se afectó fuertemente al cambiar un solo dato extremo, en cambio, la mediana sigue siendo la misma 275. Cabe remarcar que en el tianguis lo que pudo ocurrir es que llegó esporádicamente una señora que gastó diez mil pesos, pero esto no es común, y no representa al gasto que en promedio se tiene en un tianguis. Por ello, para este ejemplo, la mediana es un indicador más ajustado a la media poblacional que la misma media aritmética.

Si se toman los datos del ejercicio 3.1, así como de las referencias a las celdas donde se encuentran, para calcular la mediana de los datos se utiliza el siguiente comando y se muestra el resultado:

Mediana = MEDIANA($A$2:$A$91) = 89.3

Moda

Es una medida de tendencia central cuyo valor en la muestra es el que más se repite, o sea el más frecuente, de un conjunto de datos $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$. Para ilustrar lo anterior, observe las siguientes tres muestras:

$S_1 = \{ 200, 250, 250, 275, 275, 300, 300, 300\}$
$S_2 = \{200, 200, 250, 250, 275, 275, 300, 300\}$
$S_3 = \{200, 225, 250, 275, 300, 325, 350\}$

Nótese que en el primer caso, el valor que más se repite es 300, por lo cual este representa a la moda de S_1. En el segundo caso, existen cuatro valores que se repiten igual número de veces, por lo cual se tiene una curva multimodal y existen cuatro modas diferentes. Para evitar la duplicidad, si existen varios valores de moda, se dice por convención que no existe moda. Cuando la moda se calcula a través de Excel, usando el comando " = moda(datos)", si existe repetición de un valor que representa a la moda, lo que hace Excel es tomar el primer valor que se encuentre de izquierda a derecha como moda; para la segunda muestra S2, Excel arrojaría el valor de 200. En la tercera muestra dada, cualquiera de los valores puede representar a la moda, ya que en todos los casos la frecuencia de aparición es uno, pero de acuerdo con la anterior convención usada en la segunda muestra, se diría que no existe moda.

Si se toman los datos del ejercicio 3.1, así como de las referencias a las celdas donde se encuentran, para calcular la moda de los datos se utiliza el siguiente comando y se muestra el resultado:

Moda = MODA($A$2:$A$91) = 87.3

Semirango

Se define el semirango de una muestra $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$, como el valor promedio entre la observación o lectura más grande $x_{\max}$ y la observación o lectura más pequeña $x_{\min}$, es decir,

$$SR = \frac{x_{\max} + x_{\min}}{2} \tag{3.11}$$

Para la muestra $S_1 = \{ 200, 250, 250, 275, 275, 300, 300, 300\}$, el semirango sería $SR_{S1} = (200 + 300)/2 = 250$.

Si se toman los datos del ejercicio 3.1, así como de las referencias a las celdas donde se encuentran, para calcular la moda de los datos se utiliza el siguiente comando y se muestra el resultado:

Semirango = (MAX($A$2:$A$91) + MIN($A$2:$A$91))/2 = 90.3

3.2.3. Medidas de dispersión

Las medidas de dispersión son indicadores que proporcionan información sobre el grado de variabilidad de una muestra o de una variable.

El rango o amplitud R de una muestra

Se define como la máxima dispersión de un conjunto de datos $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$. Para calcular el rango es necesario determinar la observación o lectura más grande x_{max} y la observación o lectura más pequeña x_{min} y calcular la diferencia entre ellas:

$$R = x_{max} - x_{min} \tag{3.12}$$

Para la muestra $S_1 = \{200, 250, 250, 275, 275, 300, 300, 300\}$, el rango sería $R = 300 - 200 = 100$.

Si se toman los datos del ejercicio 3.1, así como de las referencias a las celdas donde se encuentran, para calcular el rango de los datos se utiliza el siguiente comando y se muestra el resultado:

Rango = MAX($A$2:$A$91)–MIN($A$2:$A$91) = 15.4

Desviación o sesgo

Se define una desviación o sesgo como la diferencia entre una lectura x_i y un valor que puede ser el origen o alguno de los estimadores muestrales de tendencia central anteriormente definidos. Por ejemplo, para la muestra $S_1 = \{200, 250, 250, 275, 275, 300, 300, 300\}$, la media aritmética sería 268.75, la mediana sería 275 y la moda sería 300, por lo que las respectivas desviaciones serían:

Desviaciones con respecto a la media: $\{-68.75, -18.75, -18.75, 6.25, 6.25, 32.25, 32.25, 32.25\}$. Nótese que para el caso de la desviación con respecto a la media, si se suman las desviaciones obtenidas, el resultado es cero.

Desviaciones con respecto a la mediana: $\{-75, -25, -25, 0, 0, 25, 25, 25\}$, nótese que en este caso la suma de las desviaciones no es cero.

Desviaciones con respecto a la moda: $\{-100, -50, -50, -25, -25, 0, 0, 0\}$. La suma de las desviaciones no es cero.

Desviación promedio

Para el caso de las desviaciones con respecto a la media, la suma se cancela porque existen valores tanto positivos como negativos que se tienden a cancelar. Una manera de evitar que las desviaciones se cancelen es tomar el valor absoluto de cada una de ellas, de tal manera que se puede calcular la desviación promedio de la siguiente forma:

$$Desv\ Prom = \frac{1}{n}\sum_{i=1}^{i=n} |x_i - \bar{x}| \tag{3.13}$$

El problema que presenta esta expresión matemática es que contempla a la función valor absoluto, y esta función no es derivable en todos los puntos donde $x = x_i$; en muchas ocasiones lo que se busca es minimizar a la función desviación promedio, pero con esta fórmula no sería posible. Recuerde de sus clases de cálculo diferencial la función $y = |x|$ y la función $y = x^2$, son simétricas, cóncavas hacia arriba y tienen forma aproximada para valores cercanos a cero; la principal diferencia entre ellas es que la primera no es derivable en $x = 0$ y la segunda sí es derivable en $x = 0$, como se muestra en la figura 3.12.

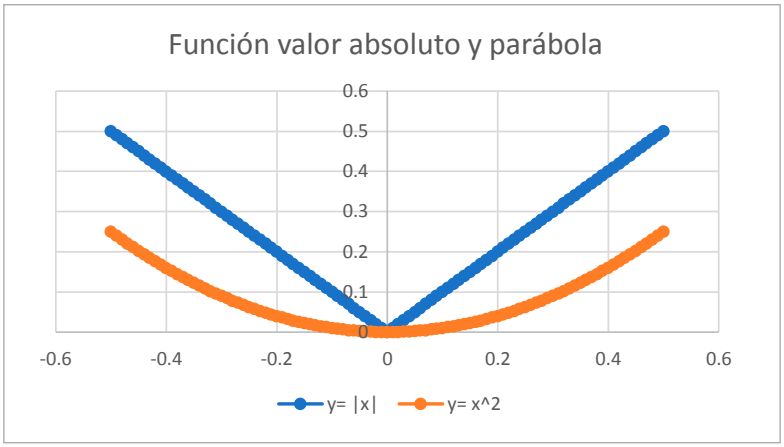


Figura 3.12

Si se toman los datos del ejercicio 3.1, así como de las referencias a las celdas donde se encuentran, para calcular la desviación promedio de los datos se utiliza el siguiente comando y se muestra el resultado:

Desviación promedio = DesvProm($A$2:$A$91) = 3.517235

Varianza o variancia

Se calcula la varianza o variancia de un conjunto de datos con la siguiente expresión:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2 \tag{3.14}$$

Otra expresión que se usa para estimar a la varianza poblacional a partir de los datos de una muestra es la siguiente:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2 \tag{3.15}$$

Cabe remarcar que la expresión (3.15) anterior es más adecuada que la (3.14), debido a que la (3.15) representa a un estimador insesgado de la varianza poblacional, ya que:

$$E\left\{S_{n-1}^2\right\} = \sigma_x^2$$

$$E\left\{S_n^2\right\} = \left(\frac{n-1}{n}\right)\sigma_x^2$$

Nótese qué ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Varianza = Var (2, -2, 5, 1) = 8.333333

De igual forma con la muestra $\{2,0,5,1\}$:

Varianza = Var(2,0,5,1) = 4.666667

De la misma forma, para calcular la varianza, tomando los datos del ejercicio 3.1 y con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando, mostrándose su resultado:

VAR($A$2:$A$91) = 17.287036

Desviación estándar

Dado que cuando se pretende minimizar la dispersión o variación de los datos no es posible usar el concepto de desviación promedio, se define la desviación estándar de los datos de una muestra como la desviación media de dichos datos y para calcularla se obtiene la raíz cuadrada de la varianza muestral, es decir,

$$S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2} \tag{3.16}$$

Para calcular la desviación estándar, si ya se tiene la varianza, basta con obtener su raíz cuadrada.

Nótese qué ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Desviación estándar = DesvEst(2,-2,5,1) = 2.886751

Que como se puede comprobar es la raíz cuadrada de la varianza calculada antes.

De igual forma con la muestra $\{2,0,5,1\}$:

Desviación Estándar = DesvEst(2,0,5,1) = 2.160247

Para calcular la desviación estándar de los datos del ejercicio 3.1 y con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando, mostrándose su resultado:

DESVEST($A$2:$A$91) = 4.157768

Coeficiente de variación de una muestra

Se define el coeficiente de variación de una muestra como el cociente entre la desviación estándar muestral y la media aritmética muestral, es decir,

$$CV = \frac{S_{n-1}}{\bar{x}} \tag{3.17}$$

Nótese qué ocurre en la siguiente muestra: $\{2, -2, 5, 1\}$, utilizando Excel:

Coeficiente de Variación = DESVEST(2,-2,5,1)/PROMEDIO(2,-2,5,1) = 1.9245

De igual forma con la muestra $\{2,0,5,1\}$:

Coeficiente de variación = DESVEST(2,0,5,1)/PROMEDIO(2,0,5,1) = 1.080123

Para calcular el coeficiente de variación de los datos del ejercicio 3.1 y con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando, mostrándose su resultado:

DESVEST($A$2:$A$91)/PROMEDIO($A$2:$A$91) = 0.046468

3.2.4. Momentos de orden k con respecto al origen y con respecto a la media

Los momentos muestrales son estimadores que caracterizan a una muestra aleatoria para aproximarse a los momentos de la distribución de probabilidad de una población, los cuales tienen la propiedad de que dos distribuciones de probabilidad son iguales si tienen todos sus momentos iguales. Los momentos muestrales de orden k con respecto al origen, m'_k, y los momentos muestrales de orden k con respecto a la media aritmética, m_k, se obtienen a través de las siguientes expresiones matemáticas:

$$m'_k = \frac{1}{n}\sum_{i=1}^{i=n} x_i^k \tag{3.18}$$

$$m_k = \frac{1}{n}\sum_{i=l}^{i=n}\left(x_i - \bar{x}\right)^k \tag{3.19}$$

Como se puede apreciar, al momento de orden $k = 1$ con respecto al origen se le conoce como media aritmética. El momento de orden $k = 1$ con respecto a la media es cero. El momento de segundo orden con respecto a la media tiene una relación directa con el concepto de varianza:

$$m_2 = \frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2 = \left(\frac{n-1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2 = \left(\frac{n-1}{n}\right)S^2_{n-1} \tag{3.20}$$

Nótese qué ocurre en la muestra $\{2, -2, 5, 1\}$ al calcular los momentos de orden 1, 2, 3, 4 y 5 con respecto a la media, utilizando Excel:

xi	(xi–media)^2	(xi–media)^3	(xi–media)^4	(xi–media)^5
2	0.25	0.125	0.0625	0.03125
–2	12.25	–42.875	150.0625	–525.2188
5	12.25	42.875	150.0625	525.2188
1	0.25	–0.125	0.0625	–0.03125
1.5	6.25	0	75.0625	0

Momento de orden uno con respecto al origen = 1.5

Momento de orden dos con respecto a la media = 6.25

Momento de orden tres con respecto a la media = 0

Momento de orden cuatro con respecto a la media = 75.0625

Momento de orden cinco con respecto a la media = 0

Para calcular los momentos de orden 1, 2, 3, 4 y 5, con respecto a la media de los datos del ejercicio 3.1 con las celdas de referencia A2:A91, se hace el mismo procedimiento que el ejemplo de arriba:

Momento de orden uno con respecto al origen = 89.4756

Momento de orden dos con respecto a la media = 17.0950

Momento de orden tres con respecto a la media = 17.7507

Momento de orden cuatro con respecto a la media = 580.8263

Momento de orden cinco con respecto a la media = 1,412.6491

3.2.5. Cuantiles o fractiles muestrales

Los cuantiles o fractiles muestrales son puntos tomados a intervalos regulares de la abscisa, de la función de distribución de frecuencias empírica de una muestra. El cuantil x_p de orden p de una distribución, con $0 < p < 1$, es el valor de la variable que se está midiendo u observando en una muestra que marca una segmentación de modo que una proporción p de valores de la muestra es menor o igual que x_p, y su complemento $1–p$ en x es mayor que este valor.

Los cuantiles más usados son:

i. Los cuartiles, que dividen a la distribución de frecuencias empíricas en cuatro partes, correspondiendo a los cuantiles 0.25; 0.50 y 0.75.

ii. Los quintiles, que dividen a la distribución en cinco partes correspondiendo a los cuantiles 0.20; 0.40; 0.60 y 0.80.

iii. Los deciles, que dividen a la distribución en diez partes;

iv. Los percentiles, que dividen a la distribución en cien partes.

En las distribuciones muestrales, donde se obtiene una muestra $Sn = \{x_1, x_2, x_3, \cdots, x_n\}$, para el cálculo de los cuantiles, las partes en que se divide el rango de los datos solo pueden ser aproximadamente iguales. Por desgracia, no hay consenso sobre cómo realizar esta aproximación; en la literatura científica que describe el software R existen hasta nueve métodos diferentes, que conducen a resultados diferentes. Por ello, al calcular cualquier cuantil de datos no agrupados por medio de calculadora, software o manualmente, es básico saber e indicar el método utilizado.

Para calcular los cuantiles de los datos del ejercicio 3.1, con las referencias a las celdas donde se encuentran, en Excel se utiliza el siguiente comando:

CUARTIL(Datos, r) donde r = 1, 2, 3, primero, segundo y tercer cuartil, es decir,

Primer cuartil = CUARTIL($A$2:$A$91, 1) = 86.175

Segundo cuartil = CUARTIL($A$2:$A$91, 2) = 89.25

Tercer cuartil = CUARTIL($A$2:$A$91, 3) = 93.1

En el software R se utiliza el comando:

- quantile(Datos, probs = seq(0, 1, 0.25), na.rm = FALSE, names = TRUE, type = 7, ⋯)

Ver la sintáxis del comando en la siguiente dirección: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html

El cual devuelve estimaciones de cuantiles de distribución subyacentes, basados en estadísticas de uno o dos órdenes de los elementos suministrados en Datos, empleando uno de los nueve algoritmos para generar cuantiles analizados en el artículo de Hyndman, R. J. and Fan, Y. (1996), "Sample quantiles in statistical packages", American Statistician 50, 361–365. Se sugiere revisar la siguiente publicación: https://en.wikipedia.org/wiki/Quantile#Estimating_quantiles_from_a_sample

Se utilizará este comando en R para generar los percentiles 12%, 24%, 36%, 48%, 60%, 72%, 84% y 96% de los datos del ejercicio 3.1. Para empezar se le deben dar los datos, a través del siguiente comando:

> datos<-c (94.1,93.2,90.6,91.4,88.2,86.1,95.1,90.0,92.4,87.3,86.6,91.2,86.1,90.4,89.1,87.3,84.1,90.1,95.2,86.1,94.3,93.2,86.7,83.0,95.3,94.1,97.8,93.1,86.4,87.6,94.1,92.1,96.4,88.2,86.4,85.0,84.9,87.3,89.6,90.3,93.1,94.6,96.3,94.7,91.1,92.4,90.6,89.1,88.8,86.4,85.1,84.0,93.7,87.7,90.6,89.4,88.6,84.1,82.6,83.1,84.6,83.6,85.4,89.7,87.6,85.1,89.6,90.0,90.1,94.3,97.3,96.8,94.4,96.1,98.0,85.4,86.6,91.7,87.5,84.2,85.1,90.5,95.6,88.3,84.1,83.7,82.9,87.3,86.4,84.5)

Otra forma de proporcionarle los datos a R es si se tienen en una columna de Excel, guardar el archivo con el nombre datosespesor con el formato csv (del inglés comma-separated values), es decir, datosespesor.csv y para leerlo en R se usa el siguiente comando:

datos<-read.csv(datosespesor.csv, header = TRUE)

Para obtener los cuantiles se utiliza el siguiente comando:

quantile(datos,probs = c(0.12,0.24,0.36,0.48,0.60,0.72,0.84,0.96),type = 7)

En donde type = 7 es el algoritmo siete de los nueve que señala el artículo ya citado de Hyndman, R. J. and Fan, Y. (1996), "Sample quantiles in statistical packages", American Statistician 50, 361–365.

Obteniéndose el siguiente resultado:

12%	24%	36%	48%	60%	72%	84%	96%
84.404	86.100	87.300	89.016	90.340	92.400	94.300	96.576

3.2.6. Medidas de forma (asimetría y curtosis)

Las medidas de asimetría de una muestra son indicadores muestrales que permiten establecer el grado de simetría (o asimetría), que presenta una distribución de frecuencias empírica de una muestra, sin tener que hacer su representación gráfica.

Coeficiente de Asimetría de Fisher

El coeficiente de asimetría de Fisher, representado por γ_1, se define como:

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} \tag{3.21}$$

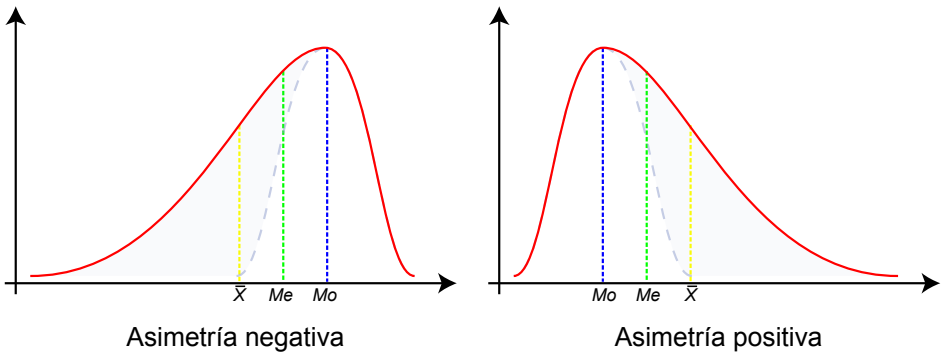
Algunos autores lo definen como:

$$\gamma_1 = \frac{m_3}{S_{n-1}^3} = \frac{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{x})^3}{S_{n-1}^3} \tag{3.22}$$

En donde m_3 es el tercer momento con respecto a la media muestral y S_{n-1} es la desviación estándar muestral.

Si $\gamma_1 < 0$ la forma de la distribución de frecuencias empírica es asimétrica positiva o a la derecha. Si $\gamma_1 > 0$, la distribución es asimétrica negativa o a la izquierda. Si la distribución de frecuencias empírica es simétrica, entonces se sabe que $\gamma_1 = 0$; el recíproco no es cierto: la condición $\gamma_1 = 0$ es necesaria, más no suficiente para asegurar que la curva es simétrica.

Figura 3.12. Tipos de asimetría de un polígono de frecuencias



Asimetría estadística 2007 Recuperado de: https://es.wikipedia.org/wiki/Asimetr%C3%ADa_estad%C3%ADstica

Para la muestra: $\{2, -2, 5, 1\}$, se calculará el coeficiente de asimetría de Fisher usando Excel:

Coeficiente de asimetría de Fisher = COEFICIENTE.ASIMETRIA(2,-2,5,1) = 0

De igual forma con la muestra $\{2,0,5,1\}$:

Coeficiente de asimetría de Fisher = COEFICIENTE.ASIMETRIA(2,0,5,1) = 1.19034

Para calcular el coeficiente de asimetría de Fisher de los datos del ejercicio 3.1 con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando mostrándose su resultado:

COEFICIENTE.ASIMETRIA($A$2:$A$91) = 0.255416

Coeficiente de asimetría de Pearson

Con base en la figura 3.12, el Coeficiente de asimetría de Pearson de una muestra se define como:

$$C_{AP} = \frac{\bar{x} - m_o}{S_{n-1}} \tag{3.23}$$

Se basa en que en distribuciones simétricas la media de la distribución de frecuencias empírica es igual a la moda. Si la distribución es simétrica $C_{AP} = 0$. Si la distribución es asimétrica positiva la media se sitúa a la derecha de la moda y, por tanto, $C_{AP} > 0$; en cambio, si la distribución es asimétrica negativa la moda se sitúa a la derecha de la media y $C_{AP} < 0$.

Para la muestra: $\{2,-2, 5, 1\}$, se calculará el Coeficiente de Asimetría de Pearson usando Excel:

Coeficiente de asimetría de Pearson = COEFICIENTE.ASIMETRIA.P(2,-2,5,1) = 0

De igual forma con la muestra $\{2,0,5,1\}$:

Coeficiente de asimetría de Pearson = COEFICIENTE.ASIMETRIA.P(2,0,5,1) = 0.687243

Para calcular el coeficiente de asimetría de Pearson de los datos del ejercicio 3.1 con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando, mostrándose su resultado:

COEFICIENTE.ASIMETRIA.P($A$2:$A$91) = 0.251139

Coeficiente de Asimetría de Bowley-Yule

Está basado en la posición de los cuartiles de una muestra, y para calcularlo utiliza la siguiente expresión:

$$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1} \tag{3.24}$$

En una distribución simétrica el primer, segundo y tercer cuartil estarán a la misma distancia uno de otro, por lo que $C_{ABY} = 0$. Si la distribución es asimétrica positiva o a la derecha, $C_{ABY} > 0$ y si es asimétrica negativa o a la izquierda, $C_{ABY} < 0$.

Para calcular el Coeficiente de Asimetría de Bowley-Yule de los datos del ejercicio 3.1 con las celdas de referencia A2:A91:

$C_{ABY} = (93.1 + 86.175\text{-}2*89.25) / (93.1\text{-}86.175) = 0.111913$

Medidas de curtosis o aplanamiento

La curtosis (apuntamiento o "picudez") de una distribución de frecuencias empíricas, o su antónimo aplanamiento, es una característica de la forma "alta y esbelta" o "aplanada y amplia" de la distribución de frecuencias empíricas que se puede trazar de una muestra.

El coeficiente de curtosis inicialmente se define como:

$$\beta_2 = \frac{m_4}{m_2^2} \tag{3.25}$$

Dado que matemáticamente se puede demostrar que para una curva normal, el cuarto momento es equivalente a tres veces su varianza y con la finalidad de comparar contra la curva normal estándar, otros autores la definen así:

$$\gamma_2 = \frac{m_4}{m_2^2} - 3 \tag{3.26}$$

Finalmente, otros autores por simplificación la definen así:

$$\gamma_2 = \frac{m_4}{S_{n-1}^4} - 3 \tag{3.27}$$

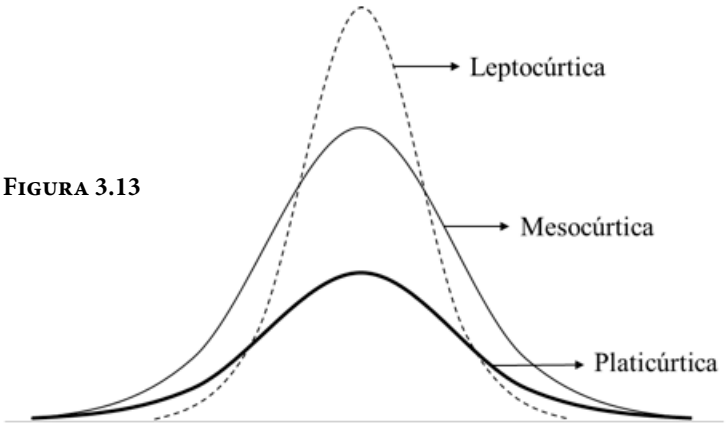


Figura 3.13

Como se puede apreciar en la figura 3.13, tomando la distribución normal como referencia, una distribución puede ser:

i. Leptocúrtica, cuando $\beta_2 > 3$ o $\gamma_2 > 0$, lo que implica que la curva de frecuencias empírica es más alta y esbelta, pero con colas más gruesas que la normal correspondiente.
ii. Platicúrtica, cuando $\beta_2 < 3$ o $\gamma_2 < 0$, lo que implica que la curva de frecuencias empírica es más baja y amplia, pero con colas menos gruesas que la normal correspondiente.
iii. Mesocúrtica, cuando $\beta_2 = 3$ o $\gamma_2 = 0$, lo que implica que la curva de frecuencias empírica tiene una distribución de frecuencias empírica similar a una normal.

Para la muestra: $\{2, -2, 5, 1\}$, se calculará el coeficiente de curtosis usando Excel:

Coeficiente de curtosis = CURTOSIS(2,-2,5,1) = 0.912

De igual forma con la muestra $\{2,0,5,1\}$:

Coeficiente de curtosis = CURTOSIS(2,0,5,1) = 1.5

Para calcular el coeficiente de curtosis de los datos del ejercicio 3.1 con las celdas de referencia A2:A91, basta ubicarse en cualquier celda libre y teclear el siguiente comando mostrándose su resultado:

CURTOSIS($A$2:$A$91) = -1.0013

Existe una variedad numerosa de software para realizar estadística descriptiva, por ejemplo, se pueden citar: Excel, MatLab, Minitab, R, Stata, StatGraphics, SPSS, Wolfram Mathematica, etcétera. Conviene hacer estos cálculos en forma resumida, en vez de hacer el cálculo de concepto por concepto; aunque los resúmenes estadísticos no calculan todo, según el software que se utilice, sí ahorran tiempo. En este volumen se utilizarán de preferencia tres de ellos, el más común de todos es Excel, por su sencillez y por ser el más comercial; otro software gráfico que requiere licencia es Minitab, muy fácil de usar y tiene un ambiente gráfico que lo hace muy amigable; y, el tercero es R, un software gratuito que tiene varias ventajas competitivas, es muy seguro, puede uno bajar de la red el programa fuente y adaptarlo a la medida de sus necesidades, es un lenguaje de desarrollo por lo cual puede programarse y se ha ido enriqueciendo con las aportaciones de todos los usuarios que lo utilizan.

Para ejemplificar cómo se aplican estos conceptos utilizando dicho software, se resolverá el ejemplo 3.1, ya mencionado antes.

Ejercicio 3.1

Se reciben lotes consecutivos de tamaño N = 800 de piezas de sustrato de cerámica a las que se les ha aplicado un revestimiento metálico, mediante un proceso de deposición por vapor. La calidad del revestimiento depende de su grosor en milésimas de pulgada. Para conocer el comportamiento probabilístico del grosor se decidió tomar una muestra aleatoria de tamaño n = 90 y los resultados se muestran a continuación:

94.1	86.6	94.3	94.1	93.1	85.1	84.6	97.3	85.1
93.2	91.2	93.2	92.1	94.6	84.0	83.6	96.8	90.5
90.6	86.1	86.7	96.4	96.3	93.7	85.4	94.4	95.6
91.4	90.4	83.0	88.2	94.7	87.7	89.7	96.1	88.3
88.2	89.1	95.3	86.4	91.1	90.6	87.6	98.0	84.1
86.1	87.3	94.1	85.0	92.4	89.4	85.1	85.4	83.7
95.1	84.1	97.8	84.9	90.6	88.6	89.6	86.6	82.9
90.0	90.1	93.1	87.3	89.1	84.1	90.0	91.7	87.3
92.4	95.2	86.4	89.6	88.8	82.6	90.1	87.5	86.4
87.4	86.1	87.6	90.3	86.4	83.1	94.3	84.2	84.5

Calcular las medidas de tendencia central, las medidas de dispersión, los momentos de orden tres y cuatro con respecto a la media y las medidas de forma (asimetría y curtosis).

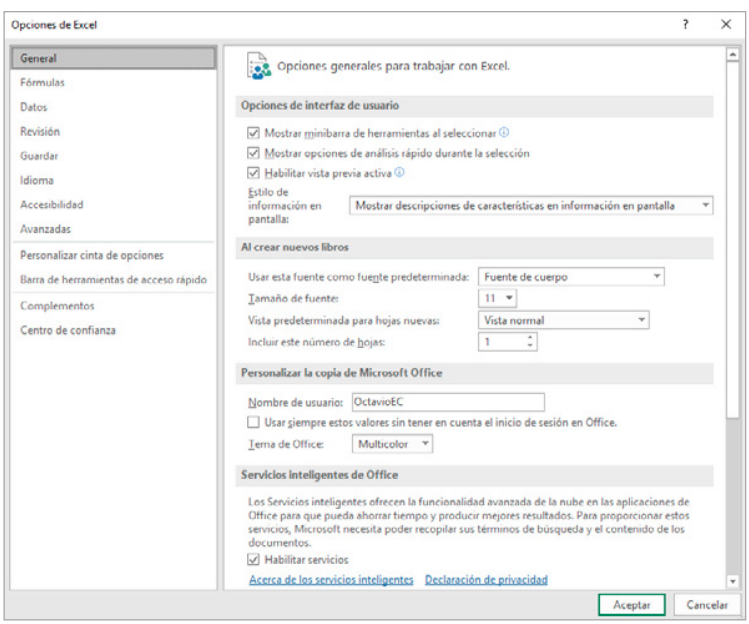
Primero se usará Excel:

a. Capturar los datos en una sola columna de Excel, por ejemplo, en la columna A el encabezado de los datos se llamará espesor, estos datos quedarán capturados en las celdas de la A1 a la A91 y la primera celda de esta columna será el encabezado o rótulo. Se capturarán columna por columna y en ese orden, es decir, primero se capturan los datos de la primera columna, luego los de la segunda y así sucesivamente.

b. En la pantalla principal de Excel, dar click en el menú que dice Datos y totalmente a la derecha debe aparecer un submenú denominado *Análisis de datos.*

   Si no aparece el submenú de *Análisis de datos*, se deberá activarlo, con las siguientes instrucciones: dar un clic en el menú *Archivo*, elija el submenú *Opciones* y aparece la siguiente pantalla:

Figura 3.14.



Elija el menú Complementos en esta pantalla y aparece la siguiente:

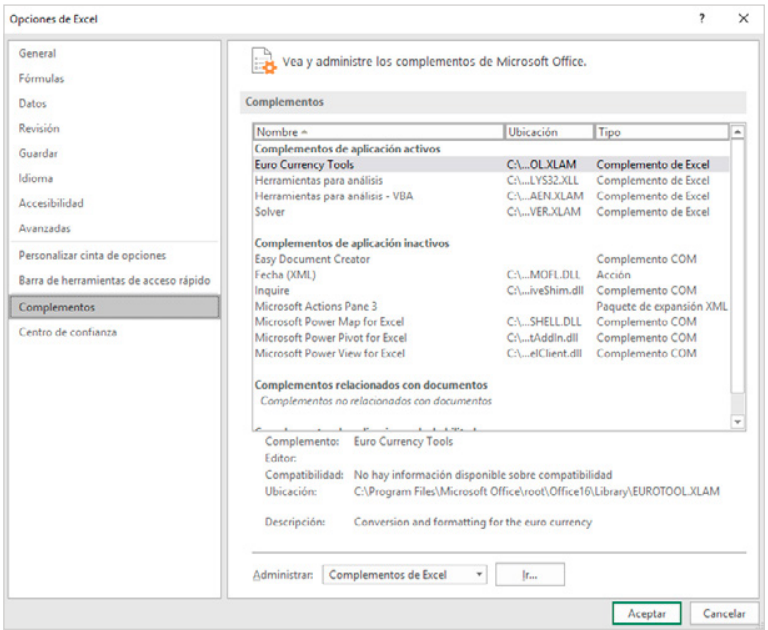


Figura 3.15.

Hasta debajo de esta pantalla aparecen los Complementos de Excel, dé un clic en Ir... y aparece la siguiente pantalla:

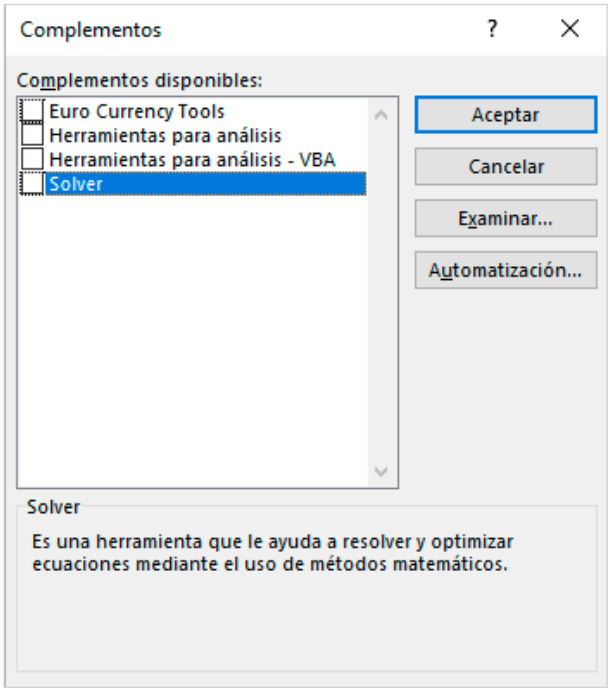


Figura 3.16.

Palomee los complementos que requiera, por lo menos debe activar Herramientas para análisis, quedando:

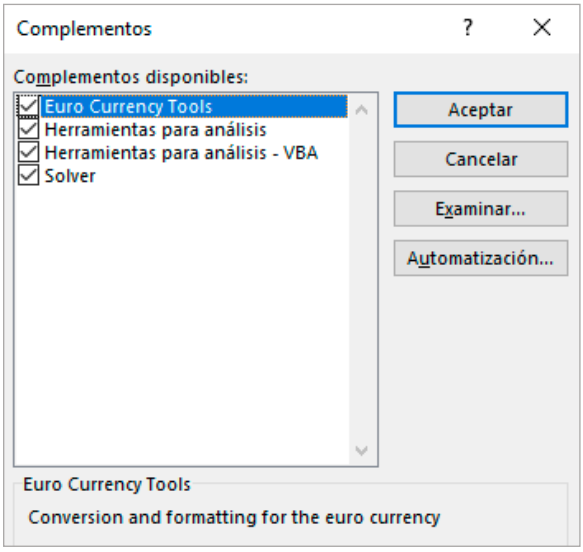
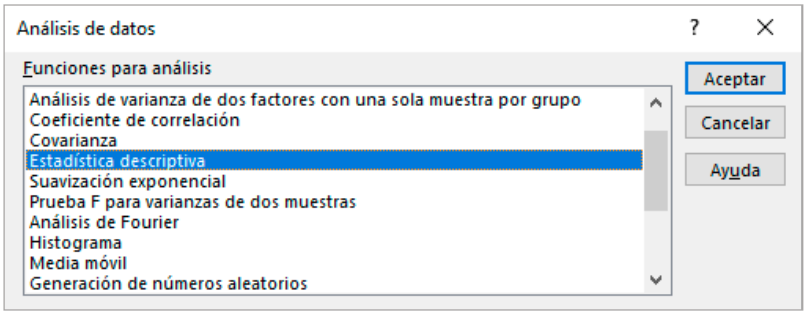


Figura 3.17.

c. Una vez activado el submenú de Análisis de datos, dándole clic aparece la siguiente pantalla:

Figura 3.18.



d. Se elige la opción Estadística descriptiva, apareciendo la pantalla de la Figura 3.19. En esta pantalla, en la Entrada, donde dice Rango de entrada, se capturan las coordenadas donde se encuentran capturados los datos, es decir, $A$1:$A$91 (nótese que se incluye hasta el encabezado), lo que implica que debe darse un click en la ventana que dice Rótulos en la primera fila. También se palomea la ventana que dice Resumen de estadísticas y finalmente se da click en Aceptar.

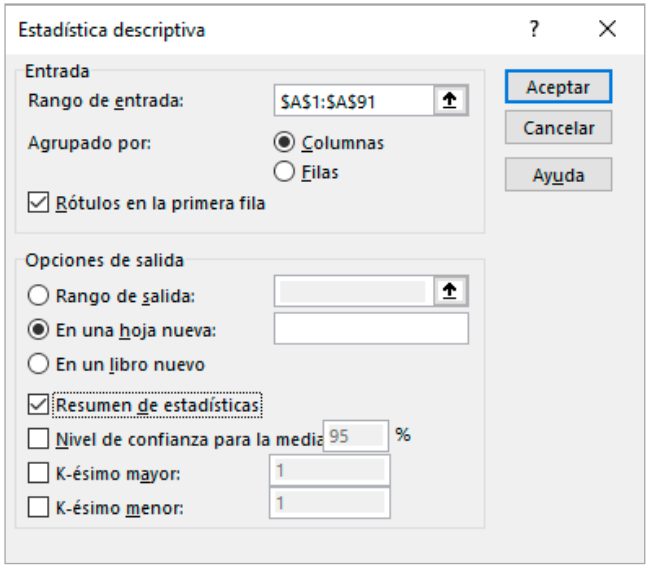


Figura 3.19.

e. Se obtienen los resultados que aparecen en la figura 3.20.

Estadístico	*Valor*
Media	89.47555556
Error típico	0.438267247
Mediana	89.25
Moda	87.3
Desviación estándar	4.157768176
Varianza de la muestra	17.2870362
Curtosis	–1.001322258
Coeficiente de asimetría	0.255415846
Rango	15.4
Mínimo	82.6
Máximo	98
Suma	8052.8
Cuenta	90

Figura 3.20.

Nótese que en este resumen estadístico no aparecen algunos conceptos que se pide calcular, por ejemplo, el concepto de media solo muestra la media aritmética; si se desea calcular la media geométrica a través de Excel, se usa el siguiente comando y se oprime enter:

Media Geométrica = MEDIA.GEOM($A$1:$A$91) = 89.380617

De la misma forma:

= MEDIA.ARMO($A$1:$A$91) = 89.286323

Para obtener el semirango, es necesario primero obtener la máxima y la mínima observación a través del siguiente comando:

Valor máximo = MAX($A$2:$A$91) = 98.0

Valor mínimo = MIN($A$2:$A$91) = 82.6

Semirango = (Valor máximo + Valor mínimo)/2 = 90.3

Para calcular la desviación promedio, en la columna B se calculan las desviaciones con respecto a la media, es decir, la diferencia entre cada valor dado menos su media aritmética, por ejemplo, en la celda B2, se teclea el siguiente comando:

Desviación1 = $A$1-Media aritmética = 4.624444

En la siguiente columna C se calcula el valor absoluto de las desviaciones con respecto a la media a través del siguiente comando:

= ABS(Desviación1) = ABS(B1) = 4.624444

Después se iluminan las celdas B1 y C1, se coloca el cursor en el punto que aparece abajo a la derecha en estas celdas y se dan dos clicks o se arrastra el cursor hasta la celda C91, posteriormente se calcula el promedio de los datos obtenidos en la referencia $C$2:$C$91,

Desviación Promedio = PROMEDIO($C$2:$C$91) = 3.517235

La Desviación Estándar es 4.157768, nótese que la desviación promedio es menor. Esto ocurrirá siempre.

Aunque el cálculo de la desviación promedio obtenido anteriormente se hizo prácticamente manual, existe el comando para hacerlo automáticamente con Excel, simplemente se obtiene:

Desviación Promedio = DesvProm($A$2:$A$91) = 3.517235

El concepto que aparece en el resumen estadístico de la figura 3.10 como Error típico se refiere a la desviación estándar entre la raíz cuadrada del total de datos, compruebe que:

Error típico = 4.157768 / raíz(90) = 0.438267

El coeficiente de variación es:

CV = Desv Est / Media aritmética = 4.157768 / 89.475556 = 0.46468

El coeficiente de asimetría de Fisher es $\gamma_1 = 0.552416 > 0$ positivo; por lo cual se tiene una asimetría negativa, es decir, la distribución de frecuencias empírica carga su media a la derecha de su moda.

El coeficiente de asimetría de Pearson sería:

$C_{AP} = (89.475556-87.3) / 4.157768 = 0.52325$

Para calcular el coeficiente de asimetría de Bowley-Yule primero deben calcularse los cuartiles tercero q_3, segundo q_2 y primero q_1, con la siguiente expresión:

q_k = CUARTIL($A$2:$A$91, k) (3.28)

Donde $A$2:$A$91 es la referencia de las celdas donde se encuentran los datos y k = 1, 2, 3 representa al primero, al segundo o al tercer cuartil:

q_3 = CUARTIL($A$2:$A$91, 3) = 93.1

q_2 = CUARTIL($A$2:$A$91, 2) = 89.25

q_1 = CUARTIL($A$2:$A$91, 1) = 86.175

De tal manera que su coeficiente de asimetría de Bowley-Yule es:

$C_{ABY} = (93.1 + 86.175-2*89.25) / (93.1-86.175) = 0.111913$

La curtosis $\gamma_2 = -1.001322 < 0$, es negativa, por lo que la curva de frecuencias empírica es más plana y amplia que su normal equivalente.

Con todos los cálculos anteriores, el resumen estadístico de Excel quedaría de la forma en la que se muestra en la figura 3.21.

Estadístico	Valor
Media aritmética	89.4755556
Media geométrica	89.380617
Media armónica	89.286323
Semirango	90.3
Mediana	89.25
Moda	87.3
Rango	15.4
Desviación promedio	3.517235
Desviación estándar	4.15776818
Error típico	0.43826725
Varianza de la muestra	17.2870362
Coeficiente de variación	0.46468
Curtosis	-1.0013223
Coef Asimetría Fisher	0.25541585
Coef Asimetría Pearson	0.52325
Tercer cuartil	93.1
Segundo Cuartil	89.25
Primer cuartil	86.175
Coef Asimetría Bowley- Yule	0.111913
Rango	15.4
Máximo	98
Mínimo	82.6
Suma	8,052.8
Cuenta	90

Figura 3.21

Ahora se empleará Minitab:

Los pasos a aplicar con Minitab son los siguientes:

1. Se copia la tabla de datos en una sola columna de Excel en la pantalla principal de Minitab.

2. Se selecciona el menú Stat dando un click en él y luego se escoge el submenú Basic Statistics, posteriormente se selecciona el submenú Display Descriptive Statistics, como se muestra en la figura 3.22.

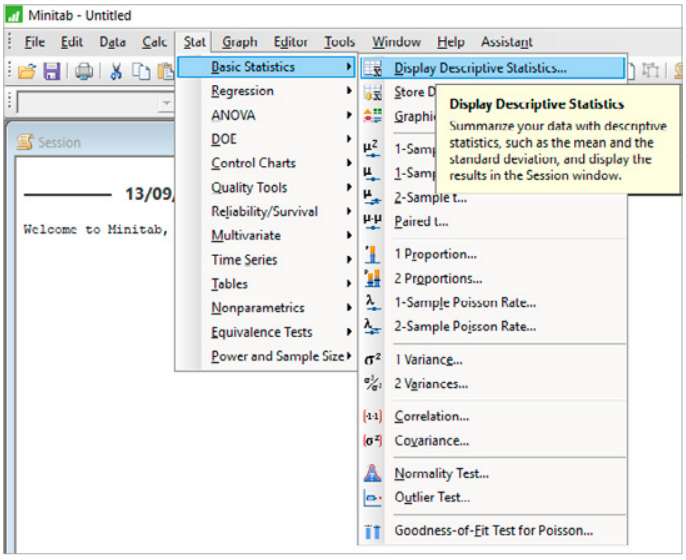


Figura 3.22.

3. Aparece la pantalla de la figura 3.23 en la cual se selecciona la columna donde se copiaron los datos de la muestra del ejercicio 3.1. Al dar un click en el botón que dice Statistics, aparecen los estadísticos que puede calcular Minitab de una muestra, como se observa en la figura 3.24, en la cual con el botón All se seleccionan todos.

4. De la pantalla de la figura 3.23, se selecciona la opción Graphs, apareciendo la pantalla de la figura 3.25, en la cual se muestran las opciones de gráficos que se pueden trazar y también se palomean todas las opciones.

5. Nuevamente en la pantalla de la figura 3.23, se oprime la opción Ok, obteniéndose los resultados que se muestran en la figura 3.26. Los gráficos que aparecen como resultado se interpretarán en el siguiente subtema.

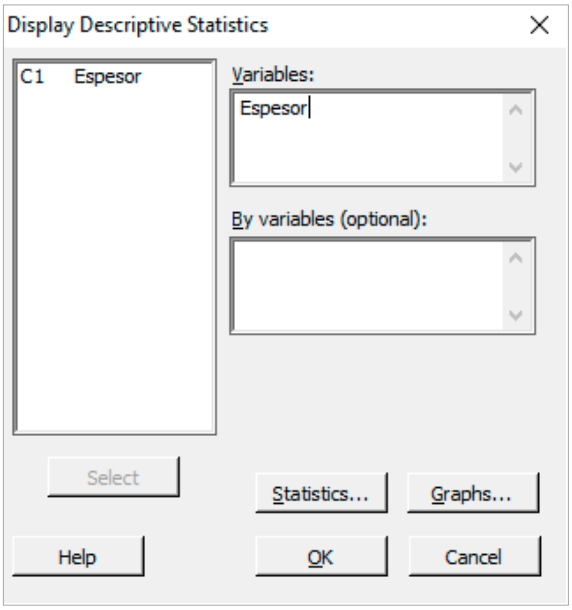


Figura 3.23.

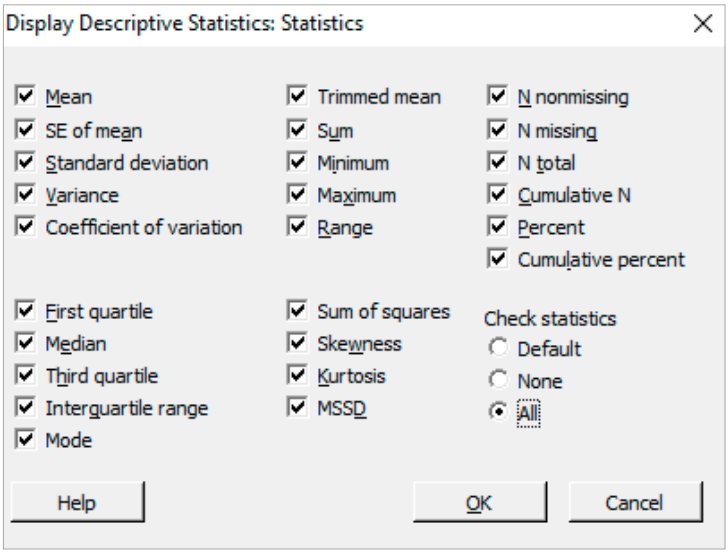


Figura 3.24.

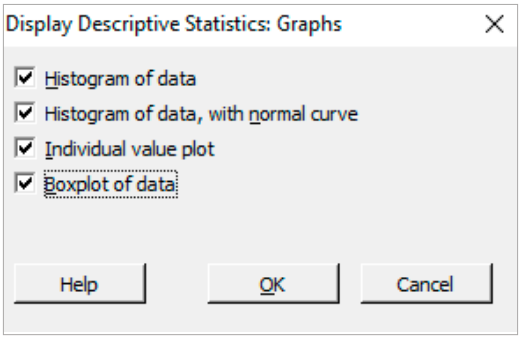


Figura 3.25.

Figura 3.26

```
13/09/2019 08:41:46 a. m.

Welcome to Minitab, press F1 for help.

Descriptive Statistics: Espesor

          Total
Variable  Count   N  N*  CumN  Percent  CumPct    Mean  SE Mean  TrMean  StDev  Variance
Espesor      90  90   0    90      100     100  89.476    0.438  89.391  4.158    17.287


Variable  CoefVar       Sum  Sum of Squares  Minimum      Q1  Median      Q3  Maximum   Range
Espesor      4.65  8052.800       722067.300   82.600  86.100  89.250  93.125   98.000  15.400

                          N for
Variable     IQR        Mode  Mode  Skewness  Kurtosis   MSSD
Espesor    7.025  86.4, 87.3     4      0.26     -1.00  8.792
```

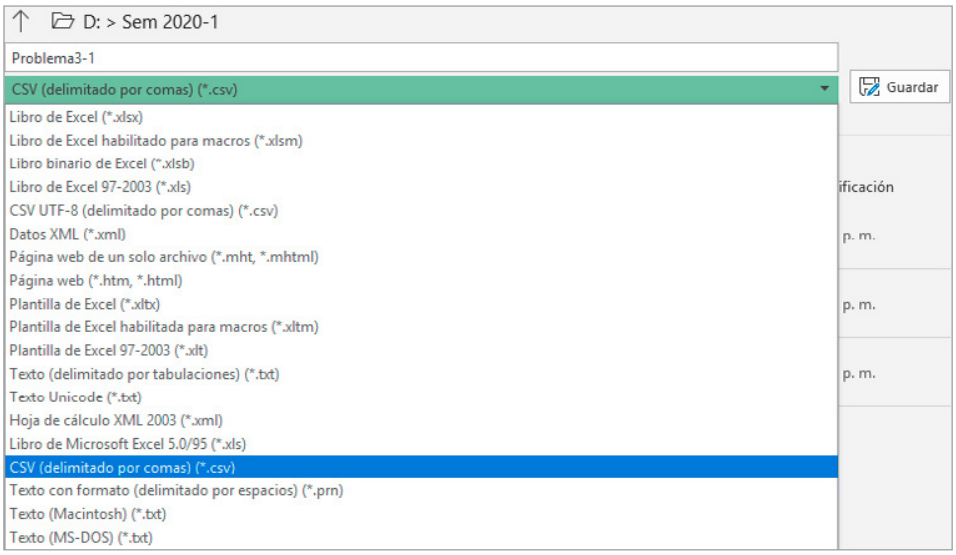
Ahora se empleará R:

Los pasos con R son los siguientes:

1. R no puede leer los datos directamente de Excel, a menos que el archivo de Excel se salve en un formato que R pueda interpretar, afortunadamente sí existe la opción. Copie los datos del Ejercicio 3.1 en un archivo limpio de Excel, que queden en una sola columna y salve dicho archivo en un formato csv con las siguientes instrucciones:

   En Excel abra el menú Guardar como y ponga el nombre del archivo donde copió los datos del Ejercicio 3.1; en mi caso personal le voy a llamar "datos-prob3-1". Luego en la opción donde dice Guardar aparece una pestaña apuntando hacia abajo, selecciónela para que le muestre los formatos o extensión en que puede guardarse el archivo, como se muestra en la figura 3.27. Elija la opción csv (delimitado por comas) (*.csv). Asegúrese que este archivo se guarde en la dirección de trabajo de R, porque si no, R nunca va a encontrar el archivo donde están los datos.

Figura 3.27



2. R es un software muy amigable tan solo con manejar los términos en inglés básico. Para conocer en qué directorio de trabajo opera el software r, teclee el comando getwd(), el cual significa "get work directory", en la dirección que le proporcione tiene que guardar el archivo de Excel que acaba de crear en el paso anterior, o en su defecto cambiar la dirección de trabajo de R con el comando setwd(), "set work directory", dando como argumento la ruta del directorio que quiere usar: setwd("C:\otro_directorio"). Si desea conocer el contenido de su directorio de trabajo, puede ejecutar la función list. files(), que le dará una lista de los archivos dentro del directorio de trabajo.

3. Para abrir el archivo de datos en R, cuando se recurre a la extensión csv, el archivo queda guardado como un vector en donde cada una de sus componentes están separadas por comas y cada una de ellas representa una lectura, por lo cual, deben guardarse estos datos en un vector, al que se le denominará datos, a través del siguiente comando

   - datos < -
     c(datos < read.csv(file = "c:/Usuarios/OctavioEC/Documentos/datos-prob3-1.csv",head = TRUE, sep = ","");

Otra forma de alimentar con datos a R es declarar un vector con todas las lecturas de la muestra separadas por comas, de la siguiente forma:

datos < -c(94.1,93.2,90.6,91.4,88.2,86.1,95.1,90.0,92.4,87.3,86.6,91.2,86.1,90.4,89.1,87.3,84.1,90.1,95.2,86.1,94.3,93.2,86.7,83.0,95.3,94.1,97.8,93.1,86.4,87.6,94.1,92.1,96.4,88.2,86.4,85.0,84.9,87.3,89.6,90.3,93.1,94.6,96.3,94.7,91.1,92.4,90.6,89.1,88.8,86.4,85.1,84.0,93.7,87.7,90.6,89.4,88.6,84.1,82.6,83.1,84.6,83.6,85.4,89.7,87.6,85.1,89.6,90.0,90.1,94.3,97.3,96.8,94.4,96.1,98.0,85.4,86.6,91.7,87.5,84.2,85.1,90.5,95.6,88.3,84.1,83.7,82.9,87.3,86.4,84.5)

Los comandos básicos en R para realizar estadística descriptiva de los datos se muestran en la figura 3.28

Figura 3.28.

Comando	Descripción
summary(datos)	Resumen estadístico
min(datos)	Valor mínimo
max(datos)	Valor máximo
range(datos)	Rango
mean(datos)	Media aritmética o promedio
geometric(datos)	Media geométrica
armonic(datos)	Media armónica
median(datos)	Mediana
lenght(datos)	Número de unidades muestrales
sd(datos)	Desviación estándar
var(datos)	Varianza
skweness(datos)	Coeficiente de asimetría
kurtosis(datos)	Coeficiente de curtosis
quantile(datos, 0.25)	Primer cuartil q1
quantile(datos, 0.75)	Tercer cuartil q3
IQR(datos)	Rango intercuartílico
sort(datos)	Ordenar
table(datos)	Tabla de frecuencias absolutas

Por ejemplo, si se teclea la instrucción

- summary(datos);

El resultado que arroja es el siguiente:

```
> datos<-read.csv(file="C:/Users/OctavioEC/Documents/datosprob3-1.csv",head=TRUE)
> summary(datos)
    Espesor
 Min.   :82.60
 1st Qu.:86.17
 Median :89.25
 Mean   :89.48
 3rd Qu.:93.10
 Max.   :98.00
```

Ejercicio 3.4

La Liga Nacional de Fútbol Americano (NFL por sus siglas en inglés) es la mayor liga de fútbol americano profesional de los Estados Unidos. Actualmente la NFL está formada por 32 franquicias establecidas en diversas ciudades y regiones estadounidenses. Se divide en dos conferencias: la Nacional (NFC) y la Americana (AFC). A su vez, cada conferencia se integra por cuatro divisiones (Norte, Sur, Este y Oeste) y cada una de ellas, por cuatro diferentes equipos. La temporada regular consiste en un calendario de diecisiete semanas durante las cuales cada equipo tiene una de descanso (denominada bye week), consistiendo en seis partidos contra rivales de la misma división, así como varios duelos interdivisionales e interconferenciales. A continuación, se muestran los resultados de la NFL del año 2018.

Resultados de la NFL en el 2018

No.	Division	Equipo	G	P	E	CTE	DIV	CONF	PF	PC
1	AFC Este	#2 New England Patriots	11	5	0	0.688	43470	43563	436	325
2	AFC Este	Miami Dolphins	7	9	0	0.438	43500	43622	319	433
3	AFC Este	Buffalo Bills	6	10	0	0.375	43557	43681	269	374
4	AFC Este	New York Jets	4	12	0	0.25	43586	43711	333	441
5	AFC Norte	#4 Baltimore Ravens	10	6	0	0.625	43527	43563	389	287
6	AFC Norte	Pittsburgh Steelers	9	6	1	0.594	36895	37017	428	360
7	AFC Norte	Cleveland Browns	7	8	1	0.459	36925	37047	359	392
8	AFC Norte	Cincinnati Bengals	6	10	0	0.375	43586	43681	368	455
9	AFC Norte	#3 Houston Texans	11	5	0	0.688	43500	43533	402	316
10	AFC Norte	#6 Indianapolis Colts	10	6	0	0.625	43500	43592	433	344
11	AFC Norte	Tennessee Titans	9	7	0	0.563	43527	43651	310	303
12	AFC Norte	Jacksonville Jaguars	5	11	0	0.313	43586	43681	245	316
13	AFC Oeste	#1 Kansas City Chiefs	12	4	0	0.75	43470	43506	565	421
14	AFC Oeste	#5 Los Angeles Chargers	12	4	0	0.75	43500	43533	428	329
15	AFC Oeste	Denver Broncos	6	10	0	0.375	43557	43681	329	349
16	AFC Oeste	Oakland Raiders	4	12	0	0.25	43586	43711	290	467
17	NFC Este	#4 Dallas Cowboys	10	6	0	0.625	43470	43533	339	324
18	NFC Este	#6 Philadelphia Eagles	9	7	0	0.562	43500	43622	367	348
19	NFC Este	Washington Redskins	7	9	0	0.437	43557	43622	281	359
20	NFC Este	New York Giants	5	11	0	0.313	43586	43681	369	412
21	NFC Norte	#3 Chicago Bears	12	4	0	0.75	43470	43506	397	273
22	NFC Norte	Minnesota Vikings	8	7	1	0.533	36925	37017	350	317
23	NFC Norte	Green Bay Packers	6	9	1	0.406	36982	37106	376	400
24	NFC Norte	Detroit Lions	6	10	0	0.375	43557	43681	324	360
25	NFC Sur	#1 New Orleans Saints	13	3	0	0.813	43500	43533	504	353
26	NFC Sur	Atlanta Falcons	7	9	0	0.438	43500	43592	414	423
27	NFC Sur	Carolina Panthers	7	9	0	0.438	43557	43651	376	382
28	NFC Sur	Tampa Bay Buccaneers	5	11	0	0.313	43557	43681	396	464
29	NFC Oeste	#2 Los Angeles Rams	13	3	0	0.8	6-0	43533	527	384
30	NFC Oeste	#5 Seattle Seahawks	10	6	0	0.6	43527	43563	428	347
31	NFC Oeste	San Francisco 49ers	4	12	0	0.25	43617	43740	342	435
32	NFC Oeste	Arizona Cardinals	3	13	0	0.2	43557	43711	225	425

Fuente: https://es.wikipedia.org/wiki/Temporada_2018_de_la_NFL

Se desea calcular algunos parámetros estadísticos de la población (censo) y obtener estimadores puntuales de dichos parámetros estadísticos, a partir de una o varias muestras tomadas de dicha población. Suponga que la característica de interés es el número de pases completados con éxito (PC).

a. Obtenga la estadística descriptiva censal de los pases completados con éxito.
b. Tome una muestra aleatoria simple de $n = 8$ de la población de $N = 32$ equipos y realice la misma estadística del inciso (a), a partir de la muestra tomada.
c. Obtenga una muestra aleatoria estratificada con $n = 8$ de la población de $N = 32$ equipos y realice la misma estadística del inciso (a), a partir de la muestra tomada.
d. Ordene la población de mayor a menor número de pases completados con éxito, obtenga una muestra aleatoria sistemática de $n = 8$ de la población de $N = 32$ equipos y realice la misma estadística del inciso (a), a partir de la muestra tomada.
e. Compare los resultados de las muestras tomadas con el resultado de la población bajo estudio y comente sus resultados.

La muestra aleatoria simple se genera aplicando ocho veces el comando aleatorio.entre(1,32)

Muestra Aleatoria Simple usando números aleatorios con excel:

No.	División	Equipo	PC
1	AFC Este	#2 New England Patriots	325
3	AFC Este	Buffalo Bills	374
13	AFC Oeste	#1 Kansas City Chiefs	421
21	NFC Norte	#3 Chicago Bears	273
26	NFC Sur	Atlanta Falcons	423
29	NFC Oeste	#2 Los Angeles Rams	384
31	NFC Oeste	San Francisco 49ers	435
32	NFC Oeste	Arizona Cardinals	425

La muestra aleatoria estratificada se genera aplicando ocho veces, una en cada estrato, el comando aleatorio.entre(a,b), por ejemplo, para el primer estrato aleatorio.entre(1,4).

Muestra Aleatoria Estratificada eligiendo aleatoriamente un equipo de cada División:

No.	División	Equipo	PC
1	AFC Este	#2 New England Patriots	325
6	AFC Norte	Pittsburgh Steelers	360
10	AFC Sur	#6 Indianapolis Colts	344
14	AFC Oeste	#5 Los Angeles Chargers	329
19	NFC Este	Washington Redskins	359
21	NFC Norte	#3 Chicago Bears	273
26	NFC Sur	Atlanta Falcons	423
30	NFC Oeste	#5 Seattle Seahawks	347

La muestra aleatoria sistemática se genera ordenando la población de equipos del mayor número de pases completados con éxito hasta el menor, después se aplica el comando aleatorio.entre(1,4) para generar un número entre 1 y 4, y posteriormente elegir cada cuatro equipos a un representante.

Muestra Aleatoria Sistemática eligiendo aleatoriamente un equipo de los primeros cuatro y luego cada 4:

No.	División	Equipo	PC
2	NFC Sur	Tampa Bay Buccaneers	464
6	AFC Este	Miami Dolphins	433
10	NFC Este	New York Giants	412
14	NFC Sur	Carolina Panthers	382
18	NFC Este	Washington Redskins	359
22	NFC Oeste	#5 Seattle Seahawks	347
26	NFC Este	#4 Dallas Cowboys	324
30	AFC Sur	Tennessee Titans	303

Con los datos de la tabla completa y con los datos de las tres muestras generadas, se obtiene la siguiente tabla de resultados:

Estadístico	*Población*	*Muestra Aleatoria Simple*	*Muestra Aleatoria Estratificada*	*Muestra Aleatoria Sistemática*
Media aritmética	372.4375	382.5	345	378
Media geométrica	368.7147	378.3064	342.7547	374.4745
Media armónica	364.9955	373.6738	340.4890	370.9932
Semirango	370	354	348	383.5
Mediana	360	402.5	345.5	370.5
Moda	360	#N/D	#N/D	#N/D
Rango	194	162	150	161
Desviación promedio	45.0898	43.875	27.25	44.75
Desviación estándar	53.4137	57.3909	42.0034	55.3637
Error típico	9.4423	20.2907	14.8504	19.5740
Varianza	2,853.0282	3,293.7143	1,764.2857	3,065.1429
Coef. Variación	0.1434	0.1500	0.1217	0.1465
Curtosis	–0.9537	0.4869	2.1719	–1.0177
Coef. Asim. Fisher	0.1338	–1.1676	0.2450	0.2535
Coef. Asim. Pearson	0.2329	#N/D	#N/D	#N/D
Tercer cuartil	421.5	423.5	359.25	417.25
Segundo Cuartil	360	402.5	345.5	370.5
Primer cuartil	328	361.75	328	341.25
Coef. Asim. Bowley- Yule	0.1119	–0.3198	–0.1200	0.2303
Máximo	467	435	423	464
Mínimo	273	273	273	303
Suma	11,918.00	3,060.00	2,760.00	3,024.00
Cuenta	32	8	8	8

3.3. Análisis de datos univariados agrupados

En este subtema se usará el concepto de frecuencia relativa de una muestra obtenida empíricamente de un fenómeno aleatorio o sujeto a incertidumbre, como un estimador del concepto de probabilidad desde el enfoque de la Escuela Frecuentista o de Von Mises. Se sabe que la frecuencia relativa tiende a la probabilidad cuando el número de ensayos tiende a infinito. Cabe señalar que este enfoque es un enfoque inductivo, parte de lo particular para llegar a lo general, es decir, a partir de conocer una muestra aleatoria se trata de inferir el comportamiento probabilístico de la población, por lo cual se verán conceptos que tienen el mismo nombre que lo calculado en el capítulo de Variables Aleatorias del Volumen de Probabilidad, con la diferencia de que en aquel capítulo se refería al comportamiento probabilístico de una población. En este subtema se hará el estudio del comportamiento en frecuencia de una muestra obtenida empíricamente. Si la muestra es representativa de la población, entonces el comportamiento en frecuencia deberá aproximarse al comportamiento probabilístico de la población.

Lo primero que se requiere es definir el rango de variación R de las lecturas obtenidas con el plan de muestreo establecido previamente en el subtema anterior. Se define este rango de variación como la máxima variación o máxima dispersión de las observaciones establecidas. Para calcular el rango, se utiliza la expresión 3.12 definida en el subtema 3.1.3.

Cuando se pretende comparar una muestra con otra, es conveniente definir el rango de variación de otra forma, se puede tomar como límite superior la lectura mayor y como límite inferior la lectura menor de todas las unidades muestrales de todas las muestras obtenidas. Otra forma, también común en la práctica, es tomar la especificación o norma que sugiere algún organismo reconocido, por ejemplo, si se pretende obtener el rango de variación de la edad de los académicos de la UNAM, el menor valor que se toma corresponde con un ayudante de profesor con 50% de sus créditos cubiertos que representa una edad aproximada de 20 años, y el máximo es 70, desde el punto de vista estatutario, que es la edad de retiro según el artículo 102 del Estatuto del Personal Académico. Así, en este último caso, el rango sería de 50 años, sin que se tenga una muestra para ello; de hecho, se da el caso de lecturas que están por debajo de 20 y por encima de 70.

Para el caso de los datos proporcionados en el Ejercicio 3.1 la lectura más grande es 98 y la lectura más pequeña es 82.6, por lo que el rango sería $R = 98 - 82.6 = 15.4$. No hay nada escrito de dónde deberá tomarse el límite inferior y el límite superior del rango, es por conveniencia; se pretende conocer la

variación de los datos, por lo que se puede tomar de 82 a 98, con lo cual el rango sería $R = 98\text{-}82 = 16$; habrá quien establezca tomar de 80 a 100, con lo cual el rango sería $R = 100\text{-}80 = 20$.

El siguiente paso es definir en cuántos subintervalos o clases se va a dividir este rango. Otra vez, en este punto, la Estadística es más arte que ciencia, algunos autores coinciden en establecer que el número de subintervalos o clases, m, esté entre 5 y 15, otros autores establecen que sea aproximadamente igual a la raíz cuadrada del total de datos; sin embargo, no deben ser muy pocas clases que no permitan conocer la dispersión de los datos o tantas que dificulten los cálculos; nuevamente se reitera en tomar lo más conveniente para hacer cálculos rápidamente y que se ilustre la variación de los datos. Un número adecuado de subintervalos o clases sería $m = 10$.

Con esta tónica lo que sigue es definir la amplitud de cada subintervalo, Δ, la cual se obtiene dividiendo el rango entre el número de clases, es decir,

$$\Delta = R/m \tag{3.29}$$

No es forzoso que la amplitud de cada subintervalo sea la misma, por facilidad se considera que sí. De esta forma, con el primer rango definido, la amplitud de cada subintervalo sería $\Delta = 15.4/10 = 1.54$; con el segundo rango, la amplitud sería $\Delta = 16/10 = 1.6$; con el tercer rango, la amplitud sería $\Delta = 20/10 = 2$.

Para definir los subintervalos o clases, es conveniente establecer criterios para determinar exactamente dónde cae una lectura. Los datos deben caer de tal forma que no se traslapen, en uno y solo un subintervalo, por lo cual algunos autores toman un dígito de más un poco antes del límite inferior del intervalo o un dígito de más un poco después del límite superior del intervalo. En Ingeniería, se establecen los conceptos de intervalo abierto o intervalo cerrado; puede convenirse que el límite inferior del intervalo sea abierto y el límite superior del intervalo sea cerrado, lo cual se indicaría así: (a, b]; también se puede hacer al revés [a, b). Con esta convención, se trata de definir una tabla, denominada Tabla de Frecuencias, con el siguiente formato:

Figura 3.29. Tabla de Frecuencias de un conjunto de datos obtenido como muestra aleatoria de una población

No. Clase	Lim Inf Clase	Lim Sup Clase	Marca de Clase	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
1	LI_1	$LS_1 = LI_1 + \Delta$	$t_1 = (LI_1 + LS_1)/2$	f_1	$F_1 = f1$	$f^*_1 = f_1/F_m$	$F^*_1 = F_1/F_m$
2	$LI_2 = LS_1$	$LS_2 = LI_2 + \Delta$	$t_2 = (LI_2 + LS_2)/2$	f_2	$F_2 = f_2 + F_1$	$f^*_2 = f_2/F_m$	$F^*_2 = F_2/F_m$
3	$LI_3 = LS_2$	$LS_3 = LI_3 + \Delta$	$t_3 = (LI_3 + LS_3)/2$	f_3	$F_3 = f_3 + F_2$	$f^*_3 = f_3/F_m$	$F^*_3 = F_3/F_m$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
m	$LI_m = LS_{m-1}$	$LS_m = LI_m + \Delta$	$t_m = (LI_m + LS_m)/2$	f_m	$F_m = f_m + F_{m-1}$	$f^*_m = f_m/F_m$	$F^*_m = F_m/F_m$

En la tabla de frecuencias anterior se puede ver que existen m subintervalos o clases. El límite inferior de cada clase es el valor mínimo de la clase, el límite superior es el valor máximo de la clase. El punto medio entre ambos límites se define como la marca de clase. Cabe señalar que la marca de clase es el valor donde se van a considerar concentrados todos los valores de ese intervalo. Al respecto, se toma como valor de concentración el punto medio, lo cual viene siendo su centro geométrico; sin embargo, no siempre el punto medio es el valor donde se consideran concentrados los datos; ocurre lo mismo que la diferencia entre centro geométrico y centro de gravedad, el centro geométrico es el centro del lugar geométrico, en cambio, el centro de gravedad es el punto donde se considera concentrada la masa de ese cuerpo. Si los puntos tienden a concentrarse en el límite inferior de cada clase, allí debería estar la marca de clase; de la misma forma, si los datos tienden a cargarse sobre el límite superior del intervalo, entonces allí debería estar la marca de clase. Como se considera que los datos se cargan uniformemente a todo lo largo de la clase, por convención, la marca de clase se coloca en el punto medio de cada intervalo.

En la tabla de frecuencias de la figura 3.29, la frecuencia absoluta es el número de observaciones o lecturas que caen en el intervalo de clase y si se suman todas las frecuencias desde la primera clase hasta el límite superior del intervalo que se esté analizando, a ese concepto se le denomina frecuencia absoluta acumulada del intervalo. Si se divide la frecuencia del intervalo entre el tamaño de la muestra n, se obtiene la frecuencia relativa del intervalo, y de la misma forma, si se divide la frecuencia absoluta acumulada entre n, se obtiene la frecuencia relativa acumulada. A esta tabla, como ya se dijo, se le conoce como tabla de frecuencias.

Cabe mencionar que el concepto de frecuencia tiene relación directa con el concepto de probabilidad. En el volumen anterior se definieron al menos cuatro

escuelas de probabilidad: clásica o de Laplace, frecuentista o de von Misses, subjetivista o de Savage y axiomática, constructivista o de Kolmogórov.

Nótese que en la segunda escuela se dice frecuentista precisamente porque la probabilidad se define como el límite de la frecuencia con que ocurre un fenómeno, cuando el tamaño de la muestra tiende a infinito; esto implica que si la muestra es representativa de la población entonces la frecuencia viene siendo un estimador de la probabilidad, es decir, debe aproximarse a la probabilidad.

Para obtener la frecuencia de cada intervalo, se puede contabilizar manualmente cuántas observaciones caen entre el límite inferior y el límite superior de cada intervalo, pero una forma rápida de contabilizarlo es utilizando Excel:

Por ejemplo, suponga que los datos del ejercicio 3.1 se encuentran capturados en una página de Excel en las celdas A1:I10, como se muestra en la figura 3.30:

Figura 3.30. Observaciones del Ejercicio 3.1

	A	B	C	D	E	F	G	H	I
1	94.1	86.6	94.3	94.1	93.1	85.1	84.6	97.3	85.1
2	93.2	91.2	93.2	92.1	94.6	84.0	83.6	96.8	90.5
3	90.6	86.1	86.7	96.4	96.3	93.7	85.4	94.4	95.6
4	91.4	90.4	83.0	88.2	94.7	87.7	89.7	96.1	88.3
5	88.2	89.1	95.3	86.4	91.1	90.6	87.6	98.0	84.1
6	86.1	87.3	94.1	85.0	92.4	89.4	85.1	85.4	83.7
7	95.1	84.1	97.8	84.9	90.6	88.6	89.6	86.6	82.9
8	90.0	90.1	93.1	87.3	89.1	84.1	90.0	91.7	87.3
9	92.4	95.2	86.4	89.6	88.8	82.6	90.1	87.5	86.4
10	87.3	86.1	87.6	90.3	86.4	83.1	94.3	84.2	84.5

Si se teclea el comando:

F_{90} = FRECUENCIA($A$1:$I$10, 90)

Se estaría obteniendo 51, lo que quiere decir que existen 51 observaciones que son menores o iguales a 90; por favor, cuente manualmente cuántos números en el cuadro de la figura 3.20 de arriba son menores o iguales a 90 y comprobará que son 51. Este concepto no es la frecuencia del intervalo, sino que representa a la frecuencia acumulada hasta el valor 90.

Para calcular la frecuencia absoluta acumulada hasta el límite superior del primer intervalo de clase se tendría que aplicar el siguiente comando:

F_1 = FRECUENCIA($A$1:$I$10, LS1) (3.30)

En este comando el valor LS1 es incluido en el intervalo.

Nótese qué si se pretende calcular la frecuencia absoluta en el primer intervalo, esta sería igual a la frecuencia absoluta acumulada del primer intervalo.

Si se pretende calcular la frecuencia absoluta de la segunda clase, el comando por aplicar sería:

f_2 = FRECUENCIA($A$1:$I$10, LS2) - FRECUENCIA($A$1:$I$10, LI2) (3.31)

La frecuencia absoluta acumulada del intervalo k sería:

F_k = FRECUENCIA($A$1:$I$10, LSk) (3.32)

En donde se puede apreciar que la frecuencia se acumula hasta el límite superior del intervalo, no hasta su marca de clase, como conceptualmente lo grafica Excel, de allí la aclaración.

La frecuencia absoluta del intervalo k sería:

f_k = FRECUENCIA($A$1:$I$10, LSk) - FRECUENCIA($A$1:$I$10, LIk) (3.33)

Con este comando, aplicado en el cuadro de datos de la figura 3.30, se obtiene la siguiente tabla de frecuencias, figura 3.31:

Figura 3.31. Tabla de frecuencias del Ejercicio 3.31

No. Clase	Lim Inf Clase	Lim Sup Clase	Marca de Clase	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
1	82.6	84.14	83.37	10	10	0.1111	0.1111
2	84.14	85.68	84.91	10	20	0.1111	0.2222
3	85.68	87.22	86.45	10	30	0.1111	0.3333
4	87.22	88.76	87.99	12	42	0.1333	0.4667
5	88.76	90.3	89.53	12	54	0.1333	0.6000
6	90.3	91.84	91.07	9	63	0.1000	0.7000
7	91.84	93.38	92.61	7	70	0.0778	0.7778
8	93.38	94.92	94.15	9	79	0.1000	0.8778
9	94.92	96.46	95.69	7	86	0.0778	0.9556
10	96.46	98	97.23	4	90	0.0444	1.0000
			Suma=	90			

Con esta tabla de frecuencias se pueden hacer estimaciones de frecuencia o de frecuencia relativa muy importantes. Por ejemplo, qué porcentaje de las lecturas es menor o igual a 87.22; en este caso observe el tercer renglón de la tabla:

$F(x \leq 87.22) = 30$

$F^*(x \leq 87.22) = 30/90 = 0.3333 = 33.33\%$

¿Qué porcentaje de las observaciones es mayor a 87.22?

$F(x > 87.22) = n - F(x \leq 87.22) = 90\text{-}30 = 60$

$F^*(x > 87.22) = 1 - F^*(x \leq 87.22) = 1 - 0.3333 = 0.6667 = 66.67\%$

¿Qué porcentaje de las observaciones se encuentra entre 90.3 y 91.84, excluyendo al límite inferior?

$F(90.3 < x \leq 91.84) = 9$

$F^*(90.3 < x \leq 91.84) = 9/90 = 0.10 = 10\%$

¿Qué porcentaje de las observaciones se encuentra entre 87.22 y 91.84, excluyendo al límite inferior?

$F(87.22 < x \leq 91.84) = F(87.22 < x \leq 88.76) + F(88.76 < x \leq 90.3) + F(90.3 < x \leq 91.84) = 12 + 12 + 9 = 33$

$F^*(87.22 < x \leq 91.84) = F^*(87.22 < x \leq 88.76) + F^*(88.76 < x \leq 90.3) + F^*(90.3 < x \leq 91.84) = 12 + 12 + 9 = 33/90 = 0.3667 = 36.67\%$

Otras preguntas que pueden responderse con la tabla de frecuencias se plantearían en modo inverso; por ejemplo, ¿para qué valor de *x* el porcentaje de lecturas es menor o igual a 70%? Observe que en las preguntas anteriores se daba un valor de *x* y se pedía la frecuencia absoluta o relativa. En esta pregunta es al contrario, se da una frecuencia relativa y se pide para qué valor de *x* se cumple.

Observe la sexta clase de la tabla de la figura 3.31:

$x_{F^* = 0.7} = 91.84$

Este valor de *x* se puede obtener también para valores no situados en los puntos definidos. Por ejemplo, ¿para qué valor de *x* el porcentaje de lecturas es menor o igual a 25%?

Nótese que el 25% en la frecuencia relativa acumulada cae en el tercer intervalo, entre la frecuencia relativa 22.22% con el valor $x = 85.68$ y la frecuencia relativa acumulada 33.33% con el valor de $x = 87.22$. En estos dos valores se tienen definidos dos puntos (85.68, 0.2222) y (87.22, 0.3333), y se pretende obtener el valor de *x* para el cual (x, 0.25). Se tienen definidos dos puntos extremos de una recta y se pretende obtener un punto intermedio de ella.

Esto se realiza por interpolación lineal:

$$F_x^* - F_{x1}^* = \frac{F_{x_2}^* - F_{x_1}^*}{x_2 - x_1}\left(x - x_1\right)$$

Despejando:

$$x = \frac{x_2 - x_1}{F_{x_2}^* - F_{x_1}^*}\left(F_x^* - F_{x1}^*\right) + x_1 \tag{3.34}$$

Con esta expresión es posible calcular los denominados Cuantiles o Fractiles; los primeros de los cuales se denominan Cuartiles, primer cuartil q_1, segundo cuartil q_2 o mediana, q_3 tercer cuartil, cada uno de ellos representa el valor de *x* para una frecuencia acumulada del 25%, del 50% y del 75% respectivamente:

$$q_1 = \frac{LSq_1 - LIq_1}{F_{LSq_1}^* - F_{\ Llq_1}^*}\left(0.25 - F^*_{\ Llq_1}\right) + Llq_1 \tag{3.35}$$

$$q_2 = \frac{LSq_2 - LIq_2}{F_{LSq_2}^* - F_{LIq_2}^*}\left(0.50 - F_{LIq_2}^*\right) + Llq_2 \tag{3.36}$$

$$q_3 = \frac{LSq_3 - LIq_3}{F_{LSq_3}^* - F_{LIq_3}^*}\left(0.75 - F_{LIq_3}^*\right) + Llq_3 \tag{3.37}$$

A la diferencia entre el tercer cuartil y el primer cuartil se le conoce como Intervalo intercuartil:

$$Iiq = q_3 - q_1 \tag{3.38}$$

También, con la expresión 3.33 es posible calcular los llamados Deciles, los cuales representan los valores para los cuales la frecuencia relativa acumulada representa un múltiplo del 10%; en este sentido, se tienen 10 deciles; por ejemplo, el primer decil y el noveno decil representan aquellos valores para los cuales la frecuencia relativa acumulada es del 10% y del 90% respectivamente, y se calcularían con las siguientes expresiones:

$$d_1 = \frac{LSd_1 - Lld_1}{F^{\star}_{LSd_1} - F^{\star}_{LId_1}}\left(0.10 - \mathrm{F}^{\star}_{Lld_1}\right) + LId_1 \tag{3.39}$$

$$d_9 = \frac{LSd_9 - LId_9}{F^{\star}_{LSd_9} - F^{\star}_{LId_9}}\left(0.90 - \mathrm{F}^{\star}_{Lld_9}\right) + LId_9 \tag{3.40}$$

A la diferencia entre el noveno decil y el primer decil se le conoce como Intervalo interdecil:

$$Iid = d_9 - d_1 \tag{3.41}$$

De la misma forma, con la expresión 3.33 es posible calcular los llamados Percentiles, los cuales representan los 100 valores para los cuales la frecuencia relativa acumulada representa un múltiplo del 1%; por ejemplo, el percentil uno y el percentil 99 representan aquellos valores para los cuales la frecuencia relativa acumulada es del 1% y del 99%, y se calcularían con las siguientes expresiones:

$$p_1 = \frac{LSp_1 - LIp_1}{F^{\star}_{LSp_1} - F^{\star}_{Ldp_1}}\left(0.01 - F^{\star}_{LIp_1}\right) + LIp_1 \tag{3.42}$$

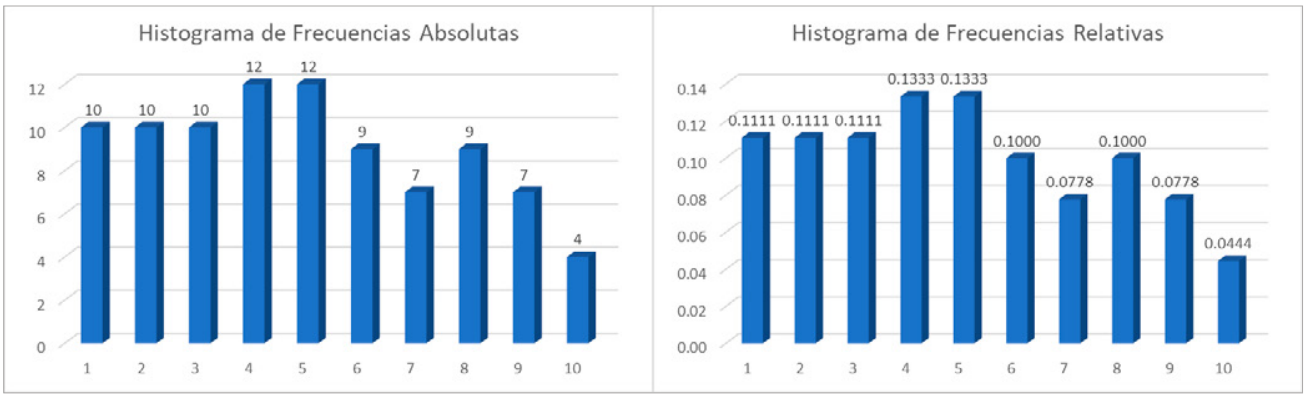
$$p_{99} = \frac{LSp_{99} - LIp_{99}}{F^{\star}_{LSp_{99}} - F^{\star}_{LIp_{99}}}\left(0.99 - F^{\star}_{LIp_{99}}\right) + LIp_{99} \tag{3.43}$$

A la diferencia entre el percentil 99 y el percentil uno se le conoce como Intervalo interpercentil:

$$Iip = p_{99} - p_1 \tag{3.44}$$

Si se grafica con un diagrama de barras la Frecuencia Absoluta o la Frecuencia Relativa, se obtiene lo que se denomina un Histograma:

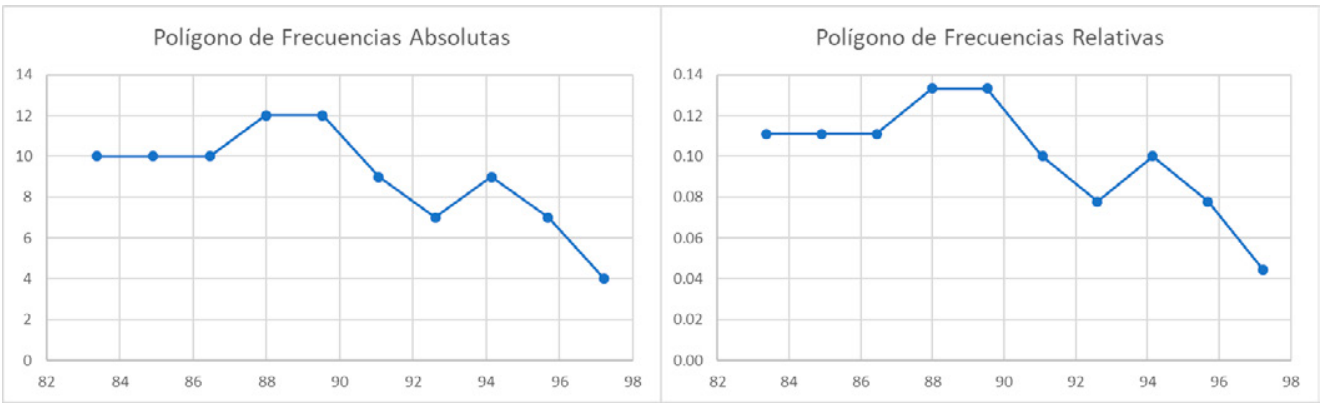
Figura 3.32. Histograma de los datos del Ejercicio 3.1



Nótese que ambos histogramas tienen la misma forma, solo cambia la escala de su eje vertical. De hecho, si los valores de cada barra de la derecha los multiplican por $n = 90$, se obtienen los valores de las barras de la izquierda.

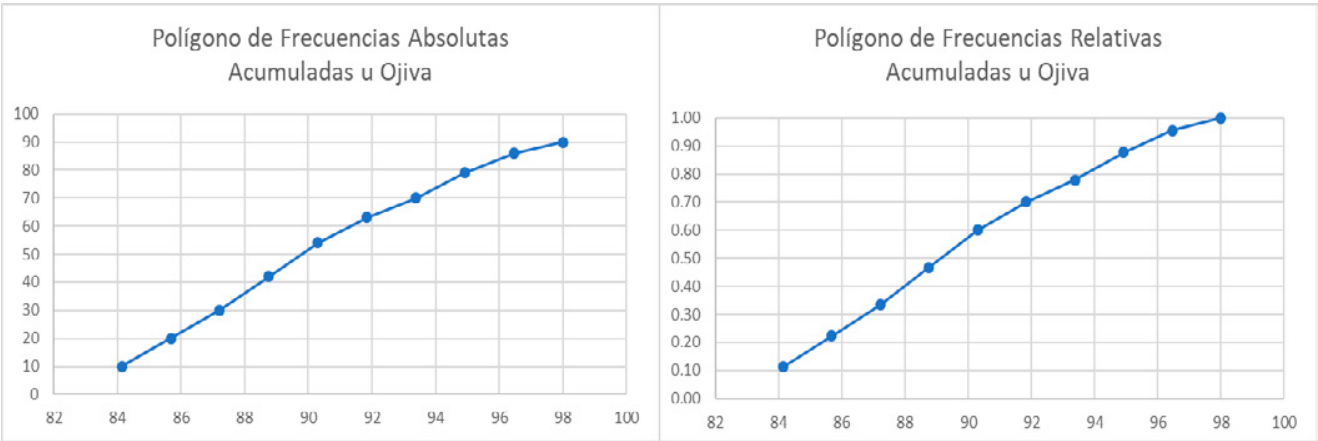
Si se trazan tomando como eje de las abscisas a la marca de clase y como eje de las ordenadas las frecuencias absolutas o relativas, se obtiene lo que se denomina el Polígono de Frecuencias Absolutas o el Polígono de Frecuencias Relativas. Nuevamente estos polígonos tienen la misma forma, la única diferencia es la escala del eje vertical.

Figura 3.33. Polígonos de Frecuencias Absolutas y Relativas de los datos del Ejercicio 3.1



Si ahora se grafica el límite superior de cada clase en el eje de las abscisas y la frecuencia absoluta acumulada o la frecuencia relativa acumulada en el eje de las ordenadas, se obtiene lo que se denomina el Polígono de Frecuencias Acumuladas, también conocido como Ojiva, como se muestra en la figura 3.34:

Figura 3.34. Polígonos de Frecuencias Acumuladas u Ojivas

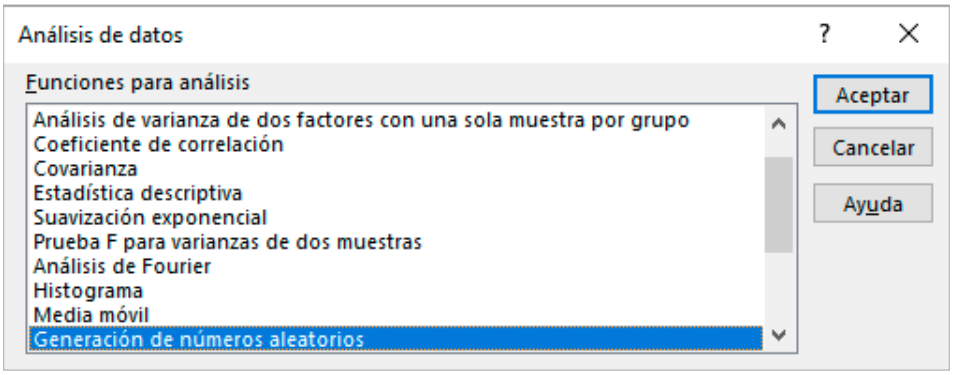


Nuevamente, se puede apreciar que tanto el polígono de frecuencias absolutas acumuladas como el polígono de frecuencias relativas acumuladas tienen la misma forma, solo difieren en la escala del eje vertical.
También se le denomina Ojiva por la forma No Decreciente de dicha curva, que la hace ver como la punta de un proyectil.

Hay una forma más rápida de obtener la tabla de frecuencias usando Excel, los pasos son los siguientes:

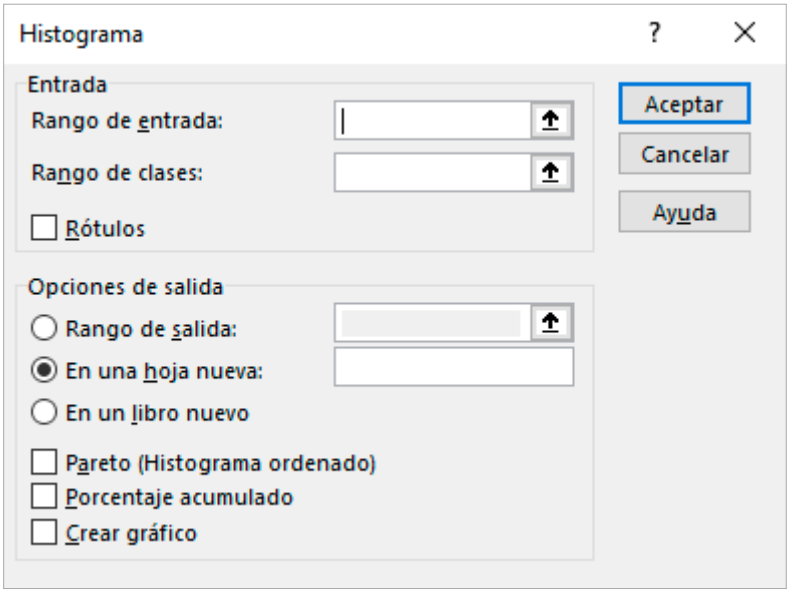
1. Se da un clic en el menú denominado Datos y luego en el submenú intitulado Análisis de datos.
2. Una vez activado el submenú de Análisis de datos, dándole clic aparece la siguiente pantalla:

Figura 3.35.



3. Se da clic en Histograma, luego en aceptar y aparece la pantalla:

Figura 3.36



En esta pantalla, en donde dice Rango de entrada, se deben teclear las coordenadas de las celdas donde se encuentran los datos $A$1: $I$10 (el símbolo pesitos se utiliza para fijar o anclar la columna y el renglón); de otra forma, sitúese donde está el primer dato y arrastre el ratón sin soltar el botón izquierdo del mismo hasta donde se encuentre el último dato. En esta misma pantalla, palomee la opción Rótulos en la primera fila si existe encabezado de los datos, si no existe encabezado déjela en blanco. En la opción Rango de clases teclee los límites de los intervalos de clase. Palomee las opciones Porcentaje acumulado, Crear Gráfico y dé un clic en Aceptar.

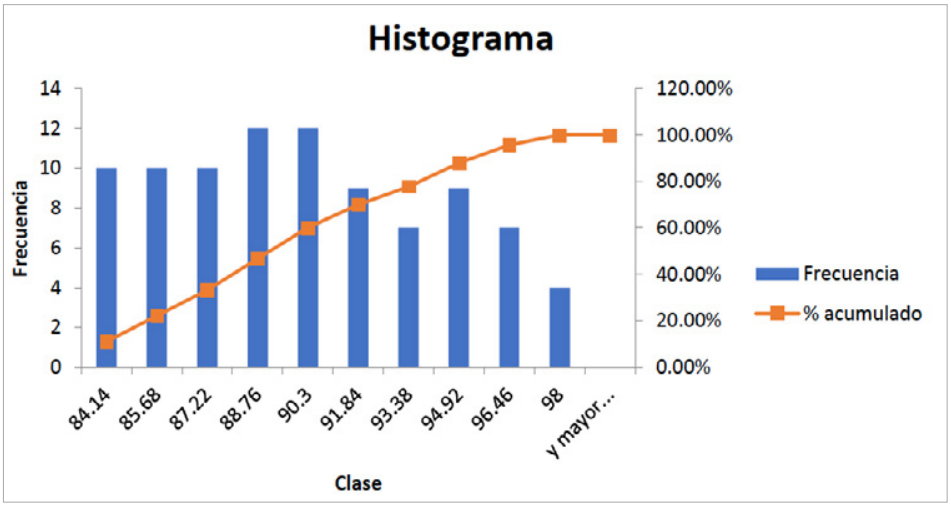
Aparece la Tabla de Frecuencias:

Figura 3.37. Tabla de Frecuencias Ejercicio 3.1

Clase	*Frecuencia*	*% acumulado*
84.14	10	11.11%
85.68	10	22.22%
87.22	10	33.33%
88.76	12	46.67%
90.3	12	60.00%
91.84	9	70.00%
93.38	7	77.78%
94.92	9	87.78%
96.46	7	95.56%
98	4	100.00%
y mayor…	0	100.00%

Y se obtiene en una sola gráfica el histograma y la ojiva de los datos del Ejercicio 3.1:

Figura 3.38. Histograma y Ojiva de los datos del Ejercicio 3.1

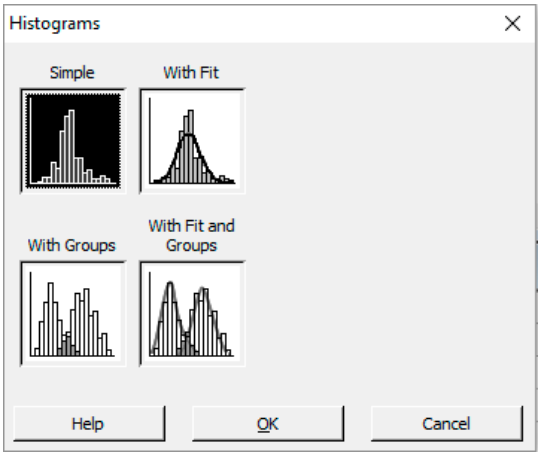


Nótese que Excel concentra la frecuencia acumulada en la marca de clase, por simplificar y facilitar su trazado, pero la frecuencia acumulada debe concentrarse en el límite superior de cada clase.

Otro software que puede ser usado para obtener la tabla de frecuencias y las gráficas relacionadas con ella es Minitab; a continuación, se explicará el procedimiento a seguir:

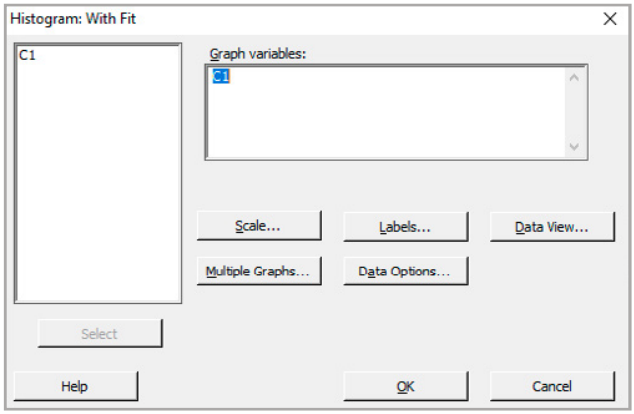
a. Si los datos están agrupados en una tabla de excel, hay que colocarlos todos en una sola columna, ya que se trata de una sola población.
b. Se copia la columna de 90 datos y se pega en la hoja electrónica de minitab, por ejemplo, en la primera columna.
c. En el menú principal de minitab se da click en Graph y luego en Histogram, apareciendo la siguiente pantalla en la cual se selecciona la segunda opción With Fit y se oprime Ok

Figura 3.39.



d. Aparece la pantalla que se muestra en la figura 3.40, en la cual se selecciona C1, luego Select y se oprime Ok

Figura 3.40.



e. Obteniéndose el Histograma que se muestra en la figura 3.41.

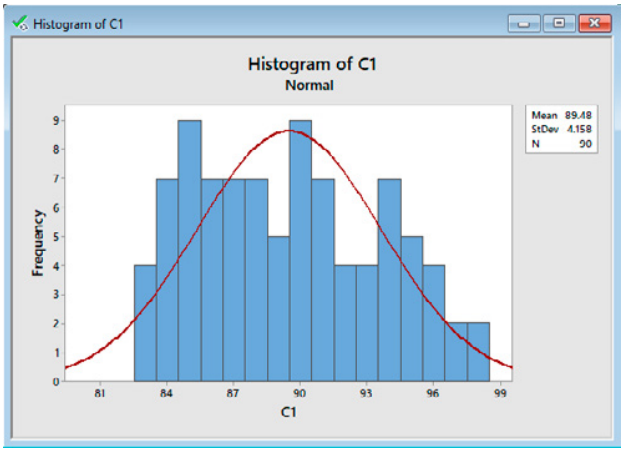


Figura 3.41. Histograma de los datos del Ejercicio 3.1

Usando un procedimiento similar, se obtiene la Ojiva:

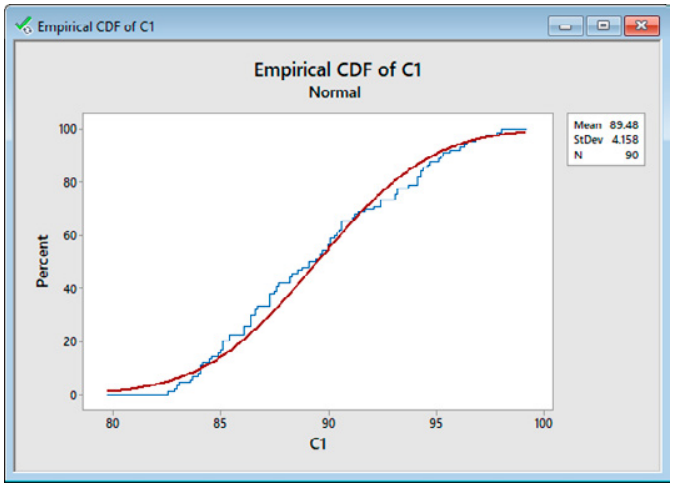


Figura 3.42. Ojiva de los datos del Ejercicio 3.1

De la misma forma, puede obtenerse la tabla de frecuencias y los gráficos asociados a ella, a través del software R, a continuación, se explicará la forma de llevarlo a cabo.

1. Alimentar a R con los datos del Ejercicio 3.1 en formato csv, como ya se explicó antes, o en su defecto capturar los datos en R con el siguiente comando:

datos < -
c(94.1,93.2,90.6,91.4,88.2,86.1,95.1,90.0,92.4,87.3,86.6,91.2,86.1,90.4,89.1,87.3,84.1,90.1,95.2,86.1,94.3,93.2,86.7,83.0,95.3,94.1,97.8,93.1,86.4,87.6,94.1,92.1,96.4,88.2,86.4,85.0,84.9,87.3,89.6,90.3,93.1,94.6,96.3,94.7,91.1,92.4,90.6,89.1,88.8,86.4,85.1,84.0,93.7,87.7,90.6,89.4,88.6,84.1,82.6,83.1,84.6,83.6,85.4,89.7,87.6,85.1,89.6,90.0,90.1,94.3,97.3,96.8,94.4,96.1,98.0,85.4,86.6,91.7,87.5,84.2,85.1,90.5,95.6,88.3,84.1,83.7,82.9,87.3,86.4,84.5)

2. Si los datos se obtienen de un archivo externo a R a través de Excel, teclear el siguiente comando:

   - Datos < -read.csv(file = ”datosprob3-1.csv”, head = TRUE, sep = ”,”)

3. Para obtener la tabla de frecuencias de la muestra dada en el Ejercicio 3.1, se utiliza el siguiente comando:

   - table < -(datos)

Dando como resultado lo siguiente:

datos

82.6	82.9	83	83.1	83.6	83.7	84	84.1	84.2	84.5	84.6	84.9	85	85.1	85.4	86.1
1	1	1	1	1	1	1	3	1	1	1	1	1	3	2	3
86.4	86.6	86.7	87.3	87.5	87.6	87.7	88.2	88.3	88.6	88.8	89.1	89.4	89.6	89.7	90
4	2	1	4	1	2	1	2	1	1	1	2	1	2	1	2
90.1	90.3	90.4	90.5	90.6	91.1	91.2	91.4	91.7	92.1	92.4	93.1	93.2	93.7	94.1	94.3
2	1	1	1	3	1	1	1	1	1	2	2	2	1	3	2
94.4	94.6	94.7	95.1	95.2	95.3	95.6	96.1	96.3	96.4	96.8	97.3	97.8	98		
1	1	1	1	1	1	1	1	1	1	1	1	1	1		

4. Para trazar el histograma de la muestra se teclea el siguiente comando:

   - hist(datos, breaks = 10, main = “Fijando el número de clases (10)”)

   Lo cual proporciona el gráfico mostrado en la figura 3.43:

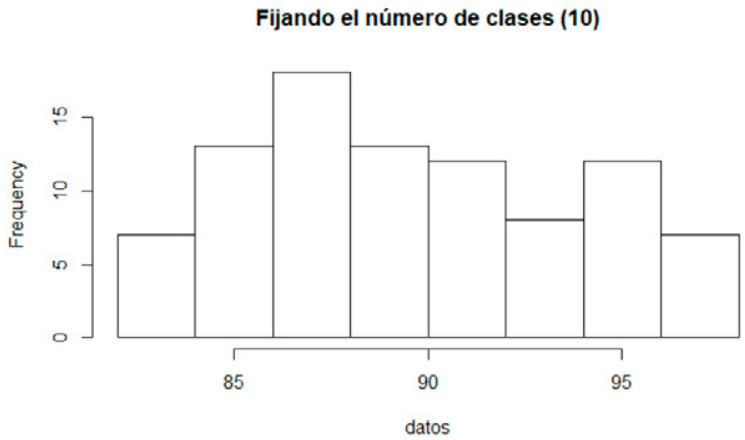


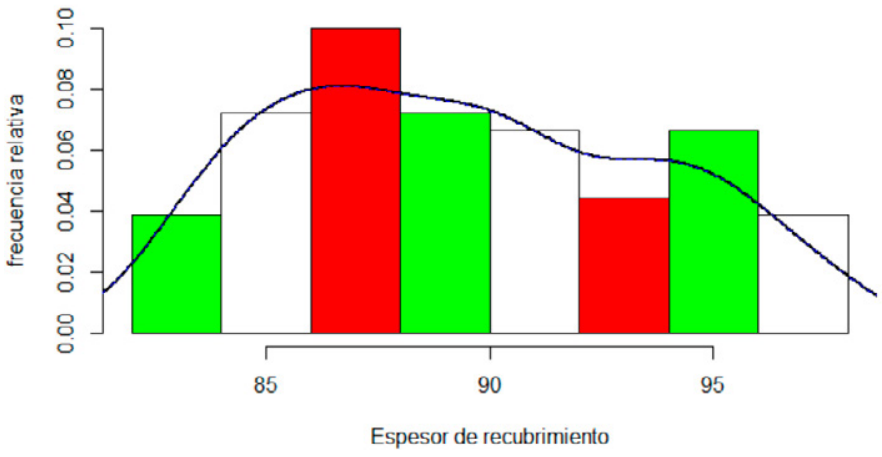
Figura 3.43. Histograma de la muestra dada en el Ejercicio 3.1

Se puede mejorar la presentación de este histograma con las siguientes instrucciones:

```
> hist( datos, breaks = 10, probability = TRUE, xlab = "Espesor de
recubrimiento", ylab = "frecuencia relativa", col = c( "green", "white", "red" ),
main = "Histograma del Ejercicio 3.1")
> lines( density( datos ), lwd = 2 ) # + su curva de densidad
> lines( density(datos), col = "blue", lty = 2, ps = 20 )
```

Lo cual arroja el histograma que se muestra en la figura 3.44:

Figura 3.44. Histograma de la muestra dada en el Ejercicio 3.1



3.4. Cálculo de los estadísticos muestrales o parámetros descriptivos de una muestra, para datos agrupados en una tabla de frecuencias

En el subtema 3.2 (ver figura 3.10), se analizaron los conceptos y las expresiones o fórmulas para el cálculo de los estadísticos muestrales o parámetros de una muestra (de tendencia central, de dispersión, de forma, etcétera). Todas las expresiones que se definieron en ese subtema se basan en el cálculo de los estadísticos a partir de los datos enumerados de la muestra tomada, pero, ¿qué ocurre si no se dispone de los datos, sino solo de una tabla de frecuencias?, ¿cómo se calcularían los parámetros muestrales?

Cabe señalar que las fórmulas que se listarán a continuación, parten del supuesto que todas las observaciones que caen en un subintervalo o clase se concentran en su marca de clase, de allí que los valores que resulten no van a coincidir con los valores de las expresiones usadas en el subtema 3.1.

Por ejemplo, ¿cómo calcular los parámetros muestrales de tendencia central, de dispersión y de forma, si solo se proporciona la tabla de frecuencias de la figura 3.29?

No. Clase	Lim Inf Clase	Lim Sup Clase	Marca de Clase	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
1	LI_1	$LS_1 = LI_1 + \Delta$	$t_1 = (LI_1 + LS_1)/2$	f_1	$F_1 = f1$	$f^*_1 = f_1/F_m$	$F^*_1 = F_1/F_m$
2	$LI_2 = LS_1$	$LS_2 = LI_2 + \Delta$	$t_2 = (LI_2 + LS_2)/2$	f_2	$F_2 = f_2 + F_1$	$f^*_2 = f_2/F_m$	$F^*_2 = F_2/F_m$
3	$LI_3 = LS_2$	$LS_3 = LI_3 + \Delta$	$t_3 = (LI_3 + LS_3)/2$	f_3	$F_3 = f_3 + F_2$	$f^*_3 = f_3/F_m$	$F^*_3 = F_3/F_m$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
m	$LI_m = LS_{m-1}$	$LS_m = LI_m + \Delta$	$t_m = (LI_m + LS_m)/2$	f_m	$F_m = f_m + F_{m-1}$	$f^*_m = f_m/F_m$	$F^*_m = F_m/F_m$

Para ilustrar la aplicación de todas las fórmulas que se verán en este subtema, se utilizará la Tabla de Frecuencias de la Figura 3.31, correspondiente a los datos del Ejercicio 3.1.

No. Clase	Lim Inf Clase	Lim Sup Clase	Marca de Clase	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
1	82.6	84.14	83.37	10	10	0.1111	0.1111
2	84.14	85.68	84.91	10	20	0.1111	0.2222
3	85.68	87.22	86.45	10	30	0.1111	0.3333
4	87.22	88.76	87.99	12	42	0.1333	0.4667
5	88.76	90.3	89.53	12	54	0.1333	0.6000
6	90.3	91.84	91.07	9	63	0.1000	0.7000
7	91.84	93.38	92.61	7	70	0.0778	0.7778
8	93.38	94.92	94.15	9	79	0.1000	0.8778
9	94.92	96.46	95.69	7	86	0.0778	0.9556
10	96.46	98	97.23	4	90	0.0444	1.0000
			Suma=	90			

También se usará la tabla de Frecuencias del Ejercicio 3.3 citado en el subtema 3.2:

Ejercicio 3.3

Suponga que se cuenta con la siguiente tabla de frecuencias, obtenida de Las Estadísticas del Personal Académico de la UNAM 2015, página 19, de la Dirección General de Asuntos del Personal Académico de la UNAM. (Fuente:http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2015.pdf)

Personal académico en la UNAM por edad											
Universo	Edad (años cumplidos)										Total
	Hasta 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 o más	
UNAM	896	2,356	3,539	4,457	4,924	4,847	5,288	5,106	3,712	2,401	1,822	39,348

Esta tabla debe completarse para formar la tabla de frecuencias, utilizando Excel, de la forma que se muestra a continuación:

Figura 3.45. Tabla de Frecuencias del Ejercicio 3.3

Personal Académico en la UNAM por Edad en el 2015

Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
	24	22	896	896	0.0228	0.0228
25	29	27	2,356	3,252	0.0599	0.0826
30	34	32	3,539	6,791	0.0899	0.1726
35	39	37	4,457	11,248	0.1133	0.2859
40	44	42	4,924	16,172	0.1251	0.4110
45	49	47	4,847	21,019	0.1232	0.5342
50	54	52	5,288	26,307	0.1344	0.6686
55	59	57	5,106	31,413	0.1298	0.7983
60	64	62	3,712	35,125	0.0943	0.8927
65	69	67	2,401	37,526	0.0610	0.9537
70	o más	72	1,822	39,348	0.0463	1.0000

Esta tabla presenta varias peculiaridades que vale la pena explicar; en primer lugar, observe el primero y el último intervalo de clase, son abiertos, el primero define como intervalo hasta 24, ¿qué significa?; el último intervalo de clase dice 70 o más, ¿hasta dónde?; si no se conoce la población, no se sabe hasta dónde. Vale la pena comentar que se trata de las edades del personal académico de la UNAM. Para ser personal académico de la UNAM, se debe contar con al menos un 50% de créditos para ser ayudante de profesor, lo que significa que tenga más o menos 20 años de edad, sería prácticamente imposible que hubiera un académico con 12 años o menos, lo más probable es que tenga una edad por arriba de 20 años, por ello se coloca la marca de clase del primer intervalo en 22 años. Por otra parte, el artículo 102 del Estatuto del Personal Académico establece que cuando un académico alcance la edad de 70 años dejará su plaza; si la institución requiere de sus servicios podrá contratarlo anualmente por honorarios; en la UNAM, en realidad existen académicos hasta mayores de 90 años, pero no son la norma, por ello, también, la marca de clase del último intervalo se considera en 72 años.

Nótese en este caso en particular, que no existe rango de variación como comúnmente se aplica, ya que no se conoce ni la edad mínima, ni la máxima. Por otra parte, también debe tomarse en cuenta que este ejercicio no aborda el análisis de una muestra, sino el análisis de toda la población de académicos de la UNAM en el 2015, por lo cual, más adelante que se hagan cálculos, la varianza se calcula dividiendo entre *n* y no entre N-1, lo cual se haría cuando se trate de una muestra, no de toda la población como es el caso.

3.4.1. Medidas de tendencia central para datos agrupados en una tabla de frecuencias

Para el cálculo de la media aritmética o promedio para datos agrupados en una tabla de frecuencias se utiliza la siguiente función matemática:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{j=m} t_j f_j = \frac{1}{n} \left[t_1 f_1 + t_2 f_2 + t_3 f_3 + \cdots + t_m f_m \right] \tag{3.45}$$

o

$$\bar{x} = \sum_{j=1}^{j=m} t_j f_j^* = \left[t_1 f_1^* + t_2 f_2^* + t_3 f_3^* + \cdots + t_m f_m^* \right] \tag{3.46}$$

donde

$$f^* = \frac{f}{n} \tag{3.47}$$

Calcule la media aritmética de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la media aritmética de la tabla 3.45, correspondiente al problema 3.3.

Problema 3.1. Espesor

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$
1	83.37	10	833.70
2	84.91	10	849.10
3	86.45	10	864.50
4	87.99	12	1,055.88
5	89.53	12	1,074.36
6	91.07	9	819.63
7	92.61	7	648.27
8	94.15	9	847.35
9	95.69	7	669.83
10	97.23	4	388.92
	Suma=	90	8,051.5400
		Media Aritm=	**89.4616**

Problema 3.3. Personal Académico en la UNAM

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$
1	22	896	19,712
2	27	2,356	63,612
3	32	3,539	113,248
4	37	4,457	164,909
5	42	4,924	206,808
6	47	4,847	227,809
7	52	5,288	274,976
8	57	5,106	291,042
9	62	3,712	230,144
10	67	2,401	160,867
11	72	1,822	131,184
	Suma=	39,348	1,884,311
		Media Aritm=	**47.8884**

Media geométrica para datos agrupados en una tabla de frecuencias

Para el cálculo de la media geométrica para datos agrupados en una tabla de frecuencias, se utiliza la siguiente función matemática:

$$\bar{x}_{geom} = \exp\left(\frac{f_1 Ln(t_1) + f_2 Ln(t_2) + \cdots f_m Ln(t_m)}{n}\right) \tag{3.48}$$

o

$$\bar{x}_{geom} = \exp\left(f_1^* Ln(t_1) + f_2^* Ln(t_2) + \cdots f_m^* Ln(t_m)\right) \tag{3.49}$$

Calcule la media geométrica de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la media geométrica de la tabla 3.45, correspondiente al problema 3.3.

Problema 3.1. Espesor

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$	$f_j \mathrm{Ln}(t_j)$
1	83.37	10	833.70	44.2329
2	84.91	10	849.10	44.4159
3	86.45	10	864.50	44.5957
4	87.99	12	1,055.88	53.7267
5	89.53	12	1,074.36	53.9349
6	91.07	9	819.63	40.6047
7	92.61	7	648.27	31.6988
8	94.15	9	847.35	40.9040
9	95.69	7	669.83	31.9278
10	97.23	4	388.92	18.3083
	Suma=	90	8,051.5400	404.3496
		Media Aritm=	89.4616	
			Media Geom=	**89.3689**

Problema 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$	$f_j \mathrm{Ln}(t_j)$
1	22	896	19,712	2,769.5740
2	27	2,356	63,612	7,764.9917
3	32	3,539	113,248	12,265.2394
4	37	4,457	164,909	16,093.8611
5	42	4,924	206,808	18,404.2852
6	47	4,847	227,809	18,661.6654
7	52	5,288	274,976	20,894.1768
8	57	5,106	291,042	20,643.8198
9	62	3,712	230,144	15,319.9228
10	67	2,401	160,867	10,095.4670
11	72	1,822	131,184	7,792.0857
	Suma=	39,348	1,884,311	150,705
		Media Aritm=	47.8884	
			Media Geom=	**46.0652**

Media armónica

Para obtener la media armónica para datos agrupados en una tabla de frecuencias se usa la siguiente función matemática:

$$\bar{x}_{armo} = \frac{n}{\sum_{j=1}^{j=m} \frac{f_j}{t_j}} = \frac{n}{\frac{f_1}{t_1} + \frac{f_2}{t_2} + \frac{f_3}{t_3} + \cdots + \frac{f_m}{t_m}} \tag{3.50}$$

o

$$\bar{x}_{armo} = \frac{n^2}{\sum_{j=1}^{j=m} \frac{f_j^*}{t_j}} = \frac{n^2}{\left[\frac{f_1^*}{t_1} + \frac{f_2^*}{t_2} + \frac{f_3^*}{t_3} + \cdots + \frac{f_m^*}{t_m} \right]} \tag{3.51}$$

Calcule la media armónica de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la media armónica de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$	$f_j \mathrm{Ln}(t_j)$	f_j / t_j
1	83.37	10	833.70	44.2329	0.1199
2	84.91	10	849.10	44.4159	0.1178
3	86.45	10	864.50	44.5957	0.1157
4	87.99	12	1,055.88	53.7267	0.1364
5	89.53	12	1,074.36	53.9349	0.1340
6	91.07	9	819.63	40.6047	0.0988
7	92.61	7	648.27	31.6988	0.0756
8	94.15	9	847.35	40.9040	0.0956
9	95.69	7	669.83	31.9278	0.0732
10	97.23	4	388.92	18.3083	0.0411
	Suma=	90	8,051.5400	404.3496	1.0081
		Media Aritm=	89.4616		
			Media Geom=	89.3689	
				Media Armo=	**89.2768**

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$t_j f_j$	$f_j \mathrm{Ln}(t_j)$	f_j / t_j
1	22	896	19,712	2,769.5740	40.7273
2	27	2,356	63,612	7,764.9917	87.2593
3	32	3,539	113,248	12,265.2394	110.5938
4	37	4,457	164,909	16,093.8611	120.4595
5	42	4,924	206,808	18,404.2852	117.2381
6	47	4,847	227,809	18,661.6654	103.1277
7	52	5,288	274,976	20,894.1768	101.6923
8	57	5,106	291,042	20,643.8198	89.5789
9	62	3,712	230,144	15,319.9228	59.8710
10	67	2,401	160,867	10,095.4670	35.8358
11	72	1,822	131,184	7,792.0857	25.3056
	Suma=	39,348	1,884,311	150,705.0889	891.6891
		Media Aritm=	47.8884		
			Media Geom=	46.0652	
				Media Armo=	**44.1275**

Mediana para datos agrupados en una tabla de frecuencias

Para el cálculo de la mediana se usa la fórmula para obtener el segundo cuartil en datos agrupados, es decir,

$$m_e = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}} \left(0.50 - F^*_{LII}\right) + LII \tag{3.52}$$

En donde LSI representa al límite superior del subintervalo o clase cuya frecuencia relativa acumulada cruza el 50% de los datos; LII representa al límite inferior de dicho intervalo; F^*_{LSI} es la frecuencia relativa acumulada en LSI y F^*_{LII} es la frecuencia relativa acumulada en LII.

Calcule la mediana de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la mediana de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	82.6	84.14	83.37	10	10	0.1111	0.1111
2	84.14	85.68	84.91	10	20	0.1111	0.2222
3	85.68	87.22	86.45	10	30	0.1111	0.3333
4	87.22	88.76	87.99	12	42	0.1333	0.4667
5	88.76	90.3	89.53	12	54	0.1333	0.6000
6	90.3	91.84	91.07	9	63	0.1000	0.7000
7	91.84	93.38	92.61	7	70	0.0778	0.7778
8	93.38	94.92	94.15	9	79	0.1000	0.8778
9	94.92	96.46	95.69	7	86	0.0778	0.9556
10	96.46	98	97.23	4	90	0.0444	1.0000
			Suma=	90		1.0000	
						Mediana=	**89.1450**

$$m_e = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}}(0.50 - F^*_{LII}) + LII$$

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	menor a 24	24	22	896	896	0.0228	0.0228
2	25	29	27	2,356	3,252	0.0599	0.0826
3	30	34	32	3,539	6,791	0.0899	0.1726
4	35	39	37	4,457	11,248	0.1133	0.2859
5	40	44	42	4,924	16,172	0.1251	0.4110
6	45	49	47	4,847	21,019	0.1232	0.5342
7	50	54	52	5,288	26,307	0.1344	0.6686
8	55	59	57	5,106	31,413	0.1298	0.7983
9	60	64	62	3,712	35,125	0.0943	0.8927
10	65	69	67	2,401	37,526	0.0610	0.9537
11	70	o más	72	1,822	39,348	0.0463	1.0000
			Suma=	39,348		1.0000	
						Mediana=	**48.6125**

$$m_e = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}}(0.50 - F^*_{LII}) + LII$$

Moda para datos agrupados en una tabla de frecuencias

Para el cálculo de la moda para datos agrupados en una tabla de frecuencias se utilizará la gráfica de la figura 3.34. La moda siempre cae en el intervalo donde se encuentra la frecuencia absoluta o relativa más alta, a esta frecuencia se le denomina fmax, por lo que una primera aproximación a la moda de una muestra es tomar la marca de clase del intervalo donde se encuentra la máxima frecuencia:

$$m_o = t_{max} \tag{3.53}$$

Nótese que según la figura 3.46, se forman dos triángulos con un vértice coincidente. Por Ley de los Triángulos opuestos por el vértice, la altura de uno entre

su base es igual a la altura del otro entre su base, dado lo cual se puede deducir el valor de la moda a partir de la siguiente expresión:

$$m_o = LII + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) (LSI - LII) \tag{3.54}$$

Donde

$$\Delta_1 = f_{max} - f_{ant}$$
$$\Delta_2 = f_{max} - f_{ant}$$

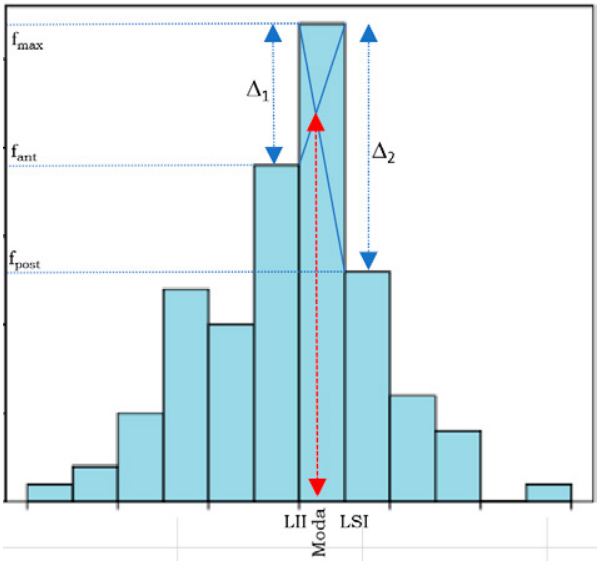


Figura 3.46

Calcule la moda de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la moda de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	82.6	84.14	83.37	10	10	0.1111	0.1111
2	84.14	85.68	84.91	10	20	0.1111	0.2222
3	85.68	87.22	86.45	10	30	0.1111	0.3333
4	87.22	88.76	87.99	12	42	0.1333	0.4667
5	88.76	90.3	89.53	12	54	0.1333	0.6000
6	90.3	91.84	91.07	9	63	0.1000	0.7000
7	91.84	93.38	92.61	7	70	0.0778	0.7778
8	93.38	94.92	94.15	9	79	0.1000	0.8778
9	94.92	96.46	95.69	7	86	0.0778	0.9556
10	96.46	98	97.23	4	90	0.0444	1.0000
			Suma=	90		1.0000	
							Media Aritm=
					Mediana=		89.1450
					Moda 1a Aprox=		**88.76**
					Moda 2a Aprox=		**88.4520**

$$m_o = LII + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) (LSI - LII)$$

Donde

$\Delta_1 = f_{max} - f_{ant}$

$\Delta_2 = f_{max} - f_{post}$

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	menor a 24	24	22	896	896	0.0228	0.0228
2	25	29	27	2,356	3,252	0.0599	0.0826
3	30	34	32	3,539	6,791	0.0899	0.1726
4	35	39	37	4,457	11,248	0.1133	0.2859
5	40	44	42	4,924	16,172	0.1251	0.4110
6	45	49	47	4,847	21,019	0.1232	0.5342
7	50	54	52	5,288	26,307	0.1344	0.6686
8	55	59	57	5,106	31,413	0.1298	0.7983
9	60	64	62	3,712	35,125	0.0943	0.8927
10	65	69	67	2,401	37,526	0.0610	0.9537
11	70	o más	72	1,822	39,348	0.0463	1.0000
			Suma=	39,348		1.0000	
							Media Aritm=
					Mediana=		48.6125
					Moda 1a Aprox=		**52**
					Moda 2a Aprox=		**53.5393**

$$m_o = LII + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) (LSI - LII)$$

Donde

$\Delta_1 = f_{max} - f_{ant}$

$\Delta_2 = f_{max} - f_{post}$

Semirango

Para obtener el semirango para datos agrupados en una tabla de frecuencias se usa la siguiente función matemática:

$$SR = \frac{LII_1 + LSI_m}{2} \tag{3.55}$$

En donde LII_1 es el límite inferior del primer subintervalo de la muestra y LSI_m es el límite superior del subintervalo m-ésimo.

Calcule el semirango de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el semirango de la tabla 3.45, correspondiente al problema 3.3.

De la tabla de frecuencias de la figura 3.31, se observa que $LII_1 = 82.6$ y $LSI_m = 98$, por lo que:

$$\text{SR} = \left(82.6 + 98\right)/2 = 90.3$$

La tabla 3.45 correspondiente al ejercicio 3.3 no indica el límite inferior del primer intervalo, pero siguiendo la secuencia lógica se fijará en 20; de la misma forma, el límite superior del último intervalo se fijará en 74, por lo que:

$$\text{SR} = \left(20 + 74\right) = 47$$

Figura 3.47. Formulario para calcular medidas de tendencia central de una muestra

CÁLCULO DE LOS PARÁMETROS ESTADÍSTICOS DE UNA MUESTRA			
Medidas	**Parámetro Estadístico**	**Datos sin agrupar**	**Datos Agrupados**
De tendencia central	Media Aritmética o Promedio	$\bar{x}=\frac{1}{n}\sum_{i=1}^{i=n}x_i=\frac{1}{n}\left[x_1+x_2+x_3+\cdots+x_n\right]$	$\bar{x}=\frac{1}{n}\sum_{j=1}^{j=m}t_j f_j=\frac{1}{n}\left[t_1 f_1+t_2 f_2+t_3 f_3+\cdots+t_m f_m\right]$ o $\bar{x}=\sum_{j=1}^{j=m}t_j f_j^{\star}=\left[t_1 f_1^{\star}+t_2 f_2^{\star}+t_3 f_3^{\star}+\cdots+t_m f_m^{\star}\right]$ donde $f^{\star}=\frac{f}{n}$
	Media Geométrica	$\bar{x}_{geom}=\sqrt[n]{\prod_{i=1}^{i=n}x_i}=\sqrt[n]{x_1\cdot x_2\cdot x_3\cdots x_n}$	$\bar{x}_{geom}=\exp\left(\frac{f_1 Ln(t_1)+f_2 Ln(t_2)+\cdots f_m Ln(t_m)}{n}\right)$ o $\bar{x}_{geom}=\exp\left(\frac{f_1^{*}Ln(t_1)+f_2^{*}Ln(t_2)+\cdots f_m^{*}Ln(t_m)}{n}\right)$
	Media Armónica	$\bar{x}_{armo}=\frac{n}{\sum_{i=1}^{i=n}\frac{1}{x_i}}=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}$	$\bar{x}_{armo}=\frac{n}{\sum_{j=1}^{j=m}\frac{f_j}{t_j}}=\frac{n}{\frac{f_1}{t_1}+\frac{f_2}{t_2}+\frac{f_3}{t_3}+\cdots+\frac{f_m}{t_m}}$ o $\bar{x}_{armo}=\frac{n^2}{\sum_{j=1}^{j=m}\frac{f_j^{\star}}{t_j}}=\frac{n^2}{\left[\frac{f_1^{\star}}{t_1}+\frac{f_2^{\star}}{t_2}+\frac{f_3^{\star}}{t_3}+\cdots+\frac{f_m^{\star}}{t_m}\right]}$
	Media Ponderada		$\bar{x}_{ponde}=\frac{1}{n}\sum_{j=1}^{j=m}f_j t_j=\sum_{j=1}^{j=m}f_j^{*}t_j=\left[f_1^{*}t_1+f_2^{*}+f_3^{*}+\cdots+f_m^{*}\right]$ Donde $f_j^{*}=\frac{f_j}{n}$
	Media Acotada	El promedio del p% de los datos centrales de una muestra	El promedio del p% de los datos centrales de una muestra
	Mediana o Segundo Cuartil	Se ordenan los valores de menor a mayor o de mayor a menor. Si n es impar se toma el de enmedio. Si n es par se toma el promedio de los dos de enmedio	$m_e=\frac{LSI-LII}{F^{\star}_{LSI}-F^{\star}_{LII}}\left(0.50-F^{\star}_{LII}\right)+LII$
	Moda	El valor más frecuente o el valor que más se repita en el conjunto de datos de la muestra. Si hay empate, se dice que no existe moda única (algunos autores toman el valor que se encuentra más a la izquierda)	$m_o=LII+\left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right)(LSI-LII)$ *Donde* $\Delta_1=f_{max}-f_{ant}$ $\Delta_2=f_{max}-f_{post}$
	Semirango	$SR=\frac{x_{max}+x_{min}}{2}$	$SR=\frac{LII_1+LSI_m}{2}$

3.4.2. Medidas de dispersión para datos agrupados en una tabla de frecuencias

El rango o amplitud R de para datos agrupados en una tabla de frecuencias

Para obtener el rango R para datos agrupados en una tabla de frecuencias se usa la siguiente función matemática:

$$R = LSI_m - LII_1 \tag{3.56}$$

En donde LII_1 es el límite inferior del primer subintervalo de la muestra y LSI_m es el límite superior del subintervalo m-ésimo.

Calcule el rango de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el rango de la tabla 3.45, correspondiente al problema 3.3.

De la tabla de frecuencias de la figura 3.31, se observa que $LII_1 = 82.6$ y $LSI_m = 98$, por lo que:

$$R = 98 - 82.6 = 15.4$$

La tabla 3.45 correspondiente al ejercicio 3.3 no indica el límite inferior del primer intervalo, pero siguiendo la secuencia lógica se fijará en 20; de la misma forma, el límite superior del último intervalo se fijará en 74, por lo que:

$$R = (74 - 20) = 54$$

Nótese que el rango es una medida de dispersión, en cambio el semirango es una medida de tendencia central.

Desviación Promedio para datos agrupados en una tabla de frecuencias

Se puede calcular la desviación promedio para datos agrupados en una tabla de frecuencias de la siguiente forma:

$$Desv\ Prom = \frac{1}{n}\sum_{j=1}^{j=m} |t_j - \bar{x}|\, f_j \tag{3.57}$$

$$Desv\ Prom = \sum_{j=1}^{j=m} |t_j - x|\, f_j^{*} \tag{3.58}$$

donde

$$f_{j=}^{*} \frac{f_j}{n}$$

Calcule la desviación promedio de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la desviación promedio de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$\lvert t_j$-media$\rvert f_j$
1	83.37	10	60.9156
2	84.91	10	45.5156
3	86.45	10	30.1156
4	87.99	12	17.6587
5	89.53	12	0.8213
6	91.07	9	14.4760
7	92.61	7	22.0391
8	94.15	9	42.1960
9	95.69	7	43.5991
10	97.23	4	31.0738
	Suma=	90	308.4107
		Desv Prom=	**3.4268**

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$\lvert t_j$-media$\rvert f_j$
1	22	896	23,195.9662
2	27	2,356	49,212.9648
3	32	3,539	56,228.8890
4	37	4,457	48,529.3991
5	42	4,924	28,994.2610
6	47	4,847	4,305.8576
7	52	5,288	21,742.3778
8	57	5,106	46,524.0584
9	62	3,712	52,382.4255
10	67	2,401	45,887.0592
11	72	1,822	43,931.4168
	Suma=	39,348	420,934.6755
		Desv Prom=	**10.6977**

Varianza o variancia para datos agrupados en una tabla de frecuencias

Se calcula el concepto de varianza o variancia para datos agrupados en una tabla de frecuencias de la siguiente forma:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j \tag{3.41}$$

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{j=1}^{j=m} t_j^2 f_j - \frac{n}{n-1}\bar{x}^2 \tag{3.59}$$

$$S_{n-1}^2 = \frac{n}{n-1} \sum_{j=1}^{j=m} \left(t_j - \bar{x}\right)^2 f_j^* \tag{3.60}$$

$$S_{n-1}^2 = \frac{n}{n-1} \left[\sum_{j=1}^{j=m} t_j^2 f_j^* - \bar{x}^2 \right] \tag{3.61}$$

Calcule la varianza de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la varianza de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$\lvert t_j\text{-media}\rvert f_j$	$(t_j\text{-media})^2 f_j$
1	83.37	10	60.9156	371.0705
2	84.91	10	45.5156	207.1666
3	86.45	10	30.1156	90.6947
4	87.99	12	17.6587	25.9857
5	89.53	12	0.8213	0.0562
6	91.07	9	14.4760	23.2838
7	92.61	7	22.0391	69.3889
8	94.15	9	42.1960	197.8336
9	95.69	7	43.5991	271.5546
10	97.23	4	31.0738	241.3949
	Suma=	90	308.4107	1,498.4296
			Varianza=	**16.6492**
			Desv Est=	4.0803
			Coef. Variación=	0.0456

Ejercicio 3.3 Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$\lvert t_j\text{-media}\rvert f_j$	$(t_j\text{-media})^2 f_j$
1	22	896	23,195.9662	600,505.41
2	27	2,356	49,212.9648	1,027,977.89
3	32	3,539	56,228.8890	893,384.56
4	37	4,457	48,529.3991	528,405.33
5	42	4,924	28,994.2610	170,728.51
6	47	4,847	4,305.8576	3,825.13
7	52	5,288	21,742.3778	89,396.93
8	57	5,106	46,524.0584	423,910.70
9	62	3,712	52,382.4255	739,202.18
10	67	2,401	45,887.0592	876,977.18
11	72	1,822	43,931.4168	1,059,258.72
	Suma=	39,348	420,934.6755	6,413,572.5443
			Varianza=	**162.9962**
			Desv Est=	12.7670
			Coef. Variación=	0.2666

En el ejercicio 3.1, para calcular la varianza se dividió entre n-1, ya que se trata de una muestra, pero en el ejercicio 3.3 se dividió entre N, ya que se trata de la población.

Desviación estándar para datos agrupados en una tabla de frecuencias

Se define la desviación estándar de los datos de una muestra como la raíz cuadrada de la varianza muestral, es decir,

$$S_{n-1} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{j=m} \left(t_j - \bar{x}\right)^2 f_j} \tag{3.62}$$

$$S_{n-1} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{j=m} t^2{}_j f_j - \frac{n}{n-1} \bar{x}^2} \tag{3.63}$$

$$S_{n-1} = \sqrt{\frac{n}{n-1} \sum_{j=1}^{j=m} \left(t_j - \bar{x}\right)^2 f_j^*} \tag{3.64}$$

$$S_{n-1} = \sqrt{\frac{n}{n-1} \left[\sum_{j=1}^{j=m} t_j^2 f_j^* - \bar{x}^2 \right]} \tag{3.65}$$

Calcule la desviación estándar de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y la desviación estándar de la tabla 3.45, correspondiente al problema 3.3.

Para el ejercicio 3.1, Sn – 1 = 4.0803
Para el ejercicio 3.3, Sn = 12.767

Coeficiente de variación para datos agrupados en una tabla de frecuencias

El cálculo se haría con la misma definición:

$$CV = \frac{S_{n-1}}{\bar{x}}$$

Calcule el coeficiente de variación de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el coeficiente de variación de la tabla 3.45, correspondiente al problema 3.3.

Para el ejercicio 3.1, CV = 0.0456
Para el ejercicio 3.3, CV = 0.2667

Momentos de Orden k con respecto al origen y con respecto a la media aritmética para datos agrupados en una tabla de frecuencias

Los momentos muestrales de orden k con respecto al origen y de orden k con respecto a la media aritmética para datos agrupados en una tabla de frecuencias se calculan con las siguientes expresiones matemáticas:

$$m'_k = \frac{1}{n}\sum_{j=1}^{j=m} t_j^k f_j = \sum_{j=1}^{j=m} t_j^k f_j^* \tag{3.66}$$

$$m_k = \frac{1}{n}\sum_{j=1}^{j=m} (t_j - \bar{x})^k f_j = \sum_{j=1}^{j=m} (t_j - \bar{x})^k f_j^* \tag{3.67}$$

Como se puede apreciar, al momento de orden $k = 1$ con respecto al origen se le conoce como media aritmética. El momento de orden $k = 1$ con respecto a la media es cero. El momento de segundo orden con respecto a la media tiene una relación directa con el concepto de varianza:

$$m_2 = \frac{1}{n}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j = \left(\frac{n-1}{n}\right) S_{n-1}^2 \tag{3.68}$$

Figura 3.48. Formulario para calcular medidas de dispersión de una muestra

CÁLCULO DE LOS PARÁMETROS ESTADÍSTICOS DE UNA MUESTRA			
Medidas	Parámetro Estadístico	Datos Sin Agrupar	Datos Agrupados
de Dispersión	Rango	$R = x_{max} - x_{min}$	$R = LSI_m - LII_1$
	Desviación Promedio	$Desv\,Prom = \frac{1}{n}\sum_{i=1}^{i=n} \lvert x_i - \bar{x} \rvert$	$Desv\,Prom = \frac{1}{n}\sum_{j=1}^{j=m} \lvert t_j - \bar{x} \rvert f_j$ $Desv\,Prom = \sum_{j=1}^{j=m} \lvert t_j - \bar{x} \rvert f_j^*$ *donde* $f_j^* = \frac{f_j}{n}$
	Varianza	$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n} (x_i - \bar{x})^2$	$S_{n-1}^2 = \frac{1}{n}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j$ $S_{n-1}^2 = \frac{1}{n-1}\sum_{j=1}^{j=m} t_j^2 f_j - \frac{n}{n-1}\bar{x}^2$ $S_{n-1}^2 = \frac{n}{n-1}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j^*$ $S_{n-1}^2 = \frac{n}{n}\left[\sum_{j=1}^{j=m} t_j^2 f_j^* - \bar{x}^2\right]$
	Desviación Estándar	$S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n} (x_i - x)^2}$	$S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j}$ $S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{j=m} t_j^2 f_j - \frac{n}{n-1}\bar{x}^2}$ $S_{n-1} = \sqrt{\frac{n}{n-1}\sum_{j=1}^{j=m} (t_j - \bar{x})^2 f_j^*}$ $S_{n-1} = \sqrt{\frac{n}{n-1}\left[\sum_{j=1}^{j=m} t_j^2 f_j^* - \bar{x}^2\right]}$
	Coeficiente de Variación	$CV = \frac{S_{n-1}}{\bar{x}}$	$CV = \frac{S_{n-1}}{\bar{x}}$
	Error Estándar	$EE = \frac{S_{n-1}}{\sqrt{n}}$	$EE = \frac{S_{n-1}}{\sqrt{n}}$

3.4.3. Medidas de forma para datos agrupados en una tabla de frecuencias

El coeficiente de asimetría de Fisher para datos agrupados en una tabla de frecuencias, representado por γ_1, se definió como:

$$\gamma_1 = \frac{m_3}{S^{3/2}_{n-1}}$$

Calcule el coeficiente de asimetría de Fisher de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el coeficiente de asimetría de Fisher de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$(t_j\text{-media})^2 f_j$	$(t_j\text{-media})^3 f_j$
1	83.37	10	371.0705	-2,260.3965
2	84.91	10	207.1666	-942.9302
3	86.45	10	90.6947	-273.1320
4	87.99	12	25.9857	-38.2394
5	89.53	12	0.0562	0.0038
6	91.07	9	23.2838	37.4508
7	92.61	7	69.3889	218.4672
8	94.15	9	197.8336	927.5319
9	95.69	7	271.5546	1,691.3630
10	97.23	4	241.3949	1,875.2630
	Suma=	90	1,498.4296	1,235.3815
		Varianza=	16.6492	
		Desv Est=	4.0803	
		CV=	0.0456	
		Coef Asimetría=		**0.2021**

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$(t_j\text{-media})^2 f_j$	$(t_j\text{-media})^3 f_j$
1	22	896	600,505.41	-15,546,097.43
2	27	2,356	1,027,977.89	-21,472,767.28
3	32	3,539	893,384.56	-14,194,411.22
4	37	4,457	528,405.33	-5,753,464.96
5	42	4,924	170,728.51	-1,005,310.09
6	47	4,847	3,825.13	-3,398.07
7	52	5,288	89,396.93	367,568.44
8	57	5,106	423,910.70	3,862,523.69
9	62	3,712	739,202.18	10,431,358.66
10	67	2,401	876,977.18	16,760,476.31
11	72	1,822	1,059,258.72	25,540,469.99
	Suma=	39,348	6,413,572.5443	-1,013,051.9709
		Varianza=	162.9962	
		Desv Est=	12.7670	
		CV=	0.2666	
		Coef Asimetría=		**-0.0124**

Coeficiente de asimetría de Pearson para datos agrupados en una tabla de frecuencias

Se definió como:

$$C_{AP} = \frac{\overline{x} - m_o}{S_{n-1}}$$

En donde m_o representa a la moda de la muestra.

Calcule el coeficiente de asimetría de Pearson de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el coeficiente de asimetría de Pearson de la tabla 3.45, correspondiente al problema 3.3.

Para el ejercicio 3.1, CAP = (89.4616–88.4520)/4.0803 = 0.254
Para el ejercicio 3.3, CAP = (47.8884–53.5393)/12.767 = –0.4426

Coeficiente de asimetría de Bowley-Yule para datos agrupados en una tabla de frecuencias

$$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}$$

Calcule el coeficiente de asimetría de Bowley-Yule de la tabla de frecuencias de la figura 3.31, correspondiente al ejercicio 3.1 y el coeficiente de asimetría de Bowley-Yule de la tabla 3.45, correspondiente al ejercicio 3.3.

Ejercicio 3.1

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	82.6	84.14	83.37	10	10	0.1111	0.1111
2	84.14	85.68	84.91	10	20	0.1111	0.2222
3	85.68	87.22	86.45	10	30	0.1111	0.3333
4	87.22	88.76	87.99	12	42	0.1333	0.4667
5	88.76	90.3	89.53	12	54	0.1333	0.6000
6	90.3	91.84	91.07	9	63	0.1000	0.7000
7	91.84	93.38	92.61	7	70	0.0778	0.7778
8	93.38	94.92	94.15	9	79	0.1000	0.8778
9	94.92	96.46	95.69	7	86	0.0778	0.9556
10	96.46	98	97.23	4	90	0.0444	1.0000
			Suma=	90		1.0000	

1er Cuartil= 86.065
3er Cuartil= 92.83

$$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}$$

$$q_1 = \frac{LSq_1 - LIq_1}{F^*_{LSq_1} - F^*_{LIq_1}}\left(0.25 - F^*_{LIq_1}\right) + LIq_1$$

$$q_2 = \frac{LSq_2 - LIq_2}{F^*_{LSq_2} - F^*_{LIq_2}}\left(0.50 - F^*_{LIq_2}\right) + LIq_2$$

$$q_3 = \frac{LSq_3 - LIq_3}{F^*_{LSq_3} - F^*_{LIq_3}}\left(0.75 - F^*_{LIq_3}\right) + LIq_3$$

Mediana= 89.1450
Moda = 88.4520
Coef Asim Bowley-Yule= 0.0894

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Límite Inferior Intervalo	Límite Superior Intervalo	Marca de Clase t_j	Frecuencia Absoluta f_j	Frecuencia Absoluta Acumulada F_j	Frecuencia Relativa f^*_j	Frecuencia Relativa Acumulada F^*_j
1	menor a 24	24	22	896	896	0.0228	0.0228
2	25	29	27	2,356	3,252	0.0599	0.0826
3	30	34	32	3,539	6,791	0.0899	0.1726
4	35	39	37	4,457	11,248	0.1133	0.2859
5	40	44	42	4,924	16,172	0.1251	0.4110
6	45	49	47	4,847	21,019	0.1232	0.5342
7	50	54	52	5,288	26,307	0.1344	0.6686
8	55	59	57	5,106	31,413	0.1298	0.7983
9	60	64	62	3,712	35,125	0.0943	0.8927
10	65	69	67	2,401	37,526	0.0610	0.9537
11	70	o más	72	1,822	39,348	0.0463	1.0000
			Suma=	39,348		1.0000	

1er Cuartil= 38.4171
3er Cuartil= 58.1375

$$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}$$

$$q_1 = \frac{LSq_1 - LIq_1}{F^*_{LSq_1} - F^*_{LIq_1}}\left(0.25 - F^*_{LIq_1}\right) + LIq_1$$

$$q_2 = \frac{LSq_2 - LIq_2}{F^*_{LSq_2} - F^*_{LIq_2}}\left(0.50 - F^*_{LIq_2}\right) + LIq_2$$

$$q_3 = \frac{LSq_3 - LIq_3}{F^*_{LSq_3} - F^*_{LIq_3}}\left(0.75 - F^*_{LIq_3}\right) + LIq_3$$

Mediana= 48.6125
Moda = 53.5393
Coef Asim Bowley-Yule= -0.0340

Medidas de curtosis o aplanamiento para datos agrupados en una tabla de frecuencias

El coeficiente de curtosis se define como:

$$\gamma_2 = \frac{\overline{m_4}}{S^4_{n-1}} - 3$$

Calcule el coeficiente de curtosis de la tabla de frecuencias de la figura 3.31, correspondiente al problema 3.1 y el coeficiente de curtosis de la tabla 3.45, correspondiente al problema 3.3.

Ejercicio 3.1

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$(t_j\text{-media})^2 f_j$	$(t_j\text{-media})^4 f_j$
1	83.37	10	371.0705	13,769.3309
2	84.91	10	207.1666	4,291.7992
3	86.45	10	90.6947	822.5523
4	87.99	12	25.9857	56.2714
5	89.53	12	0.0562	0.0003
6	91.07	9	23.2838	60.2375
7	92.61	7	69.3889	687.8317
8	94.15	9	197.8336	4,348.6816
9	95.69	7	271.5546	10,534.5605
10	97.23	4	241.3949	14,567.8764
	Suma=	90	1,498.4296	49,139.1417
		Varianza=	16.83628744	
			Coef. Curtosis=	**-1.0738**

Ejercicio 3.3. Personal Académico en la UNAM por Edad en el 2015

No. Clase	Marca de Clase t_j	Frecuencia Absoluta f_j	$(t_j\text{-media})^2 f_j$	$(t_j\text{-media})^4 f_j$
1	22	896	600,505.41	402,462,892.11
2	27	2,356	1,027,977.89	448,530,789.79
3	32	3,539	893,384.56	225,525,847.17
4	37	4,457	528,405.33	62,645,770.10
5	42	4,924	170,728.51	5,919,622.89
6	47	4,847	3,825.13	3,018.70
7	52	5,288	89,396.93	1,511,310.88
8	57	5,106	423,910.70	35,193,943.91
9	62	3,712	739,202.18	147,203,628.25
10	67	2,401	876,977.18	320,320,270.02
11	72	1,822	1,059,258.72	615,822,740.79
		Suma=	39,348	6,413,572.5443	2,265,139,835
			Varianza=	162.9962	
				Coef. Curtosis=	**-0.8332**

Con relación al ejercicio 3.1, se calcularon sus parámetros estadísticos utilizando los datos de la tabla dada y se calcularon dichos parámetros estadísticos para datos agrupados en una tabla de frecuencias. Los resultados se muestran en un cuadro comparativo.

Estadístico	*Datos Sin agrupar*	*Datos Agrupados*	*% Error Abs*
Media aritmética	89.4756	89.4616	0.02%
Media geométrica	89.3806	89.3689	0.01%
Media armónica	89.2863	89.2768	0.01%
Semirango	90.3	90.3	0.00%
Mediana	89.25	89.145	0.12%
Moda	87.3	88.452	1.32%
Rango	15.4	15.4	0.00%
Desviación promedio	3.5172	3.4268	2.57%
Desviación estándar	4.1578	4.1032	1.31%
Error típico	0.4383	0.4325	1.31%
Varianza de la muestra	17.2870	16.8363	2.61%
Coef. Variación	0.0465	0.0459	1.30%
Curtosis	–1.0013	–1.0738	7.24%
Coef. Asim. Fisher	0.2554	0.1987	22.21%
Coef. Asim. Pearson	0.5233	0.2010	61.58%
Tercer cuartil	93.1	92.83	0.29%
Segundo Cuartil	89.25	89.145	0.12%
Primer cuartil	86.175	86.065	0.13%
Coef. Asim. Bowley- Yule	0.1119	0.0894	20.09%
Rango	15.4	15.4	0.00%
Máximo	98	98	0.00%
Mínimo	82.6	82.6	0.00%
Suma	8,052.80	8,051.54	0.02%
Cuenta	90	90	0.00%

Como se puede observar, los resultados entre datos sin agrupar y datos agrupados en una tabla de frecuencias son diferentes, esto se debe a que en cada subintervalo o clase los datos correspondientes se están sustituyendo por el valor de la marca de clase, lo que implica que los resultados obtenidos con la tabla de frecuencias son resultados aproximados de los datos sin agrupar; sin embargo, la aproximación es bastante adecuada en una parte considerable de los parámetros, como se aprecia en el cuadro anterior.

Figura 3.49. Formulario para calcular otras medidas de una muestra

CÁLCULO DE LOS PARÁMETROS ESTADÍSTICOS DE UNA MUESTRA			
Medidas	Parámetro Estadístico	Datos Sin Agrupar	Datos Agrupados
	Error Estándar	$EE = \frac{S_{n-1}}{\sqrt{n}}$	$EE = \frac{S_{n-1}}{\sqrt{n}}$
de Forma	Coeficiente de Asimetría de Fisher	$\gamma_1 = \frac{m_3}{S_{n-1}^{3/2}}$	$\gamma_1 = \frac{m_3}{S_{n-1}^{3/2}}$
	Coeficiente de Asimetría de Pearson	$C_{AP} = \frac{\bar{x} - m_o}{S_{n-1}}$	$C_{AP} = \frac{\bar{x} - m_o}{S_{n-1}}$
	Coeficiente de Asimetría de Bowley- Yule	$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}$	$C_{ABY} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}$
	Coeficiente de Curtosis	$\gamma_2 = \frac{m_4}{S_{n-1}^4} - 3$	$\gamma_2 = \frac{m_4}{S_{n-1}^4} - 3$
Otras medidas	Valor Máximo	Se ordenan los valores de menor a mayor y se toma el último valor	LSI_m
	Tercer Cuartil	Es el valor de x que parte a los datos de tal manera que el 75% de ellos cae a la izquierda y el 25% a la derecha	$q_3 = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}}\left(0.75 - F^*_{LII}\right) + LII$
	Primer Cuartil	Es el valor de x que parte a los datos de tal manera que el 25% de ellos cae a la izquierda y el 75% cae a la derecha	$q_1 = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}}\left(0.25 - F^*_{LII}\right) + LII$
	Percentil p	Es el valor de x que parte a los datos de tal manera que el p% de ellos cae a la izquierda y el (1-p)% cae a la derecha	$q_p = \frac{LSI - LII}{F^*_{LSI} - F^*_{LII}}\left(p - F^*_{LII}\right) + LII$
	Valor Mínimo	Se ordenan los valores de menor a mayor y se toma el primer valor	LII_1
	Momentos de orden k con respecto a la media	$m_k = \frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^k$	$m_k = \frac{1}{n}\sum_{j=1}^{j=m}\left(t_j - \bar{x}\right)^k f_j = \sum_{j=1}^{j=m}\left(t_j - \bar{x}\right)^k f_j^*$
	Momentos de orden k con respecto al origen	$m'_k = \frac{1}{n}\sum_{i=1}^{i=n} x_i^k$	$m'_k = \frac{1}{n}\sum_{j=1}^{j=m} t_j^k f_j = \sum_{j=1}^{j=m} t_j^k f_j^*$
	Suma	$\sum_{i=1}^{i=n} x_i$	$\sum_{j=1}^{j=m} t_j f_j$
	Suma de Cuadrados	$\sum_{i=1}^{i=n} x_i^2$	$\sum_{j=1}^{j=m} t_j^2 f_j$
	Cuenta	n	$\sum_{j=1}^{j=m} f_j$

3.5. Otro tipo de diagramas: diagrama de tallo y hojas, diagrama de caja y diagrama de Pareto. Series de tiempo y sus gráficas

3.5.1. Diagrama de tallo y hojas

El diagrama de tallo y hojas (Stem-and-Leaf Diagram) permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta dividir cada dato en dos partes: un tallo, compuesto por uno o más de los primeros dígitos, y una hoja, que consiste en los dígitos restantes. Esta representación de los datos es semejante a la de un histograma, pero además de ser fáciles de elaborar, presentan más información que estos.

El software Minitab, ya trae integrado dicho diagrama. Para utilizar Minitab para obtener el diagrama de tallo y hojas se aplicarán los siguientes pasos al ejercicio 3.1:

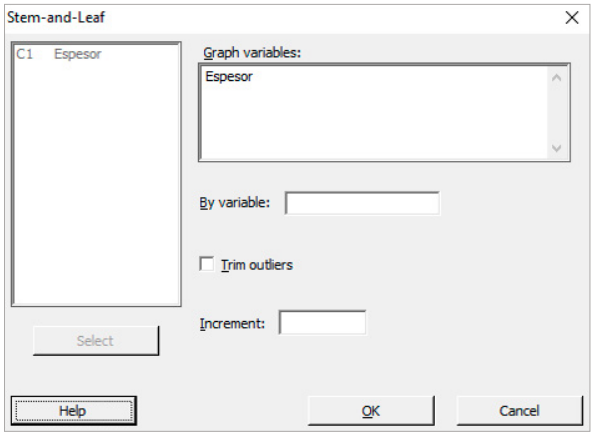
a. Se ordenan los 90 datos de la muestra de menor a mayor.

b. Se toma como tallo el valor entero de cada una de las lecturas, quedando como tallos 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98. El decimal de cada número formará las hojas de cada tallo, por ejemplo, el número 82.6 tendrá como tallo el número 82 y como hoja el seis. Se apilan todas las hojas a la derecha de cada tallo.

c. Se copian todos los datos del Ejercicio 3.1 en la primera columna de la hoja electrónica de Minitab, por ejemplo, en C1.

d. Se ordenan de menor a mayor, la parte entera del número será el tallo y el resto de dígitos serán las hojas.

e. En el menú principal de Minitab, se da un click en Graph, y luego en el submenú Stem and Leaf, obteniéndose la siguiente ventana.



En esta ventana seleccione la variable C1 Espesor y luego dé un click en Select, finalmente en Ok, obteniéndose el diagrama de Tallo y Hojas que se muestra:

Stem-and-leaf of Espesor $n = 90$
Leaf Unit = 0.10

82 69
83 0167
84 01112569
85 011144
86 1114444667
38 87 33335667
43 88 22368
89 114667
90 0011345666
91 1247
92 144
93 11227
94 11133467
95 1236
97 38
98 0

3.5.2. Diagrama de caja y bigotes

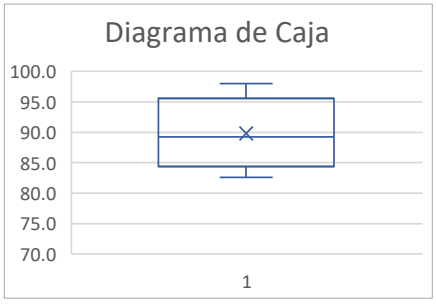
También conocido como diagrama de caja y bigote, o box plot. Es un método estandarizado para representar gráficamente los datos numéricos de una muestra a través de sus cuartiles y sus valores mínimo y máximo.

Con Excel, los pasos que se siguen son:

1. Se obtienen el valor máximo de la muestra, el tercer cuartil, el segundo cuartil o mediana, el primer cuartil y el valor mínimo de la muestra.

   Valor Máximo = Max($A$2:$A$91) = 98.0
   Tercer Cuartil = Cuartil($A$2:$A$91,3) = 93.1
   Mediana = Cuartil($A$2:$A$91,2) = 89.25
   Primer Cuartil = Cuartil($A$2:$A$91,1) = 86.175
   Valor Mínimo = Min($A$2:$A$91) = 82.6

2. Se elige el menú insertar, luego gráfico, se elige el submenú de Histograma y posteriormente Caja y bigotes:



La mejor aplicación de un diagrama de caja se lleva a cabo cuando se obtienen muestras periódicas de un proceso y se pretende conocer cómo cambia este proceso a lo largo del tiempo, como se muestra en el siguiente ejercicio.

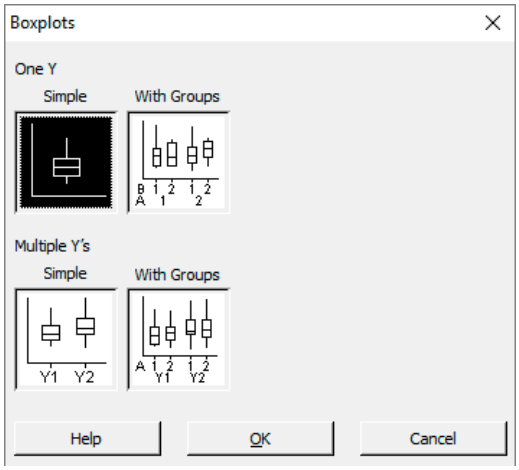
Ejercicio 3.5

Durante el proceso de manufactura de frascos de nescafé de 200 gramos, en la etapa de llenado del frasco, suponga que se producen 5000 frascos por turno de 8 horas y la característica a controlar es el peso de café que se descarga en cada frasco. La especificación del peso de café que debe tener cada frasco es de 200 ± 5 gr. Suponga que el supervisor de calidad decidió tomar las siguientes muestras de tamaño n = 8 cada media hora, de las 07:00 a las 11:00 horas.

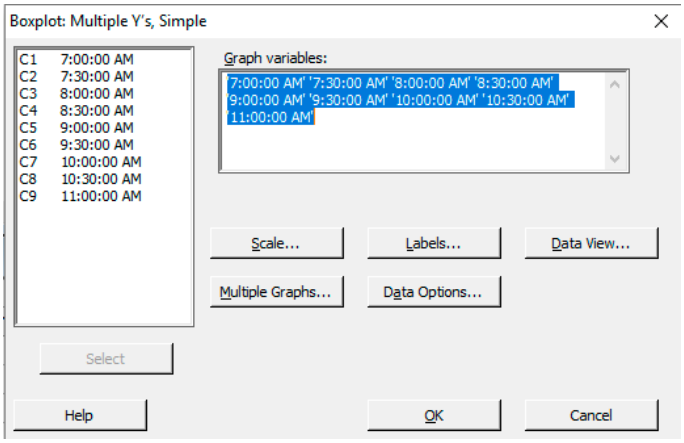
Muestra	07:00	07:30	08:00	08:30	09:00	09:30	10:00	10:30	11:00
x_1	200.5	201.9	202.1	200.4	198.4	199.7	199.4	199.1	200.9
x_2	204.4	199.4	201.5	198.9	198.9	195.1	200.2	195.6	198.2
x_3	199.1	201.5	205.3	199.6	200.7	197.0	199.4	198.7	200.0
x_4	201.2	200.8	200.6	198.5	198.0	199.8	199.9	196.4	201.0
x_5	200.2	204.5	201.5	200.5	199.9	201.6	200.1	194.4	198.8
x_6	200.9	203.3	205.4	200.4	201.6	196.5	197.5	197.3	201.8
x_7	202.1	204.3	199.7	197.5	199.9	203.7	197.9	197.6	195.2
x_8	200.5	202.0	201.2	199.3	202.8	200.6	200.3	199.1	203.2

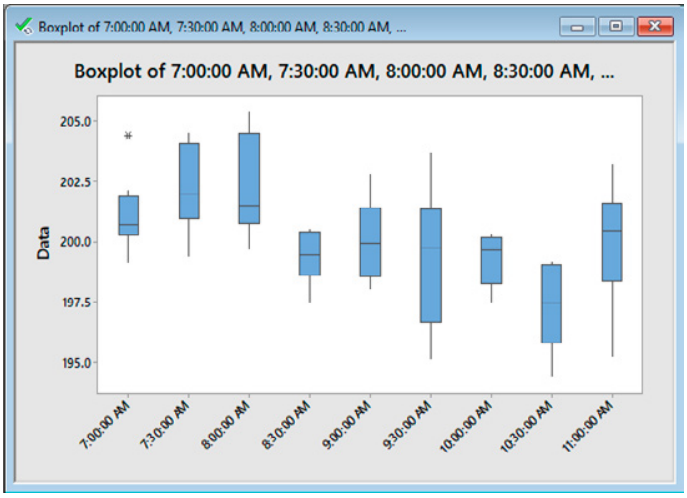
Elabore un diagrama de caja de las 9 muestras de tamaño n = 8, tomadas cada media hora, utilizando el software Minitab.

1. Se supondrá que cada columna es una muestra tomada, se copia la tabla en Minitab y se entra al menú Graph, luego al submenú Boxplot y aparece la siguiente pantalla:



2. Se selecciona la opción Multiple Y's dado que se van a graficar diversas cajas secuencialmente en la misma figura. Aparece la siguiente pantalla en la cual deben seleccionarse las nueve muestras de tamaño n = 8 cada una, después oprimir select y finalmente Ok, para que se muestre el Diagrama de cajas solicitado:





3.5.3. Diagrama de Pareto

El autor del diagrama fue el experto en calidad Dr. Joseph Juran (1904 - 2008) y le puso ese nombre en honor del economista italiano Vilfredo Pareto (1848-1923), quien realizó un estudio sobre la distribución de la riqueza, en el cual descubrió que la minoría de la población poseía la mayor parte de la riqueza y la mayoría de la población poseía la menor parte de la riqueza. Con esto estableció la llamada "Ley de Pareto", según la cual la desigualdad económica es inevitable en cualquier sociedad. De aquí surge la llamada regla del 80-20 que establece que el 80% de los defectos en una empresa se debe al 20% de los problemas de esta; a este 20% de los problemas es a lo que se le denomina "los pocos vitales" y al complemento del 80% se le conoce como "los muchos triviales", como se ilustra en la figura 3.36.

Un Diagrama de Pareto es una forma especial de gráfico de barras usado para ilustrar datos ordenados por categorías en forma descendente de derecha a izquierda, de acuerdo con su importancia. La altura de las barras representa la frecuencia o importancia relativa de los puntos que están siendo medidos.

Un diagrama de Pareto ilustra visualmente los datos con el propósito de:

» Ayudar a establecer prioridades.
» Ilustrar las oportunidades más significativas para mejorar.
» Mostrar qué categorías contribuyen con el mayor porcentaje del total.

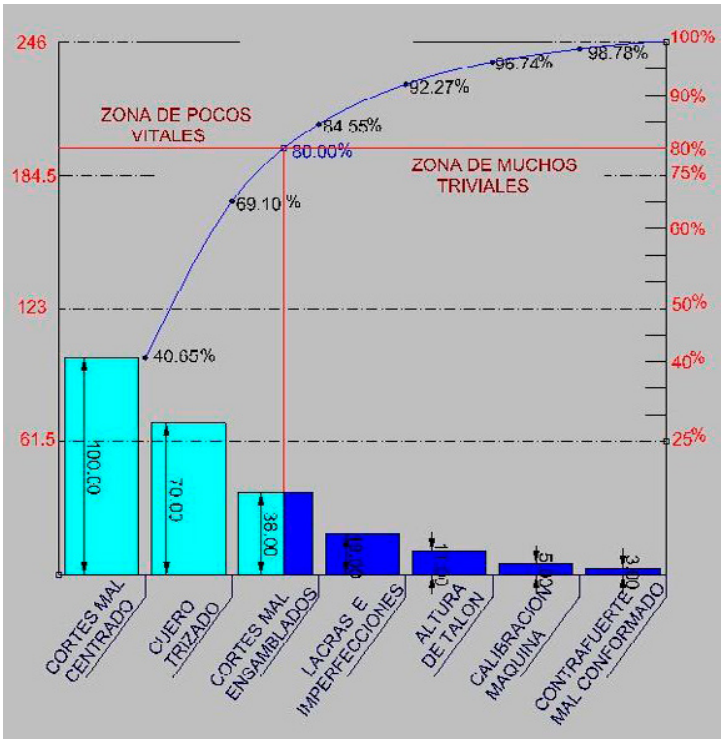


Figura 3.50. Ejemplo de diagrama de Pareto

Fuente: http://ddeprocal.blogspot.com/2015/08/diagrama-de-pareto-80-20_80.html

Los pasos que se siguen para construir un diagrama de Pareto son los siguientes:

1. Antes de coleccionar los datos se deben seleccionar las categorías de defectos, problemas o causas, etcétera, a ser comparados. Los problemas deben estar definidos lo más objetivamente posible y deben poder medirse de alguna forma.

2. Diseñar una hoja de verificación en la cual registrar los datos perfectamente y llevar a cabo su recolección.

3. Construir una tabla de frecuencias, como la que se muestra de ejemplo debajo de la figura 3.51:

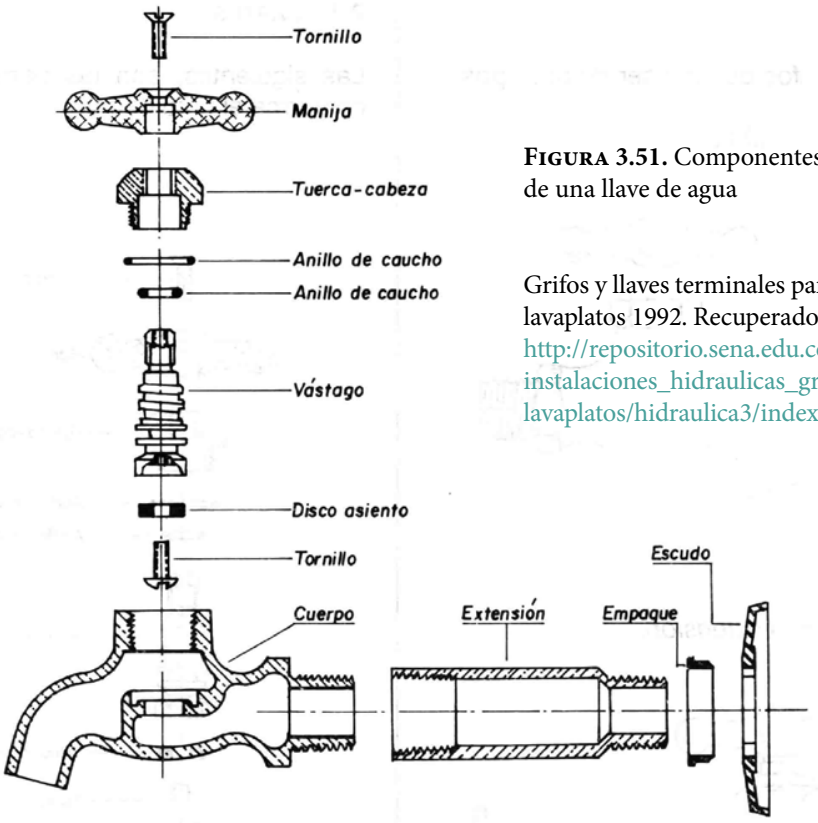


Figura 3.51. Componentes de una llave de agua

Grifos y llaves terminales para lavaplatos 1992. Recuperado de http://repositorio.sena.edu.co/sitios/instalaciones_hidraulicas_griferias_lavaplatos/hidraulica3/index.html

Piezas defectuosas (Llaves de agua)

Defecto encontrado	# piezas defectuosas	% piezas defectuosas
Rosca golpeada	216	48.0%
Empaque roto	31	6.9%
Mariposa suelta	108	24.0%
Salida ovalada	14	3.1%
Cuerpo poroso	81	18.0%
Total	450	

Cantidad de piezas inspeccionadas: 5843, del 1º al 15 de abril.

Estos datos se pueden colocar en una gráfica de barras como la que se muestra en la figura 3.52.

4. En un sistema cartesiano indicar sobre el eje horizontal las categorías, dividiendo dicho eje en segmentos iguales y ordenando dichas categorías en orden descendente de acuerdo con la importancia relativa, la cual será indicada sobre el eje horizontal. El eje vertical de la gráfica indicará los porcentajes de los valores de la característica que representa la importancia relativa de esta.

5. Dibujar barras para cada categoría, con una altura igual al valor que tome la característica que representa la importancia relativa. En cada extremo derecho de las barras se dibuja, por medio de un gráfico de líneas, la frecuencia relativa acumulada. Cada barra representa un tipo de defecto encontrado, el eje vertical muestra la importancia de cada defecto encontrado en términos de porcentaje, el eje horizontal muestra los tipos de defectos encontrados comenzando con el de mayor importancia a la izquierda, hasta el de menor importancia, que es el último a la derecha y los que quedan en medio se acomodan por orden de magnitud. En este caso el Diagrama de Pareto establece prioridades sobre los problemas que se deben analizar primero para eliminar la mala operación. Para este caso y de acuerdo con la gráfica 3.38 se debe tratar de resolver el problema de la rosca golpeada en primer lugar porque es la barra más alta; el segundo problema que se debe tratar de resolver es el de mariposa suelta, porque es la siguiente barra más alta. En esta forma la planeación de los problemas por atacar es más sencilla.

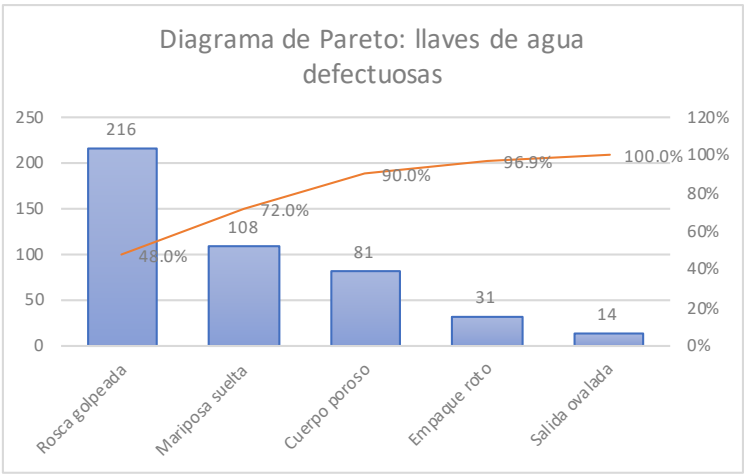


Figura 3.52. Diagrama de Pareto de llaves de agua defectuosas

Usar escalas de medida diferentes en el eje vertical de un diagrama de Pareto puede ayudar a identificar los principales problemas y sus prioridades. Los problemas más frecuentes no siempre son los más costosos. De hecho, un problema

que ocurre raras veces puede ser la mejor oportunidad para mejorar. Por ello se aconseja tomar diferentes escalas como pueden ser frecuencia de defectos, frecuencia relativa con respecto al total de productos elaborados, costos, etc.

En la figura 3.53 se muestra un ejemplo de diagrama de Pareto con un gran número de errores, pero con bajo costo. En la figura 3.54 se muestra un diagrama de Pareto con un pequeño número de errores, pero con altos costos. Finalmente, si en el diagrama de Pareto no se muestra la oportunidad para mejorar un proceso, entonces puede ser necesario reagrupar los datos y volver a trazar el diagrama, como se muestra en la figura 3.55, en donde se aprecian tres diferentes maneras de ver un diagrama de Pareto del mismo conjunto de datos, pero solo en uno de ellos se muestra una clara diferencia.

Figura 3.53. Nótese que el problema más frecuente no es el más costoso

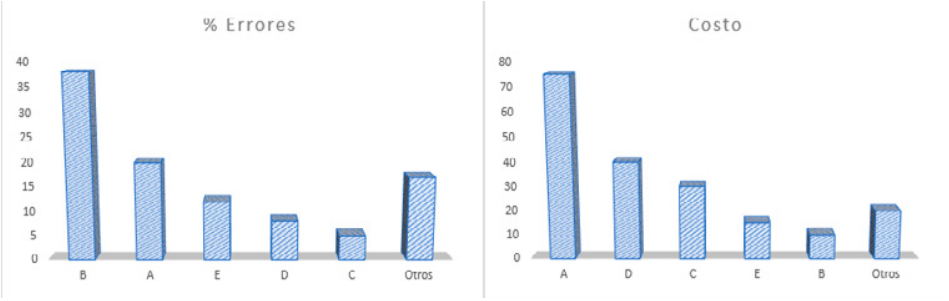


Figura 3.54. Nótese que el problema más costoso no es el más frecuente

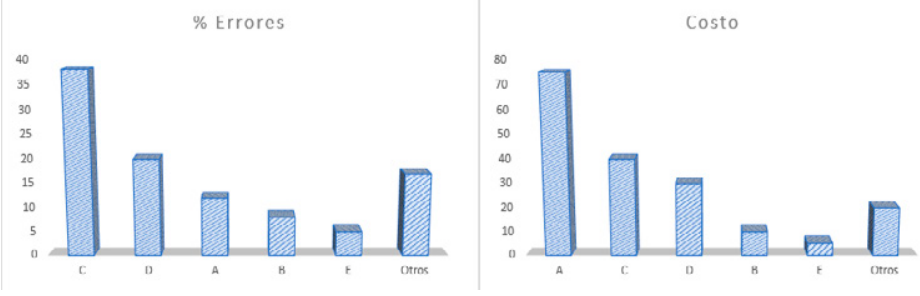
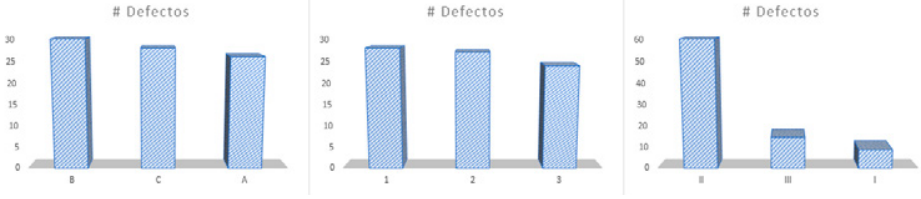
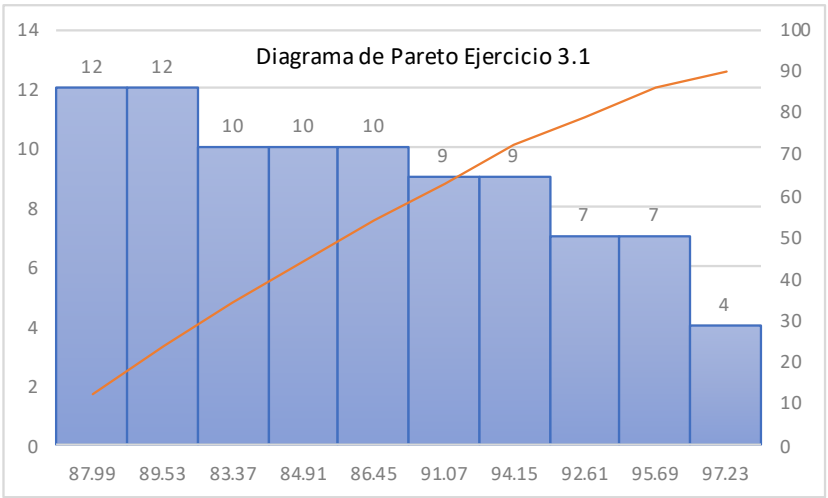


Figura 3.55. Diferentes formas de agrupar el mismo conjunto de datos



Tipo de defecto | Técnico | Desplazamiento

Con relación a la muestra del Ejercicio 3.1, trace el Diagrama de Pareto.



3.5.4. Series de tiempo

Una serie de tiempo, temporal o cronológica, es una secuencia de datos, observaciones o valores de una característica de interés particular, recopilados, obtenidos o medidos en determinados momentos y ordenados cronológicamente. Los datos pueden haberse obtenido a intervalos iguales de tiempo (como el salario tabular quincenal de un académico de la UNAM) o desiguales (como la medición de la presión arterial de una persona cada vez que asiste no regularmente con su cardiólogo).

Son innumerables las aplicaciones de las series de tiempo en muy diversos campos, por ejemplo: en Economía (índice de precios, tasas de desempleo, tasas de interés y de inflación, etcétera); en Física (Meteorología, precipitación pluvial, temperatura diaria, velocidad del viento, radiación solar, señales sísmicas, magnéticas, eléctricas, gravitacionales, etcétera); en Demografía (tasa de natalidad, tasa de mortalidad, tasa de crecimiento poblacional, censos, etcétera); en Telecomunicaciones (procesamiento de señales eléctricas; en Ingeniería de Tránsito (flujo vehicular diario).

Para ilustrar cómo es una serie de tiempo, obsérvese la figura 3.56, y su gráfico correspondiente figura (3.57), así como la figura 3.58, obtenidos de "Las Estadísticas del Personal Académico de la UNAM, 2018", elaborado por la Dirección General de Asuntos del Personal Académico de la UNAM, páginas 171, 172 y 180. http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2018.pdf

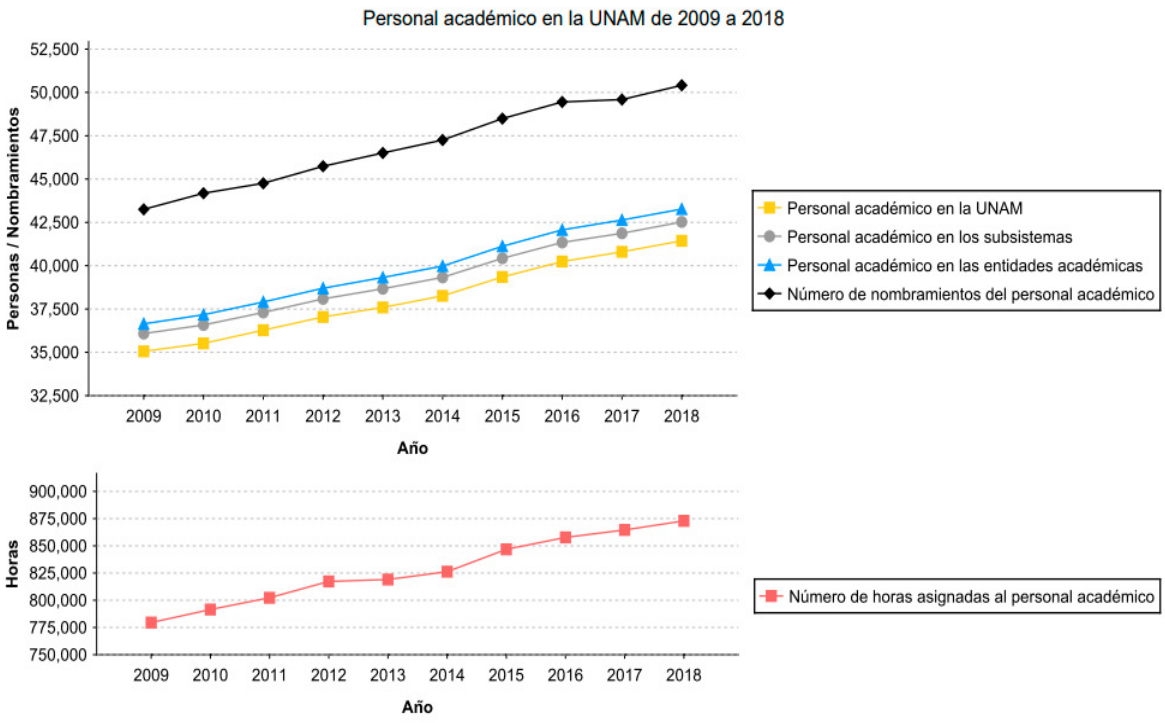
Figura 3.56.

Personal académico en la UNAM de 2009 a 2018

Año	Personal académico en la UNAM	Personal académico en los subsistemas	Personal académico en las entidades académicas	Número de nombramientos del personal académico	Número de horas asignadas al personal académico
2009	35,057	36,081	36,641	43,252	779,531.0
2010	35,516	36,578	37,170	44,181	791,493.0
2011	36,278	37,299	37,904	44,756	802,292.0
2012	37,042	38,086	38,699	45,737	817,302.0
2013	37,592	38,670	39,319	46,507	819,059.0
2014	38,259	39,325	39,978	47,253	826,238.0
2015	39,348	40,427	41,119	48,489	846,770.5
2016	40,240	41,342	42,066	49,449	857,730.5
2017	40,800	41,869	42,635	49,588	864,584.5
2018	41,436	42,525	43,268	50,412	872,931.5

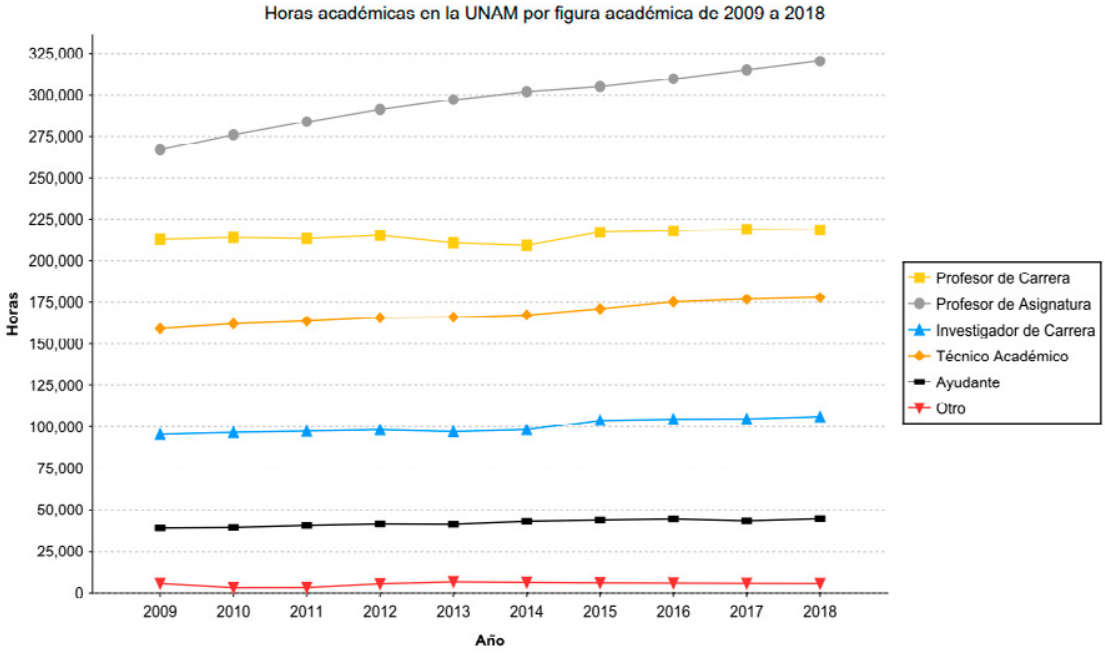
"Estadísticas del Personal Académico de la UNAM, 2018", elaborado por la Dirección General de Asuntos del Personal Académico de la Universidad Nacional Autónoma de México, DGAPA UNAM, página 171, http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2018.pdf

Figura 3.57.



"Estadísticas del Personal Académico de la UNAM, 2018", elaborado por la Dirección General de Asuntos del Personal Académico de la Universidad Nacional Autónoma de México, DGAPA UNAM, página 172 http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2018.pdf

Figura 3.58.



"Estadísticas del Personal Académico de la UNAM, 2018", elaborado por la Dirección General de Asuntos del Personal Académico de la Universidad Nacional Autónoma de México, DGAPA UNAM, página 180 http://dgapa.unam.mx/images/estadistica/anuario_estadisticas_dgapa_2018.pdf

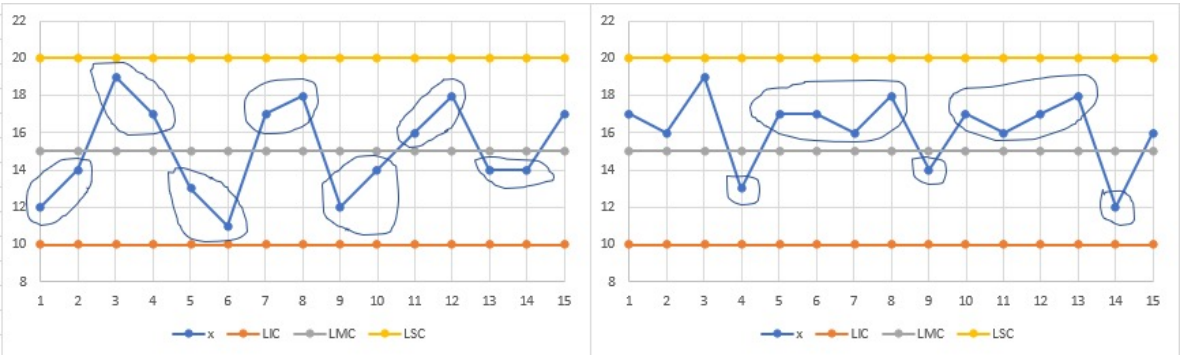
Para el análisis de las series de tiempo se usan métodos cualitativos y/o cuantitativos que ayudan a interpretarlas haciendo posible extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series y que permiten, en diferente medida y con distinta confianza, extrapolar o interpolar los datos y así predecir el comportamiento de una serie en momentos no observados, sean en el futuro (extrapolación pronóstica), en el pasado (extrapolación retrógrada) o en momentos intermedios (interpolación). Las predicciones de los hechos y condiciones futuros de una serie de tiempo se llaman pronósticos, y al acto de hacer tales predicciones se le denomina pronosticar.

Una serie de tiempo presenta diferentes comportamientos como son: a) tendencia; b) ciclo; c) variaciones estacionales; d) fluctuaciones irregulares. Una

tendencia se refiere al movimiento hacia arriba o hacia abajo que caracteriza a una serie de tiempo en un cierto período de tiempo, por ejemplo, en la figura 3.43 claramente se observa que las series de tiempo del personal académico de la UNAM son de crecimiento.

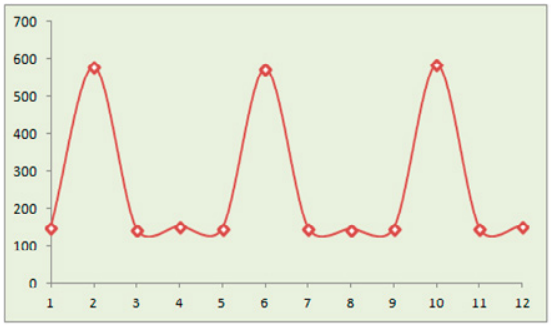
Un ciclo se refiere a los movimientos hacia arriba y/o hacia abajo alrededor de los niveles centrales de tendencia, como se puede apreciar en la figura 3.59.

Figura 3.59. Comportamiento cíclico de dos series de tiempo



Las variaciones estacionales son patrones periódicos en una serie de tiempo que se completa dentro de un año calendario y que se repiten cada año. Las principales fuerzas que causan una variación estacional son las condiciones del tiempo, por ejemplo: en invierno las ventas de helado, en verano la venta de ropa térmica; en mayo y junio la exportación de mango, etcétera. En la figura 3.60 se muestra una variación cíclica estacional a lo largo de 12 meses que dura un año calendario.

Figura 3.60. Variación cíclica estacional



Recuperado de https://image.jimcdn.com/app/cms/image/transf/none/path/s075f076504dfea8d/image/i2451b20c5c752fb2/version/1398957050/image.png

Las fluctuaciones irregulares son movimientos erráticos en una serie de tiempo que siguen un patrón indefinido o irregular.

3.6. Análisis de datos multivariados

En el ejercicio 3.2 se tomó una muestra de 64 alumnas de la Facultad de Ingeniería de la UNAM, a las cuales se les preguntó su peso y su estatura. La muestra se presenta en la figura 3.5 y se reproduce a continuación:

No.	Peso (Kg)	Estatura (m)	No.	Peso (Kg)	Estatura (m)
1	76.0	1.69	33	58.0	1.60
2	56.0	1.51	34	47.0	1.60
3	56.0	1.50	35	54.0	1.57
4	73.0	1.65	36	60.0	1.62
5	60.0	1.51	37	48.0	1.60
6	50.0	1.60	38	52.0	1.51
7	61.0	1.58	39	64.0	1.58
8	61.0	1.56	40	59.0	1.58
9	52.0	1.53	41	60.0	1.67
10	60.0	1.72	42	57.0	1.64
11	70.0	1.54	43	60.0	1.65
12	65.0	1.63	44	65.0	1.57
13	63.0	1.54	45	39.0	1.55
14	57.0	1.70	46	45.0	1.53
15	41.0	1.52	47	56.0	1.63
16	65.0	1.66	48	45.0	1.61
17	44.0	1.50	49	53.0	1.50
18	41.0	1.63	50	48.0	1.52
19	53.0	1.60	51	56.0	1.56
20	70.0	1.59	52	60.0	1.64
21	55.0	1.63	53	60.0	1.50
22	57.0	1.60	54	75.0	1.68
23	52.0	1.58	55	52.0	1.63
24	55.0	1.53	56	54.0	1.52
25	64.0	1.50	57	50.0	1.59
26	58.0	1.65	58	52.0	1.54
27	63.5	1.60	59	50.0	1.56
28	53.0	1.63	60	60.0	1.50
29	55.0	1.53	61	45.0	1.55
30	59.0	1.65	62	55.0	1.60
31	72.0	1.60	63	55.0	1.67
32	58.0	1.60	64	55.0	1.58

3.6.1. Tablas de Contingencia

Los datos anteriores pueden ser agrupados en una tabla de frecuencias bidimensional, para lo cual es necesario calcular los valores máximos, mínimos y rangos de cada una de las variables, de la siguiente forma:

Máximo peso =	76.0	Máxima estatura =	1.72
Mínimo peso =	39.0	Mínima estatura =	1.50
Rango peso =	37.0	Rango estatura =	0.22
m_{Peso} =	10	$m_{Estatura}$ =	10
Δ_{Peso} =	3.7	$\Delta_{Estatura}$ =	0.022

Con estos valores se conforman los intervalos de confianza para cada una de las variables Peso y Estatura, los cuales se muestran en la tabla de la figura 3.61 en color gris.

Figura 3.61. Tabla de frecuencias absolutas conjuntas o tabla de contingencia bidimensional o bivariada

Peso \ Estatura			< 1.5	1.5	1.522	1.544	1.566	1.588	1.61	1.632	1.654	1.676	1.698	Lim Inf Int
			1.5	1.522	1.544	1.566	1.588	1.61	1.632	1.654	1.676	1.698	1.72	Lim Sup Int
			1.489	1.511	1.533	1.555	1.577	1.599	1.621	1.643	1.665	1.687	1.709	Marca_Clase
< 39	39	37.15	0	0	0	0	0	0	0	0	0	0	0	0
39	42.7	40.85	0	1	0	1	0	0	1	0	0	0	0	3
42.7	46.4	44.55	0	1	1	1	0	1	0	0	0	0	0	4
46.4	50.1	48.25	0	1	0	1	0	4	0	0	0	0	0	6
50.1	53.8	51.95	0	2	2	0	1	1	2	0	0	0	0	8
53.8	57.5	55.65	0	3	2	1	2	2	2	1	1	0	1	15
57.5	61.2	59.35	0	3	0	1	2	2	1	4	1	0	1	15
61.2	64.9	63.05	0	1	1	0	1	1	0	0	0	0	0	4
64.9	68.6	66.75	0	0	0	0	1	0	1	0	1	0	0	3
68.6	72.3	70.45	0	0	1	0	0	2	0	0	0	0	0	3
72.3	76	74.15	0	0	0	0	0	0	0	1	0	1	1	3
Lim Inf Int	Lim Sup Int	Marca_Clase	0	12	7	5	7	13	7	6	3	1	3	64

A esta tabla de frecuencias absolutas bidimensional se le denomina tabla de contingencia bidimensional o bivariada.

Los valores que aparecen en las celdas de color blanco son las frecuencias absolutas en cada celda, $f(x_i,y_j)$ o también, f_{ij} que representa el número de veces que un número cae en la celda ubicada en la fila i y en la columna j, lo cual recibe el nombre de Frecuencia Conjunta Absoluta.

Las frecuencias absolutas de color que caen a los costados de la tabla de contingencia reciben el nombre de Frecuencias Absolutas Marginales, las cuales se calculan como:

$$h(y_j)\, f_{\cdot j} = \sum_{i=1}^{i=k} f(x_i, y_j) = \sum_{i=1}^{k} f_{ij} \tag{3.69}$$

$$g(x_i)\, f_{i\cdot} = \sum_{j=1}^{i=m} f(x_i, y_j) = \sum_{j=1}^{m} f_{ij} \tag{3.70}$$

En las anteriores expresiones matemáticas, k representa el número de intervalos de clase de la variable x y m el número de intervalos de clase de la variable y.

Nótese que existen frecuencias absolutas marginales para el peso en color rosa, las cuales se obtienen como la suma de las frecuencias conjuntas que caen en cada renglón, frecuencias absolutas marginales para la estatura en color naranja, las cuales se obtienen como la suma de las frecuencias conjuntas que caen en cada columna. La celda que aparece en color verde, en la esquina inferior derecha de la tabla de contingencia, representa la suma de todas las frecuencias conjuntas absolutas y siempre debe ser igual al total de datos o tamaño de la muestra conjunta, $n = 64$ en este caso.

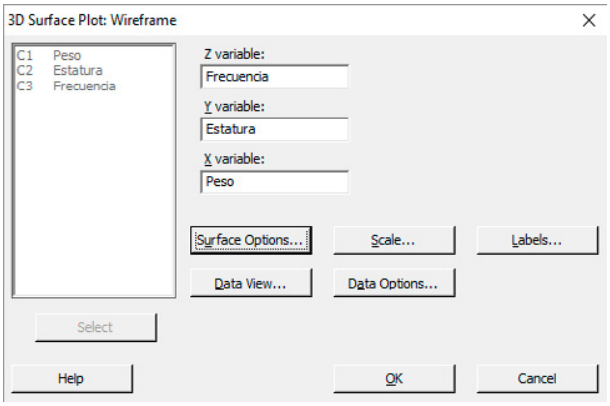
3.6.2. Poliedros de Frecuencias

Si se grafican las frecuencias absolutas conjuntas contra las marcas de clase de los pesos y de las estaturas en un gráfico de tres dimensiones, se obtiene lo que se denomina el poliedro de frecuencias absolutas conjuntas, como el que se muestra en la figura 3.62.

Para graficarlo se utiliza el software Minitab, y se aplican los siguientes pasos:

1. En la primera columna se colocan las marcas de clase del peso, en la segunda columna se colocan las marcas de clase de la estatura, y en la tercera columna se colocan las frecuencias conjuntas absolutas de las celdas que aparecen en blanco de la figura 3.61. Estas tres columnas se copian y pegan en la hoja electrónica de cálculo de Minitab, en particular en las columnas C1, C2 y C3.

2. Para graficar el poliedro de frecuencias absolutas entra al menú Graph, luego al submenú 3D Surface Plot y aparece la siguiente ventana:



En esta ventana se da un click en C3 Frecuencia y se da un click en Select, la cual representará a la variable z; luego un click en χ2 Estatura y después un click en Select, la cual representará al eje y; posteriormente un click en C1 Peso y después un click en Select, la cual representará al eje x; Finalmente un click en Ok, apareciendo la figura 3.62.

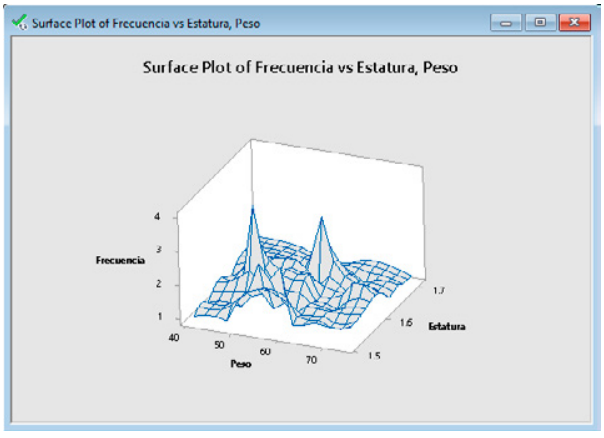


Figura 3.62. Poliedro de frecuencias absolutas conjuntas

Si se divide cada celda de la tabla de frecuencias absolutas conjuntas o tabla de contingencia bidimensional, entre el total de datos de la muestra, se obtiene la tabla de frecuencias relativas conjuntas, como la que se muestra en la figura 3.63:

Figura 3.63. Tabla de frecuencias relativas conjuntas

Peso		Estatura	< 1.5	1.5	1.522	1.544	1.566	1.588	1.61	1.632	1.654	1.676	1.698	Lim Inf Int
			1.5	1.522	1.544	1.566	1.588	1.61	1.632	1.654	1.676	1.698	1.72	Lim Sup Int
			1.489	1.511	1.533	1.555	1.577	1.599	1.621	1.643	1.665	1.687	1.709	Marca_Clase
< 39	39	37.15	0	0	0	0	0	0	0	0	0	0	0	0
39	42.7	40.85	0	0.01563	0	0.01563	0	0	0.01563	0	0	0	0	0.046875
42.7	46.4	44.55	0	0.01563	0.01563	0.01563	0	0.01563	0	0	0	0	0	0.0625
46.4	50.1	48.25	0	0.01563	0	0.01563	0	0.0625	0	0	0	0	0	0.09375
50.1	53.8	51.95	0	0.03125	0.03125	0	0.01563	0.01563	0.03125	0	0	0	0	0.125
53.8	57.5	55.65	0	0.04688	0.03125	0.01563	0.03125	0.03125	0.03125	0.015625	0.015625	0	0.015625	0.234375
57.5	61.2	59.35	0	0.04688	0	0.01563	0.03125	0.03125	0.01563	0.0625	0.015625	0	0.015625	0.234375
61.2	64.9	63.05	0	0.01563	0.01563	0	0.01563	0.01563	0	0	0	0	0	0.0625
64.9	68.6	66.75	0	0	0	0	0.01563	0	0.01563	0	0.015625	0	0	0.046875
68.6	72.3	70.45	0	0	0.01563	0	0	0.03125	0	0	0	0	0	0.046875
72.3	76	74.15	0	0	0	0	0	0	0	0.015625	0	0.01563	0.015625	0.046875
Lim Inf Int	Lim Sup Int	Marca_Clase	0	0.1875	0.10938	0.07813	0.10938	0.20313	0.10938	0.09375	0.046875	0.01563	0.046875	1

Si se grafican las frecuencias conjuntas relativas contra las marcas de clase de los pesos y de las estaturas, en un gráfico de tres dimensiones, se obtiene lo que se denomina el poliedro de frecuencias relativas que resulta ser idéntico al que se muestra en la figura 3.62, pero en otra escala en el eje vertical de frecuencias como se muestra en la figura 3.64.

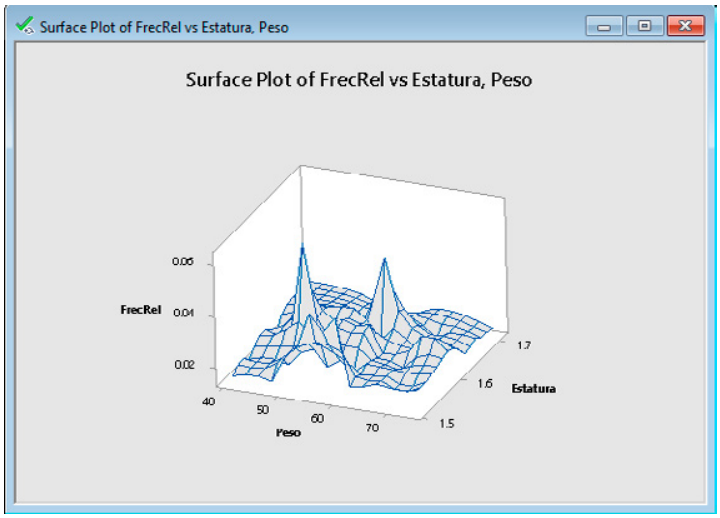


Figura 3.64. Poliedro de frecuencias relativas conjuntas

Se definen las frecuencias condicionales de la siguiente forma:

$$f(x_i \mid y_j) = \frac{f(x_i, y_j)}{h(y_j)} = \frac{f_{ij}}{f_{\cdot j}} \tag{3.71}$$

$$f(y_j \mid x_i) = \frac{f(x_i, y_i)}{g(x_i)} = \frac{f_{ij}}{f_{i\cdot}} \tag{3.72}$$

En donde $g(x_i)$ y $h(y_j)$ son las frecuencias marginales de x y de y respectivamente.

Para ejemplificar estos conceptos se obtendrán las siguientes frecuencias condicionales:

a. $f(x = 55.65 \mid y = 1.599) = f(x = 55.65, y = 1.599) / h(y = 1.599) = 0.03125 / 0.20313 = 0.153842$

b. $f(y = 1.599 \mid x = 55.65) = f(x = 55.65, y = 1.599) / / g(x = 55.65) = 0.03125 / 0.234375 = 0.133333$

c. $f(x \mid 1.588 < y < 1.61)$

x	$f(x \mid 1.588 < y < 1.61)$
< 39	0
39 - 42.7	0
42.7 - 46.4	0.0769
46.4 - 50.1	0.3077
50.1 - 53.8	0.0769
53.8 - 57.5	0.1538
57.5 - 61.2	0.1538
61.2 - 64.9	0.0769
64.9 - 68.6	0
68.6 - 72.3	0.1538
72.3 - 76	0

d. $f(y \mid 53.8 < x < 57.5)$

y	$f(y \mid 53.8 < x < 57.5)$
< 1.5	0
1.5 - 1.522	0.2000
1.522 - 1.544	0.1333
1.544 - 1.566	0.0667
1.566 - 1.588	0.1333
1.588 - 1.610	0.1333
1.610 - 1.632	0.1333
1.632 - 1.654	0.0667
1.654 - 1.676	0.0667
1.676 - 1.698	0
1.698 - 1.720	0.0667

3.6.3. Independencia estadística de frecuencias conjuntas

Se dice que dos variables *x* y y son estadísticamente independientes si se cumple que:

$f(x|y) = g(x) = f_{i.}$ o también $f(y|x) = h(y) = f_{.j}$

Esto implica que $f(x,y) = g(x) * h(y)$ para todo x y y que pertenece al rango de definición de cada una de las variables, es decir, esta condición debe cumplirse para todas las celdas que conforman a la tabla de contingencia; con que exista una celda donde no se cumpla, entonces no son estadísticamente independientes.

Para el ejemplo de pesos y estaturas nótese que:

f(peso = 55.65, estatura = 1.599) = 0.03125

En cambio

g(peso = 55.65)h(estatura = 1.599) = (0.234375)(0.20313) = 0.047608

Por lo cual se puede concluir que en este caso el peso y la estatura no son estadísticamente independientes.

3.6.4. Covarianza

En probabilidad, la covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias poblacionales. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación o la regresión.

En probabilidad, la covarianza entre dos variables aleatorias x y y se define como:

$$cov(x, y) = \sigma_{xy}^2 = E\{(x - \mu_x)(y - \mu_y)\} = E\{xy\} - E\{x\}E\{y\} \tag{3.73}$$

La covarianza cumple ciertas propiedades mismas que se deducen directamente de las propiedades de la esperanza matemática:

i. cov(x,a) = 0
ii. cov(x,x) = var(x)
iii. cov(x,y) = cov(y,x)
iv. cov(ax,by) = abcov(x,y)
v. cov(x + a,y + b) = cov(x,y)

Una covarianza positiva entre dos variables *x* y *y* significa que cuando una variable crece la otra también lo hace, y cuando una variable decrece a la otra le sucede lo mismo. Por el contrario, una covarianza negativa significa que al crecer una de las variables la otra decrece, expresando un comportamiento recíproco entre ellas. El signo de la covarianza, por lo tanto, expresa la tendencia en la relación lineal entre las variables.

En estadística, el estadístico muestral de la covarianza, el cual se denota por *Sxy*, para datos no agrupados, se define como:

$$Sxy = \frac{1}{n} \sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y}) \tag{3.74}$$

En donde x_i representa las lecturas obtenidas de la primera variable y y_i las lecturas obtenidas de la segunda variable.

Para datos agrupados en una tabla de contingencia el estadístico *Sxy* se define como:

$$Sxy = \frac{1}{n} \sum_{i=1}^{i=k} \sum_{j=1}^{j=m} f(t_i, s_j)(t_i - \bar{x})(s_j - \bar{y}) \tag{3.75}$$

$$Sxy = \frac{1}{n} \sum_{i=1}^{i=k} \sum_{j=1}^{j=m} f_{ij}(t_i - \bar{x})(s_j - \bar{y}) \tag{3.76}$$

En donde t_i es la marca de clase de la primera variable x_i y s_j, la marca de clase de la segunda variable y_j; $f(t_i,s_j)$ es la frecuencia conjunta de la celda (i,j).

Se calculará la covariancia para datos sin agrupar de la tabla de la figura 3.5, correspondiente a los datos del ejercicio 3.2.

$$Sxy = [(76–56.6)*(1.69–1.6) + (56–56.6)*(1.51–1.6) + \cdots + (64–56.6)*(1.58–1.6)]/64 = 0.1447$$

Se obtendrá la covariancia para datos agrupados de la tabla de contingencia de la figura 3.47:

$$Sxy = [(1)*(40.85–56.6)*(1.511–1.6) + (1)*(40.85–56.6)*(1.555–1.6) + \cdots + (1)*(74.15\text{-}56.6)*(1.709–1.6)]/64 = 0.15032$$

La diferencia entre ellas se debe a que los valores de la muestra en el caso de la tabla de contingencia, son reemplazados por las marcas de clase, por lo que los resultados para datos agrupados son aproximados a los resultados de los datos sin agrupar.

i. Si $Sxy > 0$ hay dependencia directa (positiva), es decir, a grandes valores de x corresponden grandes valores de y.
ii. Si $Sxy = 0$ se interpreta como la no existencia de una relación lineal entre las dos variables estudiadas.
iii. Si $Sxy < 0$ hay dependencia recíproca o negativa, es decir, a grandes valores de x corresponden pequeños valores de y.

3.6.5. Coeficiente de correlación

En probabilidad se define el coeficiente de correlación como:

$$\rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} \tag{3.77}$$

El estimador muestral del coeficiente de correlación de dos variables se define de la siguiente forma:

$$r_{xy} = \frac{Sxy}{Sx\ Sy} \tag{3.78}$$

El coeficiente de correlación entre el peso y la estatura en el ejercicio 3.2 es:

$r = 0.1447 / (8.01152 * 0.057283) = 0.3153$

El valor del índice de correlación r varía en el intervalo $[-1,1]$, indicando el signo el sentido de la relación:

i. Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
ii. Si $0 < r < 1$, existe una correlación positiva.

iii. Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.
iv. Si $-1 < r < 0$, existe una correlación negativa.
v. Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación recíproca: cuando una de ellas aumenta, la otra disminuye en proporción constante.

3.6.6. Coeficiente de contingencia cuadrática

El coeficiente de contingencia C de Pearson expresa la intensidad de la relación entre dos (o más) variables cualitativas. Se basa en la comparación de las frecuencias efectivamente calculadas de dos características, con las frecuencias que se hubiesen esperado con independencia de estas características.

Para obtenerlo, se debe calcular primero el Coeficiente de Contingencia Cuadrática χ^2, el cual se define a partir de una tabla de contingencia, de la siguiente forma:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(f_{i,j} - \frac{f_{i\cdot} f_{\cdot j}}{n} \right)^2}{\frac{f_{i\cdot} f_{\cdot j}}{n}} \tag{3.79}$$

Para el ejercicio 3.2, utilizando la tabla de contingencia de la figura 3.61

$$\chi^2 = \frac{\left[1 - \frac{3*12}{64} \right]^2}{\frac{3*12}{64}} + \frac{\left[0 - \frac{3*7}{64} \right]^2}{\frac{3*7}{64}} + \cdots + \frac{\left[1 - \frac{3*3}{64} \right]^2}{\frac{3*3}{64}} = 83.6857$$

El coeficiente de contingencia C de Pearson se define como:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{3.80}$$

Para ilustrar su aplicación se calculará para el ejercicio 3.2, usando la tabla de la figura 3.61:

$$C = \sqrt{\frac{83.6857}{83.6857 + 64}} = 0.75276$$

3.6.7. Coeficiente φ de correlación de Mathews

En estadística, el coeficiente φ (phi) o r_φ, también llamado coeficiente de correlación de Mathews, es una medida de la asociación entre dos variables binarias. Esta medida es similar al coeficiente de correlación de Pearson en su interpretación. El coeficiente phi se define como:

$$\varphi^2 = \frac{\chi^2}{n} \tag{3.81}$$

3.6.8. Tablas de contingencia multivariadas

Para ejemplificar el concepto de tablas de contingencia multivariadas se ilustrará con el siguiente ejemplo con tres variables:

Ejercicio 3.6

Se tomó una muestra de 120 estudiantes de la Facultad de Ingeniería de la UNAM, a los cuales se les preguntó su peso y su estatura. La muestra se ilustra en la figura 3.65:

Figura 3.65. Muestra de 120 estudiantes de la Facultad de Ingeniería de la UNAM

Muestra	Sexo	Peso (kg)	Estatura (m)	Muestra	Sexo	Peso (kg)	Estatura (m)	Muestra	Sexo	Peso (kg)	Estatura (m)	Muestra	Sexo	Peso (kg)	Estatura (m)
1	F	76	1.69	31	M	65	1.72	61	M	62	1.70	91	F	60	1.50
2	F	56	1.51	32	M	74	1.64	62	F	60	1.62	92	F	75	1.68
3	F	56	1.50	33	M	60	1.73	63	M	65	1.70	93	F	52	1.63
4	M	62	1.69	34	F	64	1.50	64	M	65	1.65	94	M	52	1.70
5	F	73	1.65	35	M	56	1.67	65	F	48	1.60	95	M	68	1.68
6	F	60	1.51	36	M	57	1.75	66	F	52	1.51	96	F	54	1.52
7	F	50	1.60	37	M	91	1.85	67	F	64	1.58	97	M	80	1.72
8	M	84	1.82	38	M	76	1.74	68	F	59	1.58	98	M	76	1.65
9	M	69	1.54	39	F	58	1.65	69	F	60	1.67	99	M	72	1.80
10	M	68	1.74	40	M	74	1.76	70	F	57	1.64	100	F	50	1.59
11	F	61	1.58	41	M	68	1.85	71	M	60	1.66	101	M	85	1.75
12	F	61	1.56	42	F	63	1.60	72	M	62	1.65	102	M	77	1.72
13	F	52	1.53	43	M	66	1.70	73	M	64	1.68	103	M	64	1.62
14	F	60	1.72	44	M	80	1.80	74	M	63	1.74	104	M	51	1.66
15	M	79	1.79	45	M	64	1.68	75	M	73	1.76	105	M	56	1.70
16	F	70	1.54	46	M	59	1.72	76	M	84	1.70	106	F	52	1.54
17	F	65	1.63	47	M	73	1.73	77	M	85	1.75	107	F	50	1.56
18	F	63	1.54	48	M	82	1.68	78	M	79	1.62	108	M	65	1.76
19	F	57	1.70	49	F	53	1.63	79	M	65	1.73	109	M	65	1.75
20	F	41	1.52	50	M	79	1.88	80	F	60	1.65	110	M	65	1.74
21	M	72	1.69	51	F	55	1.53	81	M	52	1.56	111	M	72	1.64
22	M	78	1.78	52	M	74	1.68	82	F	65	1.57	112	F	60	1.50
23	F	44	1.50	53	F	59	1.65	83	F	39	1.55	113	M	68	1.68
24	F	53	1.60	54	F	72	1.60	84	F	45	1.53	114	F	45	1.55
25	F	70	1.59	55	F	65	1.66	85	F	56	1.63	115	F	55	1.60
26	F	55	1.63	56	F	58	1.60	86	F	58	1.60	116	F	55	1.67
27	F	57	1.60	57	M	67	1.67	87	F	53	1.50	117	F	45	1.61
28	F	52	1.58	58	F	47	1.60	88	F	48	1.52	118	F	55	1.58
29	F	55	1.53	59	F	54	1.57	89	F	56	1.56	119	M	75	1.80
30	M	63	1.69	60	M	89	1.79	90	F	60	1.64	120	M	98	1.75

Determine una tabla de contingencia con estos datos

Primero se obtienen los valores máximos, mínimos, rango, número de intervalos y amplitud de intervalo de clase de cada una de las variables que intervienen en el problema.

Se pondrá como x la variable peso, como y la variable estatura y como z el género de cada estudiante.

Parámetro	Peso	Estatura
Val Max =	98	1.88
Val Min =	39	1.5
Rango =	59	0.38
m =	10	10
D =	6	0.04

La tabla de contingencia multivariable se muestra a continuación:

Figura 3.66. Tabla de Contingencia de la muestra de 120 estudiantes de la Facultad de Ingeniería de la UNAM

		Estatura																					
		1.50-1.54		1.54-1.58		1.58-1.62		1.62-1.66		1.66-1.70		1.70-1.74		1.74-1.78		1.78-1.82		1.82-1.86		1.86-1.90			
	s_j	1.52		1.56		1.6		1.64		1.68		1.72		1.76		1.8		1.84		1.88			
		Género		Género		Género		Género		Género		Género		Género		Género		Género		Género			
Peso	t_i	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	fi.k		fi..
39-45	43	3		2		1																6	0	6
45-51	48	1		1		4			1													6	1	7
51-57	54	9		5	1	4		6		2	3				1							26	5	31
57-62	60	3		3		3		3	2	1	1	1	2									14	5	19
62-68	65	1	1	2		1	1	2	2		9		4		2							6	19	25
68-74	71		1	1		2		1	2		2		1		2		1		1			4	10	14
74-80	77						1		1	1			3		1		3				1	1	10	11
80-86	83										2				1		1					0	4	4
86-92	89																1		1			0	2	2
92-98	95														1							0	1	1
f.jk=		17	2	14	1	15	2	12	8	4	17	1	10	0	8	0	6	0	2	0	1			
f.j.=		19		15		17		20		21		11		8		6		2		1				
f..k=		63	57																					

Observe en esta figura que existen diversas frecuencias absolutas marginales de la muestra. La definición de cada una de ellas se expresa a continuación:

$$f_{ij\cdot} = \sum_{k=1}^{k=m_3} f_{ijk} \quad f_{i\cdot k} = \sum_{j=1}^{j=m_2} f_{ijk} \quad f_{\cdot jk} = \sum_{i=1}^{i=m_1} f_{ijk} \tag{3.82}$$

$$f_{i\cdot\cdot} = \sum_{j=1}^{k=m_2} \sum_{k=1}^{k=m_3} f_{ijk} \qquad f_{\cdot j\cdot} = \sum_{i=1}^{i=m_1} \sum_{k=1}^{k=m_3} f_{ijk} \qquad f_{\cdot\cdot k} = \sum_{i=1}^{i=m_1} \sum_{j=1}^{j=m_2} f_{ijk} \tag{3.83}$$

Como se puede apreciar en la figura 3.66:

Peso	39-45	45-51	51-57	57-62	62-68	68-74	74-80	80-86	86-92	92-98
fi..	6	7	31	19	25	14	11	4	2	1

Estatura	1.50-1.54	1.54-1.58	1.58-1.62	1.62-1.66	1.66-1.70	1.70-1.74	1.74-1.78	1.78-1.82	1.82-1.86	1.86-1.90
f.j.	19	15	17	20	21	11	8	6	2	1

Género	F	M
f..k	63	57

f.jk	Estatura	1.50-1.54	1.54-1.58	1.58-1.62	1.62-1.66	1.66-1.70	1.70-1.74	1.74-1.78	1.78-1.82	1.82-1.86	1.86-1.90
Género	F	17	14	15	12	4	1	0	0	0	0
	M	2	1	2	8	17	10	8	6	2	1

fi.k	Peso	39-45	45-51	51-57	57-62	62-68	68-74	74-80	80-86	86-92	92-98
Género	F	6	6	26	14	6	4	1	0	0	0
	M	0	1	5	5	19	10	10	4	2	1

Ejercicio 3.7

(Ejercicio 5-17, página 215 del libro de Douglas C Montgomery, *Diseño y Análisis de Experimentos,* Editorial Limusa Wiley, segunda edición, 2007). El Departamento de Control de Calidad de una planta de acabados textiles estudia el efecto de varios factores sobre el teñido de una tela de algodón y fibras sintéticas, utilizada para fabricar camisas para caballero. Se seleccionaron tres operadores, tres duraciones de ciclo y dos temperaturas, y se tiñeron tres ejemplares pequeños de la tela bajo cada conjunto de mediciones. La tela terminada se comparó con un patrón y se le asignó una evaluación numérica. Los datos obtenidos se presentan a continuación:

	Temperatura					
	300			350		
	Operador			Operador		
Tiempo de Ciclo	1	2	3	1	2	3
40	23	27	31	24	38	34
	24	28	32	23	36	36
	25	26	29	28	35	39
50	36	34	33	37	34	34
	35	38	34	39	38	36
	36	39	35	35	36	31
60	28	35	26	26	36	28
	24	35	27	29	37	26
	27	34	25	25	34	24

Esta tabla contiene cuatro variables: la primera que se situaría sobre el eje de las *x* sería el tiempo de ciclo, la segunda que se situaría sobre el eje de las y sería la temperatura, la tercera que se situaría sobre el eje de las *z* sería el operador, y la variable de respuesta en este caso, llamémosle u, que son los valores dentro de cada celda sería el teñido de la tela.

La tabla anterior también se puede presentar como una tabla de cuaternas ordenadas como la que se muestra a continuación:

x	y	z	u	x	y	z	u	x	y	z	u
40	300	1	23	50	300	1	36	60	300	1	28
40	300	1	24	50	300	1	35	60	300	1	24
40	300	1	25	50	300	1	36	60	300	1	27
40	300	2	27	50	300	2	34	60	300	2	35
40	300	2	28	50	300	2	38	60	300	2	35
40	300	2	26	50	300	2	39	60	300	2	34
40	300	3	31	50	300	3	33	60	300	3	26
40	300	3	32	50	300	3	34	60	300	3	27
40	300	3	29	50	300	3	35	60	300	3	25
40	350	1	24	50	350	1	37	60	350	1	26
40	350	1	23	50	350	1	39	60	350	1	29
40	350	1	28	50	350	1	35	60	350	1	25
40	350	2	38	50	350	2	34	60	350	2	36
40	350	2	36	50	350	2	38	60	350	2	37
40	350	2	35	50	350	2	36	60	350	2	34
40	350	3	34	50	350	3	34	60	350	3	28
40	350	3	36	50	350	3	36	60	350	3	26
40	350	3	39	50	350	3	31	60	350	3	24

Para llevar a cabo la tabla de contingencia se tendrían que definir los valores máximos, mínimos, marca de clase, número de intervalos y amplitud de cada intervalo de cada una de las variables que intervienen en el problema, como se muestra a continuación:

Valor máximo de $u = 39$
Valor mínimo de $u = 23$
Rango = 39-23 = 16
$m = 8$
$\Delta = 2$

La tabla de contingencia para este problema sería la siguiente:

Temperatura	Operador	Tiempo de Ciclo	Teñido										
			23-25	25-27	27-29	29-31	31-33	33-35	35-37	37-39			
			24	26	28	30	32	34	36	38	f..k.	f.j..	fi…
300	1	40	3								21	17	28
		50						1	2		14	22	26
		60	1	1	1						19	15	
	2	40		2	1								
		50						1		2			
		60		1				3					
	3	40			1	1	1						
		50						2					
		60	1	1	1		1						
350	1	40	2		1								
		50						1	1	1			
		60	1		1								
	2	40						2	2	2			
		50						1	1	1			
		60						1	2				
	3	40						1	1	1			
		50											
		60	1	1	1								
		f…1	9	6	7	1	2	13	9	7	54		

Ejercicios del Capítulo 3

1. En la siguiente tabla se muestran el número de alumnos inscritos, el número de horas-semana-mes pagadas a los profesores de carrera y el número de horas-semana-mes pagadas a los profesores de asignatura de la UNAM, en el período 1999-2018.

No.	Periodo	No. Alumnos	Horas Prof Carr	Horas Prof Asig
1	1999-2000	255,226	195,800	251,215
2	2000-2001	245,317	195,540	236,389
3	2001-2002	251,149	197,740	244,960
4	2002-2003	259,036	206,260	244,572
5	2003-2004	269,143	207,040	248,529
6	2004-2005	279,054	208,440	250,810
7	2005-2006	286,484	210,420	253,534
8	2006-2007	292,889	212,020	257,705
9	2007-2008	299,688	213,020	262,283
10	2008-2009	305,969	212,220	265,940
11	2009-2010	314,557	213,000	266,912
12	2010-2011	316,589	214,040	275,971
13	2011-2012	324,413	213,440	283,829
14	2012-2013	330,382	215,340	291,163
15	2013-2014	337,763	210,980	297,097
16	2014-2015	342,542	209,520	301,935
17	2015-2016	346,730	217,280	304,932
18	2016-2017	349,539	218,080	309,703
19	2017-2018	349,515		
20	2018-2019	356,530		

   a. Para cada una de las últimas tres columnas elabore una estadística descriptiva completa sobre sus medidas de tendencia central, de dispersión, de forma y de otro tipo.

b. Para cada una de las últimas tres columnas elabore una tabla de frecuencias, trace sus histogramas, sus polígonos de frecuencias y sus ojivas.
c. Para cada una de las últimas tres columnas obtenga sus intervalos intercuartiles, interdeciles e interpercentiles.
d. Calcule las horas-semana-mes de docencia considerando que en promedio un profesor de carrera debe dar un 25% de su tiempo en horas de docencia y el profesor de asignatura debe cubrir el 100% de su tiempo contratado en horas frente a grupo, es decir horas de docencia = 0.25*Horas Prof Carr + Horas Prof Asig y agregue esta variable como una columna a la izquierda de la tabla anterior. Para esta columna obtenga una estadística descriptiva completa sobre sus medidas de tendencia central, de dispersión, de forma y de otro tipo; elabore una tabla de frecuencias, trace sus histogramas, sus polígonos de frecuencias y sus ojivas; obtenga sus intervalos intercuartiles, interdeciles e interpercentiles.
e. Trace los diagramas de tallo y hojas, de caja y bigote y de Pareto para cada una de las variables número de alumnos, número de horas de profesor de carrera, número de horas de profesor de asignatura y número de horas de docencia.
f. Trace la serie de tiempo de cada una de las variables número de alumnos, número de horas de profesor de carrera, número de horas de profesor de asignatura y número de horas de docencia.
g. Elabore una tabla de contingencia bivariada entre el número de alumnos y el número de horas de docencia. Calcule las frecuencias marginales y las frecuencias condicionales.
h. Trace el poliedro de frecuencias conjuntas entre el número de alumnos y el número de horas de docencia.
i. Obtenga la covarianza entre el número de alumnos y el número de horas de docencia; calcule los coeficientes de correlación de Pearson y el coeficiente phi.

2. El hilo sintético monofilamento obtenido de la extrusión de pellets de poliamida, debido a su resistencia aún con calibres muy delgados, es utilizado en microcirugía y oftalmología. El hilo es pigmentado de negro o azul con colorantes inocuos. Una característica de calidad importante de este hilo es su nivel de encogimiento. Para probar su nivel de encogimiento se corta un metro de este hilo en el carrete, se sumerge en agua caliente a cierta temperatura durante cierto tiempo y se saca el hilo volviendo a medirlo. El encogimiento que sufre, según la norma internacional correspondiente, no debe

ser mayor del 2%. Suponga que se realizó un experimento con una muestra de trozos de un metro de $n = 30$ carretes elegidos aleatoriamente de un lote de $N = 300$. Los resultados obtenidos se muestran en la siguiente tabla.

98.97	99.35	97.40
99.00	98.08	98.89
99.39	98.02	99.31
98.12	99.28	99.80
97.28	97.92	98.93
97.01	98.96	97.81
99.21	98.44	98.13
98.35	99.19	99.85
99.93	99.97	98.19
99.77	97.20	98.37

a. Elabore una estadística descriptiva completa sobre las medidas de tendencia central, de dispersión, de forma y de otro tipo del nivel de encogimiento del monofilamento.
b. Obtenga una tabla de frecuencias del nivel de encogimiento del monofilamento, trace sus histogramas, sus polígonos de frecuencias y sus ojivas.
c. Determine los intervalos intercuartiles, interdeciles e interpercentiles del nivel de encogimiento del monofilamento.
d. Trace los diagramas de tallo y hojas, de caja y bigote y de Pareto del nivel de encogimiento del monofilamento.
e. Estime la fracción de carretes de monofilamento que no cumple la especificación.
f. Calcule la fracción de carretes de monofilamento que tiene un nivel de encogimiento de entre uno y tres centímetros.

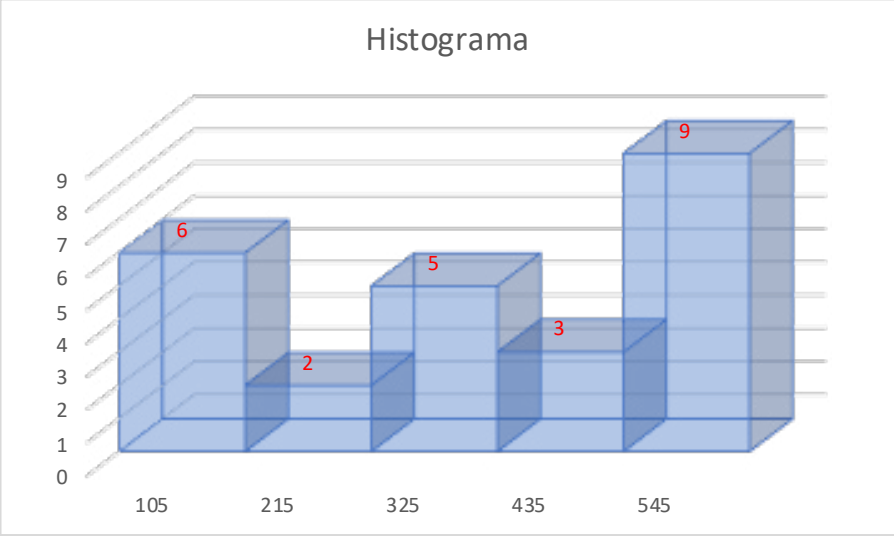
3. Una compañía geohidrológica realizó 50 mediciones del flujo en metros cúbicos por segundo del río Papaloapan, a pedido de una oficina del gobierno federal, pero como no le han pagado entregó como reporte la siguiente tabla de frecuencias.

a. Usted como experto en Estadística complete la tabla de frecuencias.
b. Trace su histograma, su polígono de frecuencias absolutas, su polígono de frecuencias relativas, su polígono de frecuencias absolutas acumuladas y su polígono de frecuencias relativas acumuladas.

c. Calcule sus medidas de tendencia central, de dispersión, de forma, y de otro tipo con base en los resultados de la tabla de frecuencias.
d. Estime los cuartiles y deciles y obtenga sus intervalos intercuartil, interdecil e interpercentil.
e. Elabore un diagrama de caja y bigotes y un diagrama de Pareto.

No.	Intervalo de clase						
	Lim Inf	Lim Sup	Marca de Clase	Frec Abs	Frec Rel	Frec Abs Acum	Frec Rel Acum
1							0.1
2	722		759.5	12			0.34
3		872			0.28	31	
4					0.2		
5				5			
6		1097	1059.5				0.98
7				1			
			Suma =	50	1		

4. Un profesor de laboratorio de electrónica le pidió de mal modo al taller de manufactura avanzada la medición de la resistencia a la tensión de una muestra de tamaño n = 25 de cables. El Jefe del Taller no pudo negarse y le mandó como resultado el siguiente histograma.



El profesor del laboratorio de electrónica le solicitó a Usted, por ser su alumno, le brindara más información al respecto, por lo cual le pidió lo siguiente:

a. Con base en este histograma complete la siguiente tabla de frecuencias.

No.	Intervalo de clase						
	Lim Inf	Lim Sup	Marca de Clase	Frec Abs	Frec Rel	Frec Abs Acum	Frec Rel Acum
1							
2							
3	270						
4							
5							
Suma =							

b. Trace su histograma, su polígono de frecuencias absolutas, su polígono de frecuencias relativas, su polígono de frecuencias absolutas acumuladas y su polígono de frecuencias relativas acumuladas.
c. Calcule sus medidas de tendencia central, de dispersión, de forma, y de otro tipo con base en los resultados de la tabla de frecuencias.
d. Estime los cuartiles y deciles y obtenga sus intervalos intercuartil, interdecil e interpercentil.
e. Elabore un diagrama de caja y bigotes y un diagrama de Pareto.

5. (Problema 5-22 de la página 216 del libro de Douglas C Montgomery, *Diseño y Análisis de Experimentos,* Editorial Limusa Wiley, segunda edición, 2007). En un artículo del Journal of Testing and Evaluation (vol. 16, no. 2, pp 508-515) se investigaron los efectos de la frecuencia de carga cíclica y de las condiciones ambientales sobre el crecimiento de las fisuras por fatiga con un esfuerzo constante de 22 MPa para un material particular. Los datos del experimento se presentan en la siguiente tabla; la respuesta es el índice de crecimiento de las fisuras por fatiga.

a. Elabore una tabla de contingencia multivariable, considerando las variables x = frecuencia, y = medio ambiente y z = índice de crecimiento de las fisuras por fatiga.

b. Calcule las frecuencias marginales de cada variable.
c. Obtenga las frecuencias condicionales $f(x|y), f(x|z), f(y|z), f(x,y|z)$.
d. Determine las covarianzas cov(x,y), cov(x,z), cov(y,z) y sus correspondientes coeficientes de correlación.

Frecuencia	Medio ambiente		
	Aire	H_2O	H_2O salada
10	2.29	2.06	1.90
	2.47	2.05	1.93
	2.48	2.23	1.75
	2.12	2.03	2.06
1	2.65	3.20	3.10
	2.68	3.18	3.24
	2.06	3.96	3.98
	2.38	3.64	3.24
0.1	2.24	11.00	9.96
	2.71	11.00	10.01
	2.81	9.06	9.36
	2.08	11.30	10.40

4. Estimación de parámetros poblacionales

4.1. Conceptos básicos de inferencia estadística

Tal como ya se mencionó antes, la Inferencia Estadística es la rama de la Estadística que proporciona las reglas para estimar ciertos valores de una población, con base en los resultados de una muestra, así como formular y probar hipótesis sobre la verdad de estas estimaciones y tomar decisiones con base en estos resultados.

Un parámetro poblacional es el valor verdadero de la medida de tendencia central, de la dispersión o de la forma de alguna característica de la población y que, cuando es desconocido, puede estimarse mediante un estadístico asociado a una muestra.

La primera finalidad del muestreo es obtener muestras representativas de la población bajo estudio. Una muestra es representativa de la población si es obtenida aleatoriamente.

Tal como ya se había mencionado previamente, se dice que el muestreo es aleatorio si cumple las siguientes características:

» Todos los posibles resultados del experimento deben tener la misma posibilidad de ocurrir.
» Los resultados deben ser independientes entre sí.

¿Cómo se puede determinar cuál procedimiento usar y el número de elementos a elegir de la muestra? La respuesta depende de dos factores: ¿qué tanta representatividad se desea? y ¿qué tan seguro se requiere estar de esta representatividad?, es decir,

1. Si u es la variable de interés y $\hat{u}$ es un estimador puntual de u entonces se debe especificar un límite superior para el error de estimación, esto es

$$|u - \hat{u}| < \varepsilon \tag{4.1}$$

La variable de interés u a estudiar puede ser la media de la población μ, el total poblacional τ, la fracción de interés p o cualquier otro parámetro poblacional de interés. La variable $\hat{u}$ representa un estimador puntual del parámetro poblacional u. Para el caso de la media poblacional, un estimador puntual podría ser la media muestral, la mediana me, o la moda m_o, según cuál sea más representativa de la tendencia central de los datos.

Nótese que la desigualdad, 4.1 también puede ser escrita como:

$$u - \varepsilon < \hat{u} < u + \varepsilon \tag{4.2}$$

2. Se debe fijar la probabilidad de que efectivamente el error de estimación sea menor de ε, esto es, la fracción de las veces en que el muestreo tiene como error de estimación un valor menor a ε,

$$p\left[\text{Error de Estimación} < \varepsilon\right] = 1 - \alpha \tag{4.3}$$

Nótese que la probabilidad $p\left[\text{Error de Estimación} < \varepsilon\right] = 1 - \alpha$ también puede ser escrita como:

$$p\left[\,|u - \hat{u}| < \varepsilon\right] = 1 - \alpha$$

$$p\left(u - \varepsilon < \hat{u} < u + \varepsilon\right) = 1 - \alpha \tag{4.4}$$

Generalmente el criterio que se adopta para fijar un valor al límite superior del error de estimación es definirlo como un múltiplo de la desviación estándar del estimador $\hat{u}$, es decir,

$$\varepsilon = k\sigma_{\hat{u}} \tag{4.5}$$

En donde k depende de la función de probabilidad que tenga el estimador û y del nivel de confianza (1–α) que se desee tener. Si la función de probabilidad del estimador de la variable de interés û es normal, en este caso $k = z_{\alpha/2}$, en donde z representa a una variable aleatoria normal estándar.

La estimación de parámetros poblacionales a partir de muestras se lleva a efecto a través de estimadores puntuales, de los cuales se debe determinar qué cualidades deben reunir para ser válidos y representativos, o a través de estimadores

por intervalos de confianza, de los cuales se debe determinar la distribución de probabilidad que presentan y el nivel de confianza que se desea tener. Las hipótesis estadísticas que se formulen deben ser probadas para comprobar su validez y representatividad. Tanto la estimación por intervalos de confianza como las pruebas de hipótesis se pueden clasificar para una población o para dos poblaciones o más.

Para una población, generalmente los parámetros que se estiman son la media, varianza, desviación estándar, fracción de éxitos, fracción de defectuosos, fracción de defectos, tamaño poblacional, etcétera. Para dos poblaciones generalmente lo que se estima es el cociente entre varianzas (para determinar la igualdad de estas varianzas), la diferencia de medias (para determinar la igualdad de estas medias), la diferencia de proporciones, la covarianza y el coeficiente de correlación. También dentro de la Estadística Inferencial se ven otro tipo de pruebas estadísticas como lo son las pruebas de bondad de ajuste, que permiten probar la adecuación de un conjunto de datos obtenidos empíricamente a un modelo probabilístico específico.

Estadístico muestral

Un estadístico muestral es una medida cuantitativa, obtenida a partir de un conjunto de datos, observaciones o mediciones de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico. Más formalmente, un estadístico muestral es una función medible $f: R^n \to R$ en la que, para una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$, le corresponde un número real $f(x_1, x_2, x_3, \cdots, x_n)$, que permite estimar un determinado parámetro de la población de la que procede la muestra.

Para ilustrar lo anterior, suponga que se requiere escoger un estadístico $\hat{\theta}$ del parámetro poblacional θ. Suponga a su vez, que existen varios estadísticos que pudieran usarse, ¿qué criterios existen para seleccionar al estadístico más adecuado?

4.2. Criterios para seleccionar estimadores puntuales

Las cualidades que deben reunir los estimadores puntuales se pueden resumir en las siguientes:

a. Insesgabilidad o imparcialidad;
b. Eficiencia o precisión;
c. Consistencia;
d. Robustez;
e. Suficiencia; e,
f. Invariancia.

4.2.1. Insesgabilidad o imparcialidad

Se dice que un estimador $\hat{\theta}$ es insesgado si la esperanza matemática del estimador es igual al parámetro poblacional θ, es decir,

$$\mu_{\hat{\theta}} = E\{\hat{\theta}\} = \theta \tag{4.6}$$

Suponga que se tiene una muestra aleatoria ordenada de menor a mayor $\{x_1, x_2, x_3, \cdots, x_n\}$, obtenida de una población cuya variable aleatoria es x. Suponga que cada una de las variables x_i son estadísticamente independientes entre sí y que todas ellas pertenecen a la misma población, lo que implica que:

$$E\{x_i\} = \mu_x$$

$$var\{x_i\} = \sigma_x^2$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2$$

Se requieren definir estimadores de la media poblacional μ_x y de la desviación estándar poblacional σ_x.

Es importante no confundir la media y la varianza muestral de cada observación x_i, con la media y la varianza poblacional. La media y la varianza poblacionales son constantes, en cambio, las medias y las varianzas para cada x_i son variables aleatorias.

Ejercicio 4.1 Tres estimadores de la media poblacional podrían ser:

a. SemiRango: $SR = (x_1 + x_n)/2$

Observe que $\mu_{SR} = E\{(x_1 + x_n)/2\} = [E\{x_1\} + E\{x_2\}]/2 = (\mu_x + \mu_x)/2 = \mu_x$

Por lo tanto, el semirango es insesgado.

b. $(x_1 + x_2 + x_{n-1} + x_n)/3$

$E\{(x_1 + x_2 + x_{n-1} + x_n)/3\} = [E\{x_1\} + E\{x_2\} + E\{x_{n-1}\} + E\{x_n\}]/2 =$
$(\mu_x + \mu_x + \mu_x + \mu_x)/3 = 4\mu_x/3 \neq \mu_x$

Por lo que este estimador es sesgado; observe que si se hubiera dividido entre cuatro sería insesgado.

c. Media aritmética: $\bar{x} = E\{(x_1 + x_2 + x_3 + \cdots + x_n)/n\} = n\mu_x/n = \mu_x$

Por lo que el promedio muestral es un estimador insesgado.

Ejercicio 4.2 Se acostumbra definir a la varianza de dos formas diferentes:

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - \bar{x})^2$$

$$S^2_n = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{x})^2$$

Nótese que el primer estimador definido arriba es insesgado, ya que:

$$\mu_{S^2_{n-1}} = E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} = \frac{1}{(n-1)}\sum_{i=1}^{i=n}E\{(x_i - \bar{x})^2\}$$

$$E\{(x_i-\bar{x})^2\}=E\{(x_i-\mu_x-\bar{x}+\mu_x)^2\}$$

$$E\{(x_i-\bar{x})^2\}=E\{(x_i-\mu_x)^2-2(x_i-\mu_x)(\bar{x}-\mu_x)+(\bar{x}-\mu_x)^2\}$$

$$E\{(x_i-\bar{x})^2\}=E\{(x_i-\mu_x)^2\}-2E\{(x_i-\mu_x)\ (\bar{x}-\mu_x)\}+E\{(\bar{x}-\mu_x)^2\}$$

$$E\{(x_i-\bar{x})^2\}=\sigma_x^2-2E\left\{(x_i-\mu_x)\left(\frac{1}{n}\sum_{j=1}^{j=n}(x_j-\mu_x)\right)\right\}+\frac{\sigma_x^2}{n}$$

$$E\{(x_i-\bar{x})^2\}=\sigma_x^2-\frac{2}{n}\sum_{j=1}^{n}E\{(x_i-\mu_x)(x_j-\mu_x)\}+\frac{\sigma_x^2}{n}$$

$$E\{(x_i-\bar{x})^2\}=\sigma_x^2-\frac{2}{n}\sigma_x^2+\frac{\sigma_x^2}{n}=\frac{n-1}{n}\sigma_x^2$$

$$\mu_{S_{n-1}^2}=\frac{n}{n-1}\left(\frac{n-1}{n}\sigma_x^2\right)=\sigma_x^2$$

En cambio, el segundo estimador de la varianza poblacional es sesgado, ya que:

$$\mu_{S_n^2}=E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}=E\left\{\frac{(n-1)}{1}\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}$$

$$\mu_{S_n^2}=\frac{(n-1)}{n}E\{S_{n-1}^2\}=\frac{(n-1)}{n}\sigma_x^2\neq\sigma_x^2$$

4.2.2. Eficiencia

Se dice que un estimador $\hat{\theta}_1$ es más eficiente o más preciso que otro estimador $\hat{\theta}_2$, si la varianza del primero es menor que la del segundo, es decir,

$$\sigma^2_{\hat{\theta}_1}<\sigma^2_{\hat{\theta}_2} \tag{4.7}$$

En la medida en que un estimador presenta menor varianza, es más eficiente o más preciso. El estimador más eficiente u óptimo será aquel que presenta la mínima variancia. Para el caso particular de la media poblacional, su estimador óptimo es la media muestral.

4.2.3. Error cuadrático medio mínimo

En estadística, el error cuadrático medio (ECM) de un estimador se define como la esperanza matemática o el valor esperado de los errores al cuadrado, es decir, la diferencia entre el estimador y el parámetro poblacional que se estima,

$$ECM(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\} \tag{4.8}$$

$$ECM(\hat{\theta}) = E\{(\hat{\theta} - \mu_{\hat{\theta}} + \mu_{\hat{\theta}} - \theta)^2\}$$

$$ECM(\hat{\theta}) = E\{(\hat{\theta} - \mu_{\hat{\theta}})^2\} + 2E\{(\hat{\theta} - \mu_{\hat{\theta}})(\mu_{\hat{\theta}} - \theta)\} + E\{(\mu_{\hat{\theta}} - \theta)^2\}$$

$$E\{(\hat{\theta} - \mu_{\hat{\theta}})^2\} = \sigma^2_{\hat{\theta}}$$

$$E\{(\mu_{\hat{\theta}} - \theta)^2\} = Sesgo^2$$

$$ECM(\hat{\theta}) = \sigma^2_{\hat{\theta}} + Sesgo^2 \tag{4.9}$$

El error cuadrático medio es un criterio importante para comparar dos estimadores. Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores puntuales del parámetro θ, y $ECM(\hat{\theta}_1)$ y $ECM(\hat{\theta}_2)$ sus errores cuadráticos medios, entonces, se define la eficiencia relativa de $\hat{\theta}_2$ sobre $\hat{\theta}_1$ como el siguiente cociente:

$$Efic\ Rel\ (\hat{\theta}_2, \hat{\theta}_1) = \frac{ECM(\hat{\theta}_1)}{ECM(\hat{\theta}_2)} \tag{4.10}$$

Si esta eficiencia relativa es menor que uno entonces el primer estimador es más eficiente que el segundo, pero si la eficiencia relativa es mayor que uno entonces el segundo estimador es más eficiente que el primero.

Ejercicio 4.3

Para el caso de los tres estimadores de la media poblacional definidos en el ejercicio 4.1, nótese que sus varianzas son:

$$\sigma^2_{SR} = \text{var}\left\{\frac{x_1 + x_n}{2}\right\} = \frac{1}{4}\text{var}\{x_1 + x_n\}$$

$$\sigma^2_{SR} = \frac{1}{4}\left[\text{var}\{x_1\} + \text{var}\{x_2\} + 2\,\text{cov}\{x_1, x_2\}\right]$$

$$\sigma^2_{SR} = \frac{1}{4}\left[\sigma^2_x + \sigma^2_x + 0\right] = \frac{\sigma^2_x}{2}$$

Por lo que su error cuadrático medio es ECM(SR) $= \sigma_x^2 / 2$, ya que no presenta sesgo al haberse comprobado que es insesgado.

En cambio

$$\sigma_2^2 = \text{var}\left\{\frac{x_1 + x_2 + x_{n-1+}x_n}{4}\right\} = \frac{1}{16}\text{var}\left\{x_1 + x_{2+}x_{n-1} + x_n\right\}$$

$$\sigma_2^2 = \frac{1}{16}\left[\text{var}\left\{x_1\right\} + \text{var}\left\{x_2\right\} + \text{var}\left\{x_{n-1}\right\} + \text{var}\left\{x_n\right\}\right]$$

$$\sigma_2^2 = \frac{1}{16}\left[4\sigma_x^2\right] = \frac{\sigma_x^2}{4}$$

Por lo que su error cuadrático medio es ECM(2) $= \sigma_x^2 / 4$

Para el caso de la media aritmética

$$\sigma_{\bar{x}}^2 = var\left\{\frac{1}{n}\sum_{i=1}^{i=n} x_i\right\} = \frac{1}{n^2}\ var\left\{x_1 + x_2 + \cdots + x_n\right\}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2}\left[var\left\{x_1\right\} + var\left\{x_2\right\} + \cdots + var\left\{x_n\right\}\right]$$

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2}\left[n\sigma_x^2\right] = \frac{\sigma_x^2}{n}$$

Por lo que su error cuadrático medio es ECM(media aritmética) $= \sigma_x^2 / n$

Sus eficiencias relativas implican que para $n > 4$ el estimador más eficiente o más preciso de la media poblacional, de los tres definidos anteriormente, es la media aritmética de las observaciones.

El estimador insesgado de varianza mínima es aquel que presenta la mínima varianza de todos los estimadores insesgados que existen.

Para conocerlo, se puede aplicar la desigualdad de la Cota Inferior de Cramér-Rao. La cota inferior de Cramér-Rao (CRLB por sus iniciales en inglés), llamada así en honor a Harald Cramér (1893-1985) y Calyampudi Radhakrishna Rao (1920-), expresa una cota inferior para la varianza de un estimador insesgado.

4.2.4. Teorema de la Cota Inferior de Cramér-Rao

Bajo condiciones muy generales, suponga que se tiene una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$ obtenida de una población cuya variable aleatoria es x, procedente de una distribución con función de densidad $f(x, \theta)$; entonces, la varianza de cualquier estimador insesgado $\hat{\theta}$ de θ cumplirá con la siguiente desigualdad:

$$\sigma^2_{\hat{\theta}} = \text{var}(\hat{\theta}) \geq \frac{1}{nE\left\{\left[\frac{\partial}{\partial\theta} Ln\left(f(x;\theta)\right)\right]^2\right\}} \tag{4.11}$$

Al amable lector que esté interesado en revisar la demostración de la desigualdad de Cramér-Rao, se le sugiere consultar el tema 8.3.2 de la página 426, de Alberto León García, *Probability, Statistics and Random Processes for Electrical Engineering,* Pearson, Prentice Hall, third edition, 2008.

La parte derecha de la anterior desigualdad es la denominada cota inferior de Cramér-Rao. Evidentemente, si se dispone de un estimador insesgado cuya varianza coincide con dicha cota se tendrá un estimador eficiente de θ (no se podrá obtener un estimador mejor según los criterios de sesgo y eficiencia). Conviene destacar la existencia de casos en los que dicha cota no es alcanzable; es decir, puede suceder que exista un estimador con la mínima varianza posible dentro del conjunto de estimadores insesgados de un parámetro, pero la varianza será superior a la cota de Cramér-Rao.

Ejercicio 4.4

Suponga que se tiene una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$ obtenida de una población cuya variable aleatoria x es normal de media μ y desviación estándar σ la cual se supondrá conocida. Con la aplicación de la desigualdad de Cramér-Rao obtenga un estimador insesgado de varianza mínima de la media poblacional.

En este caso

$$f(x;\mu)\ \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Se aplica el logaritmo de ambos lados de la igualdad

$$Ln\left[f(x,\mu)\right] = -Ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$\frac{\partial}{\partial\mu}\left[Ln\left(f(x,\mu)\right)\right] = \frac{1}{\sigma}\left(\frac{x-\mu}{\sigma}\right)$$

$$\left\{\frac{\partial}{\partial\mu}\left[Ln\left(f(x,\mu)\right)\right]\right\}^2 = \frac{1}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2$$

Sustituyendo en la expresión (4.11)

$$\hat{\sigma}^2_{\hat{\mu}} \geq \frac{1}{nE\left\{\frac{1}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}}$$

$$\hat{\sigma}^2_{\hat{\mu}} \geq \frac{1}{\frac{n}{\sigma^4}E\left\{(x-\mu)^2\right\}}$$

$$\hat{\sigma}^2_{\hat{\mu}} \geq \frac{\sigma^2}{n}$$

Se sabe que

$$\hat{\sigma}^2_{\bar{x}} \geq \frac{\sigma^2}{n}$$

Este estimador satisface la cota inferior de Cramér-Rao con una igualdad, en consecuencia, se puede concluir que la media aritmética muestral es un estimador insesgado de mínima varianza de la media poblacional.

4.2.5. Consistencia

Un estimador $\hat{\theta}$, es consistente en la medida en que el valor de dicho estimador tiende a aproximarse cada vez más al valor del parámetro poblacional θ, cuando el tamaño de muestra tiende a crecer, esto implica,

$$Lim_{n\to\infty} E\left\{\hat{\theta}\right\} = \theta \qquad (4.12)$$

$$Lim_{n\to\infty}\left\{\sigma^2_{\hat{\theta}}\right\}=\theta \tag{4.13}$$

Ejercicio 4.5

Nótese que en el caso de la media aritmética como estimador de la media poblacional:

$$\mathrm{Lim}_{n\to\infty}\left(\sigma_{\bar{x}}\right)=\mathrm{Lim}_{n\to\infty}\left(\frac{\sigma^2_x}{n}\right)=0$$

Por lo que se puede concluir que la media aritmética muestral es un estimador consistente, ya que en la medida en que se toma una muestra más y más grande, el estimador se acerca cada vez más a la media poblacional.

4.2.6. Robustez

El estimador $\hat{\theta}$ será un estimador robusto del parámetro θ si el no cumplimiento de algunas hipótesis previas en las que se basa la estimación (por ejemplo, suponer que la distribución de probabilidad de los datos es normal, o que la muestra fue obtenida aleatoriamente), no altera de manera significativa los resultados que este proporciona.

Ejercicio 4.6

La media aritmética muestral como estimador puntual de la media poblacional no depende de la distribución de probabilidad que presenten los datos de la población, por lo cual se trata de un estimador robusto.

4.2.7. Suficiencia

Se dice que un estimador es suficiente cuando resume toda la información relevante contenida en la muestra, de forma que ningún otro estimador pueda proporcionar información adicional sobre el parámetro desconocido de la población. Formalmente, se dice que un estadístico $\hat{\theta}$ es suficiente para θ si la distribución condicionada de x dado el valor $\hat{\theta}(x)$, no depende de θ.

Teorema de Neyman-Fisher: Suponga que se tiene una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$ obtenida de una población cuya variable aleatoria es x. Suponga que cada una de las variables x_i son estadísticamente independientes entre sí y que todas ellas pertenecen a la misma población. Si para un estimador $\hat{\theta}$, su función de densidad o de probabilidad de la muestra puede descomponerse en la forma

$$f(x_1, x_2, \cdots, x_n, \theta) = g(t, \theta)\ h(x_1, x_2, \cdots, x_n) \tag{4.14}$$

Donde $\hat{\theta}$ y h no dependen de θ, entonces $\hat{\theta}$ es un estimador suficiente.

Ejercicio 4.7

Para ilustrar un ejemplo de un estimador suficiente, suponga que cada una de las variables aleatorias estadísticamente independientes x_i presentan una función de densidad tipo Poisson con parámetro λ; entonces su función de probabilidad conjunta, por el hecho de ser independientes las variables, se obtiene como el producto de cada una de sus funciones de probabilidad, es decir,

$$f(x_1, x_2, \cdots, x_n, \lambda) = \left(\frac{e^{-\lambda}\ \lambda^{x_1}}{x_1!}\right)\left(\frac{e^{-\lambda}\ \lambda^{x_2}}{x_2!}\right)\cdots\left(\frac{e^{-\lambda}\ \lambda^{x_n}}{x_n!}\right)$$

$$f(x_1, x_2, \cdots, x_n, \lambda) = \frac{e^{-n\lambda}\ \lambda^{\sum_{i=1}^{i=n} x_i}}{x_1!\ x_2!\ \cdots\ x_n!} = \left[e^{-n\lambda}\ \lambda^{\sum_{i=1}^{i=n} x_i}\right]\left[\frac{1}{x_1!\ x_2!\ \cdots\ x_n!}\right]$$

$$f(x_1, x_2, \cdots, x_n, \lambda) = g(\bar{x}, \lambda)\ h(x_1, x_2, \cdots, x_n)$$

Donde

$$n\bar{x} = \sum_{i=1}^{i=n} x_i$$

Por lo que el estadístico media muestral para el caso de la distribución de Poisson es suficiente.

4.2.8. Invariancia

Se dice que un estimador es invariante cuando el estimador de la función del parámetro coincide con la función del estimador del parámetro, es decir,

$$[f(\theta)]^* = f(\theta^*) \tag{4.15}$$

Por ejemplo, si para estimar la varianza poblacional se utiliza la varianza muestral, entonces para estimar la desviación estándar poblacional será razonable utilizar la desviación estándar muestral.

4.3. Métodos de obtención de estimadores puntuales

Históricamente, para la obtención de estimadores puntuales se han diseñado procedimientos matemáticos, entre los cuales se pueden citar:

1. Principio de los Momentos.
2. Método de la χ^2 mínima.
3. Método de los mínimos cuadrados.
4. Principio de la máxima verosimilitud.
5. Principio de Bayes.

La aplicación de estos métodos a casos particulares de funciones de probabilidad específicas conduce a estimadores que pueden diferir según el método que se aplique y por lo mismo presentar diferentes atributos de bondad; por ello, solo se explicarán los más usuales y los que mejores atributos arrojan; estos son, el método de los momentos y el método de máxima verosimilitud.

4.4. Método de los momentos

El método de los momentos está sustentado en el siguiente principio: si una muestra es representativa perfectamente de una población, los momentos muestrales y poblacionales deben coincidir.

Suponga que x es una variable continua con función densidad de probabilidad $f(x; \theta_1, \theta_1, \cdots, \theta_k)$ o x es una variable aleatoria discreta con función de probabilidad $p(x; \theta_1, \theta_1, \cdots, \theta_k)$, los momentos de orden k con respecto al origen, están definidos como:

$$\mu'_t = E\{x^t\} = \int_{x \in Rx} x^t f(x; \theta_1, \theta_2, \cdots, \theta_k)\, dx \qquad t = 1,2,\cdots, k \qquad x\_\ Continua$$

$$\mu'_t = E\{x^t\} = \sum_{x \in Rx} x^t f(x; \theta_1, \theta_2, \cdots, \theta_k) \qquad t = 1, 2, \cdots, k \qquad x\_\ Discreta$$

Suponga que se tiene una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$ obtenida de una población cuya variable aleatoria es x. Suponga que cada una de las variables x_i son estadísticamente independientes entre sí y que todas ellas pertenecen a la misma población, lo que implica que todas ellas presentan la misma media y la misma varianza. Los momentos muestrales de orden k con respecto al origen están definidos como:

$$m'_t = \frac{1}{n} \sum_{i=1}^{i=n} x_i^t \qquad t = 1, 2, \cdots, k$$

Al igualar los primeros momentos de orden k de la población, con los primeros momentos de orden k de la muestra, se obtiene un sistema de k ecuaciones con k incógnitas, el cual al ser resuelto proporciona la solución estimada de los k parámetros $\theta_1, \theta_2, \cdots, \theta_k$.

Para ilustrar la aplicación del método de los momentos se tomarán ciertos ejemplos con una distribución de probabilidad bien definida.

Ejercicio 4.8

Suponga que x representa el número de llegadas o de éxitos en un cierto período de tiempo, lo cual se modela con la distribución de Poisson:

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad x = 0, 1, 2, \cdots$$

El primer momento con respecto al origen de esta variable aleatoria es su media, la cual es λ. Por lo tanto,

$$\lambda = \bar{x} = \frac{1}{n} \sum_{i=1}^{i=n} x_i \tag{4.16}$$

Esto implica que un estimador de λ para la distribución de Poisson es la media aritmética de una muestra aleatoria de las llegadas.

Ejercicio 4.9

El CrossFit es un sistema de entrenamiento de fuerza y acondicionamiento físico basado en ejercicios funcionales constantemente variados, realizados a una alta intensidad. Suponga que se realizan cuatro rutinas de ejercicios: barra, lagartijas, sentadillas y fondos sin parar durante cuatro minutos en promedio cada uno de ellos. Suponga que el tiempo de cada rutina se comporta probabilísticamente como una función exponencial negativa. El tiempo total de ejercicio de una vuelta completa sería la suma de los tiempos de cada uno de los ejercicios, esto implicaría que una vuelta completa de los ejercicios se tardaría en promedio 16 minutos. En probabilidad se demuestra que la suma de r tiempos exponenciales negativos con el mismo parámetro λ genera una distribución tipo gamma, a la cual se le denomina Erlang, cuando r es natural. Esto implicaría que $\lambda = 4$ minutos y $r = 4$. Se partirá del hecho de que esto no se sabe y que se desea obtener estimadores puntuales para λ y para r, utilizando el método de los momentos.

La función de probabilidad gamma presenta la siguiente expresión:

$$f(t) = \frac{\lambda}{\Gamma(r)} (\lambda t)^{(r-1)} e^{-\lambda t} \qquad t > 0$$

Su media y su varianza están dadas por las siguientes expresiones:

$$\mu_t = \frac{r}{\lambda}$$

$$\sigma_t^2 = \mu_2 - \mu_1^2 = \frac{r}{\lambda^2}$$

Nótese que son dos parámetros, por lo cual se igualarán los momentos de orden uno y dos de la población con los de una muestra $\{x_1, x_2, x_3, \cdots, x_n\}$, es decir,

$$\mu_1 = \frac{r}{\lambda} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

$$\mu_2 = \frac{r}{\lambda^2} + \left(\frac{r}{\lambda}\right)^2 = \frac{1}{n}\sum_{i=1}^{i=n} x_i^2$$

Como se puede apreciar, se tiene un sistema de dos ecuaciones con dos incógnitas. Se despeja r de la primera y se sustituye en la segunda para dejar una sola incógnita y despejarla:

$$r = \frac{\lambda}{n}\sum_{i=1}^{i=n} x_i$$

$$\left(\frac{\lambda}{n}\sum_{i=1}^{i=n} x_i\right) + \left(\frac{\lambda}{n}\sum_{i=1}^{i=n} x_i\right)^2 = \frac{\lambda^2}{n}\sum_{i=1}^{i=n} x_i^2$$

$$\frac{\lambda}{n}\sum_{i=1}^{i=n} x_i + \frac{\lambda^2}{n^2}\left(\sum_{i=1}^{i=n} x_i\right)^2 = \frac{\lambda^2}{n}\sum_{i=1}^{i=n} x_i^2$$

$$\lambda = \frac{\sum_{i=1}^{i=n} x_i}{\left[\sum_{i=1}^{i=n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{i=n} x_i\right)^2\right]} = \frac{n\bar{x}}{(n-1)S_{n-1}^2} \tag{4.17}$$

$$r = \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{\left[n\sum_{i=1}^{i=n} x_i^2 - \left(\sum_{i=1}^{i=n} x_i\right)^2\right]} = \frac{n\bar{x}^2}{(n-1)S_{n-1}^2} \tag{4.18}$$

Para probar la bondad del método de los momentos, se generará una muestra aleatoria de $n = 100$ números, con una distribución gamma con $\lambda = 4$ y $r = 4$, y se calcularán los estimadores de λ y r con las expresiones 4.16 y 4.17. Para generar la muestra aleatoria se usará el software R con el siguiente (comando, se muestran los datos generados).

```
> datos<-c(rgamma(100,4,rate=4))
> print(datos)
  [1] 1.5934668 0.7574606 0.7858794 1.2257872 0.7815434 1.1460247 0.4359186
  [8] 0.4561479 0.7698735 0.4330150 0.6701616 1.3912629 0.3323386 0.4501510
 [15] 1.9835707 0.7156921 0.8000181 1.0099823 1.6248275 2.0216266 1.4556469
 [22] 0.8935156 0.6530962 1.5326037 0.8709999 0.6036104 0.6046257 1.1486221
 [29] 0.2908402 0.9649620 0.7224906 1.7273219 0.8474997 0.2248012 1.5245099
 [36] 1.4390566 0.6693289 0.5829331 0.6972350 1.6551081 1.0901931 1.4662126
 [43] 0.8438658 1.4238541 0.3841886 2.2050520 0.7355364 0.4981745 2.5834837
 [50] 1.0560409 1.1705668 0.5312207 2.2633367 1.2835803 1.1432091 0.4660503
 [57] 1.0970871 0.5226934 2.2684550 0.8852529 1.5121324 2.1834056 1.4798392
 [64] 0.6354269 1.7108894 3.3166699 1.0204960 0.9553502 0.7045112 0.8712951
 [71] 1.6308135 1.6314052 0.6066914 1.5144655 0.6934443 0.5707609 0.5864769
 [78] 1.1555409 1.0508832 1.3953168 1.2375083 1.0549612 1.0294185 0.7582207
 [85] 0.4487017 0.5688171 0.4391919 0.6402583 0.5377065 0.4870611 1.5033831
 [92] 1.2955123 1.6054766 0.4893042 0.4149199 1.1766419 1.8286200 1.6743515
 [99] 1.1831398 0.2862535
```

Para calcular la suma de todos estos datos generados con el software R basta con aplicar el comando > sum(datos), y para calcular la suma de los cuadrados de estos datos basta con aplicar el comando > sum(datos^2), lo cual da los siguientes resultados:

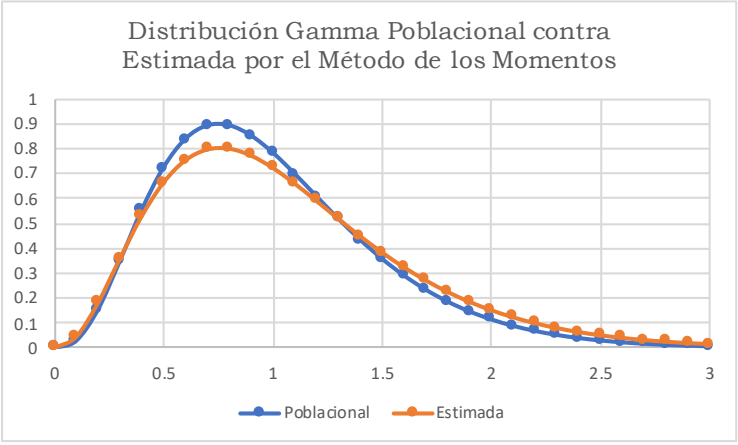
sum(datos) = 106.2969
sum(datos^2) = 145.2646

Por lo que los estimadores son:

$\lambda = 106.2969/(145.2646-106.2969^2/100) = 3.294 \rightarrow$ Error Rel = 17.66%
$r = (106.2969)^2/(100*145.2646-106.2969^2) = 3.5 \rightarrow$ Error Rel = 12.48%

En la siguiente figura se muestra un comparativo entre el modelo poblacional con $\lambda = 4$ y $r = 4$ contra el modelo estimado por el método de los momentos con $\lambda^ = 3.294$ y $r^ = 3.5$.

Figura 4.1



Ejercicio 4.10

El método de los momentos no siempre conduce a una solución sencilla de obtener. Suponga que x es una variable aleatoria con distribución Weibull, con función de densidad de probabilidad

$$f(t;\beta,\delta)=\frac{\beta}{\delta}\left(\frac{t}{\delta}\right)^{(\beta-1)}e^{-\left(\frac{t}{\delta}\right)^{\beta}} \qquad t>0$$

La media y la varianza de esta distribución son:

$$\mu_x=\delta\Gamma\left(\frac{1}{\beta}-1\right)$$

$$\sigma_x^2=\delta^2\left\{\Gamma\left(\frac{2}{\beta}+1\right)-\left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2\right\}$$

El sistema de dos ecuaciones con dos incógnitas que resulta es el siguiente:

$$\delta\Gamma\left(\frac{1}{\beta}-1\right)=\frac{1}{n}\sum_{i=1}^{i=n}x_i$$

$$\delta^2\Gamma\left(\frac{2}{\beta}+1\right)=\frac{1}{n}\sum_{i=1}^{i=n}x_i^2$$

Intentar obtener estimadores puntuales de los parámetros β y δ, utilizando el método de los momentos para esta distribución de probabilidad, da origen a un sistema no lineal de dos ecuaciones con dos incógnitas prácticamente imposible de resolver analíticamente, aunque se puede resolver a través de métodos numéricos, como se ilustrará más adelante.

4.5. Método de máxima verosimilitud

Un principio muy en uso, que conduce a estimadores con muchos atributos deseables de "bondad", por procedimientos matemáticos rutinarios fácilmente aplicables, es el de la máxima verosimilitud, establecido por Sir Ronald Fisher.

El procedimiento para determinar la estimación de máxima verosimilitud de un parámetro q de una población es el siguiente:

1. Suponga que se tiene una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$ obtenida de una población cuya variable aleatoria es x con función densidad de probabilidad $f(x; \theta_1, \theta_2, \cdots, \theta_k)$, la cual presenta k parámetros poblacionales. Suponga que cada una de las variables x_i son estadísticamente independientes entre sí y que todas ellas pertenecen a la misma población. Se forma la función de probabilidad conjunta

$$g(x_1, x_2, \cdots, x_n, \theta_1, \theta_2, \cdots, \theta_k)$$

$$= f(x_1; \theta_1, \theta_2, \cdots, \theta_k)\, f(x_2; \theta_1, \theta_2, \cdots, \theta_k) \cdots f(x_n; \theta_1, \theta_2, \cdots, \theta_k)$$

$$g(x_1, x_2, \cdots, x_n, \theta_1, \theta_2, \cdots, \theta_k) = \prod_{i=1}^{i=n} f(x_1; \theta_1, \theta_2, \cdots, \theta_k)$$

2. Dado que se trata de productos de funciones de probabilidad, y debido a que se pretende maximizar esta función con respecto al parámetro q, es conveniente por facilidad, trabajar con el logaritmo de la función de probabilidad conjunta, a la cual se le denominará función de verosimilitud L.

$$L = Ln\{g(x_1, x_2, \cdots, x_n, \theta_1, \theta_2, \cdots, \theta_k)\}$$

$$= Ln\{f(x_1, \theta_1, \theta_2, \cdots, \theta_k)\, f(x_2, \theta_1, \theta_2, \cdots, \theta_k) \cdots f(x_n, \theta_1, \theta_2, \cdots, \theta_k)\}$$

$$L = Ln\left\{\prod_{i=1}^{i=n} f(x_1, \theta_1, \theta_2, \cdots, \theta_k)\right\} \tag{4.19}$$

3. Se obtiene el máximo de esta función por el criterio de la primera derivada, resolviendo el sistema de ecuaciones

$$\frac{\partial L}{\partial \theta_1}=0$$
$$\frac{\partial L}{\partial \theta_2}=0$$
$$\vdots$$
$$\frac{\partial L}{\partial \theta_k}=0$$

Ejercicio 4.11 Obtenga un estimador de máxima verosimilitud del parámetro p en la distribución binomial.

Su función de probabilidad es:

$$p(x)=\binom{n}{x}p^x(1-p)^{n-x} \qquad x=0,\,1,\,2,\,\cdots,\,n$$

La función de probabilidad conjunta es:

$$p(x_1,x_2,\cdots,x_m,p)=\left\{\binom{n}{x_1}p^{x_1}(1-p)^{n-x_1}\right\}\left\{\binom{n}{x_2}p^{x_2}(1-p)^{n-x_2}\right\}\cdots\left\{\binom{n}{x_m}p^{x_m}(1-p)^{n-x_m}\right\}$$

Su función de verosimilitud es:

$$L=Ln\left\{\binom{n}{x_1}\binom{n}{x_2}\cdots\binom{n}{x_m}p^{\sum_{i=1}^{m}x_i}(1-p)^{mn-\sum_{i=1}^{i=m}x_i}1\right\}$$

$$L=Ln\left\{\binom{n}{x_1}\binom{n}{x_2}\cdots\binom{n}{x_m}\right\}+\left(\sum_{i=1}^{i=m}x_i\right)Ln(p)+\left(mn-\sum_{i=1}^{i=m}x_i\right)Ln(1-p)$$

Derivando esta última expresión con respecto al parámetro p:

$$\frac{\partial L}{\partial p}=\frac{\left(\sum_{i=1}^{i=m}x_i\right)}{p}-\frac{\left(mn-\sum_{i=1}^{i=m}x_i\right)}{1-p}=0$$

$$(1-p)\left(\sum_{i=1}^{i=m}x_i\right)=p\left(mn-\sum_{i=1}^{i=m}x_i\right)$$

$$\left(\sum_{i=1}^{i=m} x_i\right) - p\left(\sum_{i=1}^{i=m} x_i\right) = pmn - p\left(\sum_{i=1}^{i=m} x_i\right)$$

$$p = \frac{\left(\sum_{i=1}^{i=m} x_i\right)}{mn} \tag{4.20}$$

Este estimador que se dedujo fue partiendo del hecho de que todas las muestras tomadas tuvieran el mismo tamaño n. Si en cada muestra el tamaño fuera diferente, entonces el estimador sería:

$$p = \frac{\sum_{i=1}^{m} x_i}{\left(\sum_{i=1}^{i=m} n_i\right)} \tag{4.21}$$

Ejercicio 4.12 Obtenga el estimador de máxima verosimilitud del parámetro l en la distribución exponencial negativa.

Su función de densidad de probabilidad está dada por la expresión:

$$f(x;\lambda) = \lambda e^{-\lambda x} \qquad x > 0$$

Su función de probabilidad conjunta:

$$g(x_1, x_2, \cdots, x_n; \lambda) = \left(\lambda e^{-\lambda x_1}\right)\left(\lambda e^{-\lambda x_2}\right)\cdots\left(\lambda e^{-\lambda x_n}\right)$$

$$g(x_1, x_2, \cdots, x_n; \lambda) = \lambda^n e^{-\lambda(x_1 + x_2 + \cdots + x_n)}$$

Su función de verosimilitud es:

$$L = Ln\left[\lambda^n e^{-\lambda(x_1 + x_2 + \cdots + x_n)}\right]$$

$$L = nLn(\lambda) - \lambda\left(\sum_{i=1}^{n} x_i\right)$$

Derivando esta última expresión con respecto al parámetro λ

$$\frac{\partial L}{\partial \lambda} = \frac{n}{\lambda} - \left(\sum_{i=1}^{n} x_i \right)$$

$$\frac{n}{\lambda} = \left(\sum_{i=1}^{n} x_i \right)$$

$$\lambda = \frac{n}{\left(\sum_{i=1}^{n} x_i \right)} = \frac{1}{\bar{x}} \tag{4.22}$$

Ejercicio 4.13 Determine los estimadores de máxima verosimilitud de los parámetros μ y σ en la distribución normal.

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Su función de densidad conjunta es:

$$f(x_1, x_2, \cdots, x_n \,;\, \sigma) = \frac{1}{\sigma^n (2\pi)^{n/2}} \; e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu}{\sigma}\right)^2 + \left(\frac{x_2-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{x_n-\mu}{\sigma}\right)^2\right]}$$

Su función de verosimilitud

$$L = Ln\left\{ \frac{1}{\sigma^n (2\pi)^{n/2}} \; e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu}{\sigma}\right)^2 + \left(\frac{x_2-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{x_n-\mu}{\sigma}\right)^2\right]} \right\}$$

$$L = nLn(\sigma) - \frac{n}{2} Ln(2\pi) - \frac{1}{2} \sum_{i=1}^{i=n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

$$L = -nLn(\sigma) - \frac{n}{2} Ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{i=n} (x_i - \mu)^2$$

Derivando parcialmente esta última expresión con respecto a los parámetros μ y σ e igualando a cero dichas derivadas, se obtiene un sistema de dos ecuaciones con dos incógnitas, el cual al resolverlo proporciona los estimadores solicitados:

$$\frac{\partial L}{\partial \mu} = \frac{-1}{\sigma^2} \sum_{i=1}^{i=n} (x_i - \mu) = 0$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{i=n} (x_i - \mu)^2 = 0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \tag{4.23}$$

$$\sigma^2 = \frac{\sum_{i=1}^{i=n} (x_i - \mu)^2}{n} \tag{4.24}$$

El método de máxima verosimilitud es muy utilizado en la práctica debido a que proporciona estimadores consistentes, asintóticamente eficientes, insesgados y normalmente distribuidos en muestras grandes. Sin embargo, en ocasiones la ecuación de verosimilitud (la derivada de la función de verosimilitud, o su logaritmo, igualada a cero) presenta más de una solución lo cual llevaría a tener varios posibles estimadores para un mismo parámetro. También, es frecuente que sea imposible obtener una solución, analítica del problema de optimización que plantea la estimación por máxima verosimilitud. En tales situaciones será necesario aplicar un método de solución numérica. Una seria desventaja que se puede cometer es utilizar una distribución equivocada, pues el estimador depende de la distribución en el proceso de optimización. Por otra parte, no se puede asegurar que las propiedades de estos estimadores sean válidas para el caso de muestras pequeñas.

Ejercicio 4.14 Obtenga estimadores de los parámetros β y δ, por el método de máxima verosimilitud de la función densidad de probabilidad tipo Weibull, dada por la expresión:

$$f(t; \beta, \delta) = \frac{\beta}{\delta} \left(\frac{t}{\delta} \right)^{(\beta - 1)} e^{-\left(\frac{t}{\delta}\right)^{\beta}} \qquad t > 0$$

Se obtiene la función de verosimilitud:

$$f(t_1, t_2, \cdots, t_n; \beta, \delta) = \left[\frac{\beta}{\delta}\left(\frac{t_1}{\delta}\right)^{(\beta-1)} e^{-\left(\frac{t_1}{\delta}\right)^{\beta}}\right]\left[\frac{\beta}{\delta}\left(\frac{t_2}{\delta}\right)^{(\beta-1)} e^{-\left(\frac{t_2}{\delta}\right)^{\beta}}\right] \cdots \left[\frac{\beta}{\delta}\left(\frac{t_n}{\delta}\right)^{(\beta-1)} e^{-\left(\frac{t_n}{\delta}\right)^{\beta}}\right]$$

$$f(t_1, t_2, \cdots, t_n; \beta, \delta) = \left(\frac{\beta}{\delta}\right)^n \left(\frac{1}{\delta}\right)^{n(\beta-1)} [t_1\, t_2 \cdots t_n]^{(\beta-1)}\, e^{-\frac{1}{\delta^{\beta}}\left[t_1^{\beta} + t_2^{\beta} + \cdots + t_n^{\beta}\right]}$$

$$L = nLn(\beta) - nLn(\delta) - n(\beta-1)\, Ln(\delta) + (\beta-1) \sum_{i=1}^{i=n} Ln(t_i) - \sum_{i=1}^{n}\left(\frac{t_i}{\delta}\right)^{\beta}$$

$$L = nLn(\beta) - n\beta Ln(\delta) + (\beta-1)\sum_{i=1}^{i=n} Ln(t_i) - \sum_{i=1}^{n}\left(\frac{t_i}{\delta}\right)^{\beta}$$

Se deriva con respecto a β y δ y se iguala a cero, obteniéndose un sistema de ecuaciones de dos por dos:

$$\frac{\partial L}{\partial \beta} = \frac{n}{\beta} - nLn(\delta) + \sum_{i=1}^{i=n} Ln(t_i) - \sum_{i=1}^{n}\left(\frac{t_i}{\delta}\right)^{\beta} Ln\left(\frac{t_i}{\delta}\right) = 0$$

$$\frac{\partial L}{\partial \delta} = -\frac{n\beta}{\delta} + \frac{\beta}{\delta^{n+1}} \sum_{i=1}^{i=n} t_i^{\beta} = 0$$

Despejando δ de la segunda expresión y sustituyendo en la primera, se obtiene el siguiente sistema:

$$\frac{\sum_{i=1}^{i=n} t_i^{\hat{\beta}}\, Ln(t_i)}{\sum_{i=1}^{i=n} t_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} - \frac{1}{n}\sum_{i=1}^{i=n} Ln(t_i) = 0 \qquad (4.25)$$

$$\hat{\delta} = \left(\frac{\sum_{i=1}^{i=n} t_i^{\hat{\beta}}}{n}\right)^{\frac{1}{\hat{\beta}}} \qquad (4.26)$$

para $\gamma = 0$

Para resolver la primera ecuación en términos de β, es necesario usar algún método numérico, como puede ser el Método de Newton Raphson.

El método de Newton-Raphson obtiene la solución en forma iterativa, aplicando el siguiente algoritmo:

$$\beta_{j+1} = \beta_j - \frac{f(\beta_j)}{f(\beta_j)} \tag{4.27}$$

Donde

$$f(\beta_j) = \frac{\sum_{i=1}^{i=n} t_i^{\beta_j} \, Ln(t_i)}{\sum_{i=1}^{i=n} t_i^{\beta_j}} - \frac{1}{\beta_j} - \frac{1}{n} \sum_{i=1}^{i=n} Ln(t_i) \tag{4.28}$$

y

$$f'(\beta_j) = \frac{\left[\sum_{i=1}^{i=n} t_i^{\beta_j}\right]\left[\sum_{i=1}^{i=n} t_i^{\beta_j} \left(Ln(t_i)\right)^2\right] - \left[\sum_{i=1}^{i=n} t_i^{\beta_j} Ln(t_i)\right]^2}{\left[\sum_{i=1}^{i=n} t_i^{\beta_j}\right]^2} + \frac{1}{\beta_j^2} \tag{4.29}$$

Para ilustrar la forma en que se aplica lo anterior, se generará una muestra aleatoria de $n = 100$ datos con una distribución tipo Weibull, con el parámetro de forma $\beta = 1.5$ y el parámetro de escala $\delta = 1$. Con esta muestra de 100 datos se obtendrán los estimadores.

Para generar los cien datos se utilizará R con el comando

- Datos < -c(rweibull(100, 1.5, 1))

La muestra generada es la siguiente:

```
0.61426443 0.27121587 0.76517094 1.45757352 0.50736249 0.51468041
0.41443048 0.26660073 0.37567745 0.51194583 1.39209232 3.22679587
0.23780816 0.92201253 1.42903943 0.34234534 2.17527070 0.93798388
0.61305840 0.71692269 2.30698626 0.41103681 0.80216707 1.20059332
1.37971802 0.56307206 0.13278185 0.95776679 0.32000036 0.68107704
2.57946931 0.88399725 0.24671286 0.70774615 0.57675426 0.27839202
1.05934010 0.51506626 0.34484812 1.04850524 1.91538078 0.95123633
0.79891864 0.85689794 0.24125934 1.24501569 0.55906922 0.46538162
1.29219491 1.73313249 1.24778207 1.32795727 0.99981188 1.40136699
1.12038221 0.74978423 1.11050228 1.26613184 0.91062838 0.59141873
0.81055767 0.93121847 0.02984827 1.25197084 2.72143844 1.45070569
1.77313803 1.05675106 2.00528730 0.56884391 0.76306232 1.94209828
0.53093321 0.25276064 0.20262997 0.44333940 0.26986516 0.65391789
0.77357316 1.11527485 0.76472767 1.62648617 1.48202115 0.22383349
0.29704295 0.81042155 0.99492232 1.03593738 1.03978876 0.40902901
1.77758444 0.71200106 0.56100915 0.70889950 0.55310233 2.29939989
0.34529644 0.66307579 2.13744432 0.35385034
```

Para resolver la primera ecuación se corrió el algoritmo de Newton-Raphson, utilizando Excel durante cinco iteraciones, obteniendo los siguientes resultados:

Figura 4.2

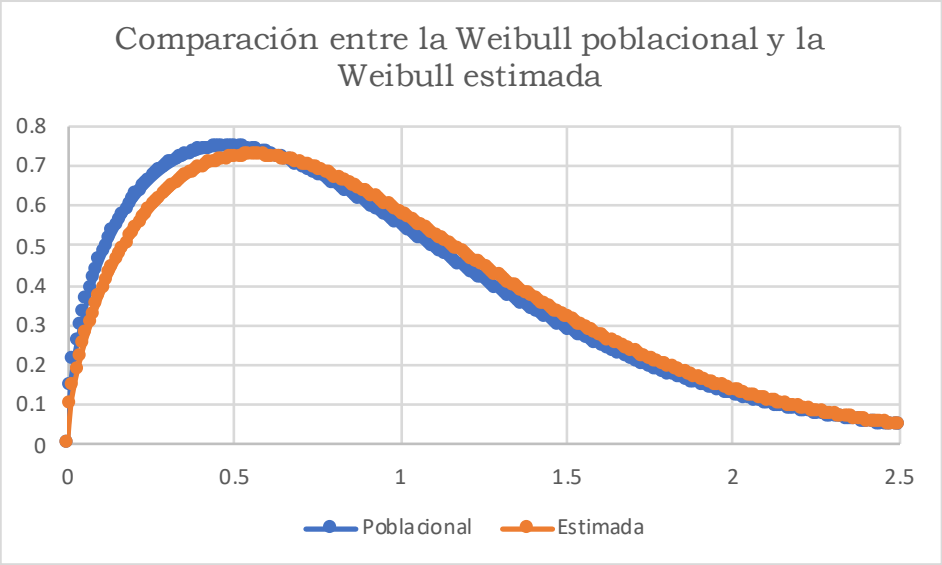
Iteración	β_i	β_{i+1}	Error
0	1	1.412825659	0.41282566
1	1.412825659	1.573600906	0.16077525
2	1.573600906	1.58737438	0.01377347
3	1.58737438	1.587458693	8.4313E-05
4	1.587458693	1.587458696	3.1129E-09

Con el último resultado de beta se calcula el valor de delta, obteniéndose que los estimadores más adecuados para este conjunto de datos son:

$\beta = 1.587458696$
$\delta = 1.049503301$

En la siguiente figura se comparan la distribución Weibull poblacional utilizada para generar los números aleatorios, con la distribución Weibull estimada.

Figura 4.3



Ejercicio 4.15

Sea x una variable aleatoria con distribución de probabilidad tipo beta, con función de densidad dada por la siguiente expresión:

$$f(x;a,b,\lambda,r)=\frac{\Gamma(\lambda+r)(x-a)^{(\lambda-1)}(b-x)^{(r-1)}}{\Gamma(\lambda)\Gamma(r)(b-a)^{(\lambda+r-1)}} \quad a\le x\le b$$

En este caso su función de máxima verosimilitud está dada por la expresión:

$$L=n\left[Ln\left(\Gamma(\lambda+r)\right)-Ln\left(\Gamma(r)\right)+(1-\lambda-r)\,Ln(b-a)\right]+$$
$$(\lambda-1)\sum_{i=1}^{i=n}Ln(x_i-a)+(r-1)\sum_{i=1}^{i=n}Ln(b-x_i)$$

Derivando con respecto a los parámetros a, b, λ y r respectivamente, e igualando a cero, se obtiene el siguiente sistema de cuatro ecuaciones con cuatro incógnitas:

$$\psi(\hat{\lambda})-\psi(\hat{\lambda}+\hat{r})=\frac{1}{n}\sum_{i=1}^{i=n}Ln(x_i-\hat{a})-nLn(\hat{b}-\hat{a})$$

$$\psi(\hat{r})-\psi(\hat{\lambda}+\hat{r})=\frac{1}{n}\sum_{i=1}^{i=n}Ln(\hat{b}-x_i)-nLn(\hat{b}-\hat{a})$$

$$\frac{\hat{\lambda}+\hat{r}-1}{\hat{\lambda}-1}+\left(\hat{b}-\hat{a}\right)=\frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{\left(x_i-\hat{a}\right)} \tag{4.30}$$

$$\frac{\hat{\lambda}+\hat{r}-1}{\hat{r}-1}+\left(\hat{b}-\hat{a}\right)=\frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{\left(\hat{b}-x_i\right)}$$

Donde a la función

$$\psi(x)=\frac{d}{dx}\left[Ln\left(\Gamma(x)\right)\right]=\frac{\Gamma'(x)}{\Gamma(x)} \tag{4.31}$$

Se le conoce como la función digamma.

Este sistema de cuatro ecuaciones con cuatro incógnitas puede ser resuelto por alguno de los métodos numéricos para resolver sistemas de ecuaciones no lineales. Para resolverlo Carnahan, J. V., en su artículo intitulado "Maximum likelihood estimation for the four parameter beta distribution, Comunications in Statistics Simulation and Computation" (1989), hace el análisis del mismo.

En las figuras 4.4., 4.5 y 4.6 se ilustran ejemplos de estimadores puntuales para parámetros estadísticos de diversas distribuciones de probabilidad discretas, continuas y de distribuciones muestrales.

Figura 4.4. Estimadores puntuales de distribuciones de probabilidad discretas

Distribución	Parámetros	Función de Probabilidad: p(x)	Estimadores
Bernoulli	0 < p < 1, probabilidad de éxito en cada ensayo	$p(x)=\begin{cases} p^x(1-p)^{(1-x)} & x=0,1 \\ 0 & en\_otro\_caso \end{cases}$	$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$
Binomial	n= 1, 2, ..., número de ensayos o tamaño de muestra m número de muestras 0 < p < 1, probabilidad constante de éxito en cada ensayo	$p(x)=\begin{cases} \binom{n}{x}p^x(1-p)^{(n-x)} & x=0,1,2,...,n \\ 0 & en\_otro\_caso \end{cases}$	$\hat{p}=\frac{\sum_{i=1}^{m}x_i}{mn}$ $n\_constante$ $\hat{p}=\frac{\sum_{i=1}^{m}x_i}{\left(\sum_{i=1}^{i=m}n_i\right)}$ $n\_variable$
Geométrica	0 < p < 1, probabilidad de éxito en cada ensayo	$p(x)=\begin{cases} p(1-p)^{(x-1)} & x=0,1,2,... \\ 0 & en\_otro\_caso \end{cases}$	$\hat{p}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\bar{x}}$
Pascal (Binomial Negativa)	0 < p < 1, probabilidad de éxito en cada ensayo r=1, 2, ... (r>0)	$p(x)=\begin{cases} \binom{x-1}{r-1}p^r(1-p)^{(x-1)} & x=0,1,2,... \\ 0 & en\_otro\_caso \end{cases}$	$\hat{p}=\frac{nr}{\sum_{i=1}^{i=n}x_i}=\frac{r}{\bar{x}}$ $r\_constante$ $\hat{p}=\frac{\bar{x}}{\frac{(n-1)}{n}S_{n-1}^2+\bar{x}}$ $\hat{r}=\frac{\bar{x}^2}{\frac{(n-1)}{n}S_{n-1}^2+\bar{x}}$
Hipergeométrica	N=1, 2, ..., tamaño del lote o población n= 1, 2, ..., N, tamaño de muestra D= 1, 2, ..., N, Número de Defectuosos en el Lote o Población	$p(x)=\begin{cases} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} & x=0,1,2,...,mín(n,D) \\ 0 & en\_otro\_caso \end{cases}$	$\hat{D}=\frac{N}{n^2}\sum_{i=1}^{i=n}x_i=\frac{N}{n}\bar{x}$ $n\_constante$ $N\_constante$
Poisson	c > 0, número de ocurrencias en promedio por unidad	$p(x)=\begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & x=0,1 \\ 0 & en\_otro\_caso \end{cases}$	$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$

Figura 4.5. Estimadores puntuales de distribuciones de probabilidad continuas

Distribución	Parámetros	Función Densidad de Probabilidad: f(x)	Estimadores
Uniforme	a, b, b>a	$f(x)=\begin{cases}\frac{1}{b-a} & a\le x\le b\\ 0 & en\_otro\_caso\end{cases}$	$\hat{a}=\bar{x}-\sqrt{8\bar{x}^2-\frac{3}{n}\sum_{i=1}^{n}x_i^2}$ $\hat{b}=\bar{x}+\sqrt{8\bar{x}^2-\frac{3}{n}\sum_{i=1}^{n}x_i^2}$
Triangular	a, b, c c > b > a	$f(x)=\begin{cases}\frac{2(x-a)}{(c-a)(b-a)} & si \quad a\le x\le b\\ \frac{2(c-x)}{(c-a)(c-b)} & si \quad b<x\le c\\ 0 & en\_cualquier\_otro\_caso\end{cases}$	
Exponencial	λ > 0	$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & en\_otro\_caso\end{cases}$	$\hat{\lambda}=\frac{n}{\sum_{i=1}^{i=n}x_i}=\frac{1}{\bar{x}}$
Normal	$\mu\in\Re$ $\sigma\in\Re$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	$\hat{\mu}_x=\frac{1}{n}\sum_{i=1}^{i=n}x_i=\bar{x}$ $\hat{\sigma}_x^2=\frac{(n-1)}{n}S_{n-1}^2+\bar{x}^2$
LogNormal	$\mu\in\Re$ $\sigma\in\Re$	$f(x)=\frac{1}{x\sigma_y\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(Ln(x)-\mu_y)^2}{\sigma_y^2}}$	$\hat{\mu}_x=\frac{1}{n}\sum_{i=1}^{i=n}Ln(x_i)$ $\hat{\sigma}_x^2=\frac{1}{n}\sum_{i=1}^{n}\left[Ln(x_i)-\hat{\mu}_x\right]^2$
Gamma	$r>0$ $\lambda>0$	$f(x)=\frac{\lambda}{\Gamma(r)}(\lambda x)^{(r-1)}e^{-\lambda x}$	$\hat{\lambda}=\frac{n\bar{x}}{(n-1)S_{n-1}^2}$ $\hat{r}=\frac{n\bar{x}^2}{(n-1)S_{n-1}^2}$
Beta	$r>0$ $\lambda>0$	$f(x)=\frac{\Gamma(\lambda+r)}{\Gamma(\lambda)\Gamma(r)}x^{\lambda-1}(1-x)^{r-1}$ $0\le x\le 1 \quad \lambda>0 \quad r>0$	$\psi(\hat{\lambda})-\psi(\hat{\lambda}+\hat{r})=\frac{1}{n}\sum_{i=1}^{i=n}Ln(x_i-\hat{a})-nLn(\hat{b}-\hat{a})$ $\psi(\hat{r})-\psi(\hat{\lambda}+\hat{r})=\frac{1}{n}\sum_{i=1}^{i=n}Ln(\hat{b}-x_i)-nLn(\hat{b}-\hat{a})$ $\frac{\hat{\lambda}+\hat{r}-1}{\hat{\lambda}-1}+(\hat{b}-\hat{a})=\frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{(x_i-\hat{a})}$ $\frac{\hat{\lambda}+\hat{r}-1}{\hat{r}-1}+(\hat{b}-\hat{a})=\frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{(\hat{b}-x_i)}$
Weibull	$\gamma\in\Re$ $\delta>0$ $\beta>0$	$f(x)=\frac{\beta}{\delta}\left(\frac{x-\gamma}{\delta}\right)^{(\beta-1)}e^{-\left(\frac{x-\gamma}{\delta}\right)^{\beta}}$	$\frac{\sum_{i=1}^{i=n}t_i^{\hat{\beta}}Ln(t_i)}{\sum_{i=1}^{i=n}t_i^{\hat{\beta}}}-\frac{1}{\hat{\beta}}-\frac{1}{n}\sum_{i=1}^{i=n}Ln(t_i)=0$ $\hat{\delta}=\left(\frac{\sum_{i=1}^{i=n}t_i^{\hat{\beta}}}{n}\right)^{\frac{1}{\hat{\beta}}}$ $para \quad \gamma=0$

Figura 4.6. Estimadores puntuales de distribuciones muestrales

Distribución	Parámetros	Función Densidad de Probabilidad: f(x)	Estimadores
Ji Cuadrada (χ^2)	$k \in \aleph$	$f(u) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} u^{\left(\frac{k}{2}-1\right)} e^{-\frac{u}{2}}$	$\hat{k} = \frac{1}{n}\sum_{i=1}^{i=n} \chi_i^2$
t de Student	$k \in \aleph$	$f(t) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\sqrt{\pi k}\Gamma\left[\frac{k}{2}\right]} \frac{1}{\left[\frac{t^2}{k}+1\right]^{\left(\frac{k+1}{2}\right)}}$	$\sum_{i=1}^{n} t_i = 0$ $\hat{k} = \frac{2\sum_{i=1}^{i=n} t_i^2}{\sum_{i=1}^{i=n} t_i^2 - 1}$
F de Fisher	$u, v \in \aleph$ $u > 0$ $v > 0$	$h(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\left(\frac{u}{2}\right)} f^{\left(\frac{u}{2}-1\right)}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\frac{u}{v}f+1\right]^{\frac{(u+v)}{2}}}$	

4.6. Estimación por intervalos de confianza para un parámetro poblacional

En el subtema anterior se analizaron los estimadores puntuales de parámetros poblacionales de algunas distribuciones de probabilidad específicas, las características deseables de dichos estimadores y al menos dos métodos para obtenerlos. Sin embargo, para muchas aplicaciones prácticas, una estimación puntual no es suficiente, ya que esta no proporciona indicadores para medir el grado de certeza o certidumbre sobre su valor. En la práctica profesional de muchas disciplinas, es de mayor utilidad estimar el parámetro poblacional θ, a partir de un intervalo de la forma

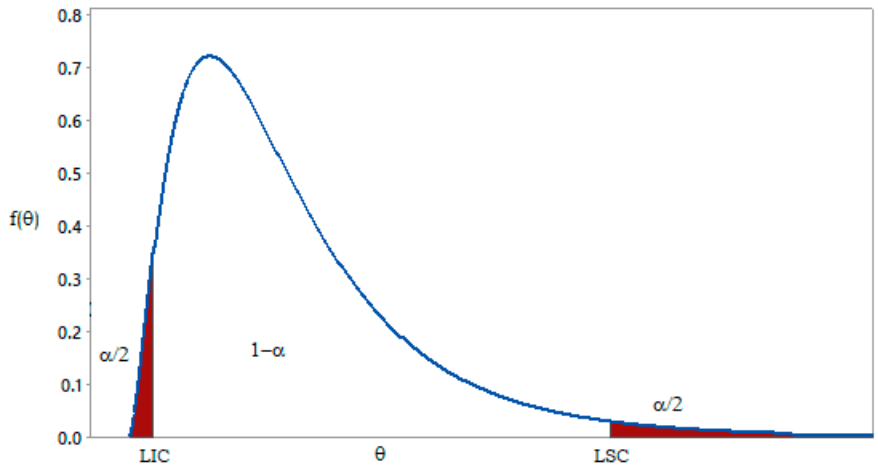
$$LIC \leq \theta \leq LSC \tag{4.32}$$

En donde los puntos extremos LIC y LSC de este intervalo, son variables aleatorias, ya que se van a obtener como estadísticos muestrales, es decir, funciones de datos de muestra. Para determinar el intervalo del parámetro desconocido q, se deben obtener los estadísticos LIC y LSC, que cumplan que la probabilidad de que q caiga entre esos puntos es (1-α), es decir,

$$p(LIC \leq \theta \leq LSC) = 1-\alpha \tag{4.33}$$

La representación gráfica de la expresión anterior se muestra en la figura 4.7 y se explica a continuación:

Figura 4.7. Representación gráfica de un intervalo de confianza



A $\text{LIC} \leq \theta \leq \text{LSC}$ se le denomina Intervalo de Confianza al $100(1-\alpha)\%$ del nivel de confianza del parámetro desconocido θ. LIC es el Límite Inferior del Intervalo de Confianza, LSC es el Límite Superior del Intervalo de Confianza. $(1-\alpha)$ es el coeficiente de confianza. Suponga que la función de densidad de probabilidad de q es $f(\theta)$ y $1-\alpha$ representa al área bajo la curva $f(\theta)$ en el intervalo de LIC a LSC.

La estimación por intervalos consiste en establecer el intervalo de valores donde es más probable que se encuentre el parámetro θ. La obtención del intervalo se basa en las siguientes consideraciones:

a. Si se conoce la distribución muestral del estimador se pueden obtener las probabilidades de ocurrencia de los estadísticos muestrales.

b. Si se conoce el valor del parámetro poblacional, se puede establecer la probabilidad de que el estimador se encuentre dentro de los intervalos de la distribución muestral.

c. El problema es que el parámetro poblacional q es desconocido, y por ello el intervalo se establece alrededor del estimador puntual. Si se repite el muestreo un gran número de veces y se define un intervalo alrededor de cada valor del estadístico muestral, el parámetro se sitúa dentro de cada intervalo en un porcentaje conocido de ocasiones. Este intervalo es el denominado intervalo de confianza.

El intervalo de confianza de la expresión 4.32 se denomina un intervalo de confianza bilateral o de dos lados, ya que especifica tanto un límite inferior como un límite superior para θ. En algunas aplicaciones prácticas un intervalo de confianza unilateral o de un solo lado podría ser más apropiado.

Por ejemplo, cuando se habla de la resistencia θ a la tensión de un cable o malacate, dicho cable debe tener una resistencia mínima, por lo cual presenta un solo límite de confianza inferior

$$\text{LIC} \leq \theta \tag{4.34}$$

donde el límite de confianza inferior LIC se elige de modo que

$$p\left(\text{LIC} \leq \theta\right) = 1-\alpha \tag{4.35}$$

En otro ejemplo, cuando se habla del tiempo de espera q, dicho tiempo de espera debe tener un tiempo máximo, por lo cual presenta un solo límite de confianza superior

$$\theta \leq \text{LSC} \tag{4.36}$$

Donde el límite superior LSC se elige de manera que

$$P(\theta \leq \text{LSC}) = 1-\alpha \tag{4.37}$$

Se denomina Imprecisión del Estimador θ (una parte considerable de los expertos en estadística le denomina precisión del estimador) a la longitud de medio intervalo de confianza θ-LIE o LSC-θ. Cuanto más grande es la imprecisión del estimador mayor es el nivel de confianza 100(1–α)%; sin embargo, mientras más grande sea la imprecisión del estimador, menos información se tiene acerca del valor verdadero de θ. Esto implica que existe una relación funcional directa entre la imprecisión del estimador y el valor de α. Lo ideal sería contar con estimadores más precisos (es decir, con una semilongitud del intervalo baja y con un nivel de confianza elevado).

En términos prácticos, la imprecisión del estimador se acostumbra fijar como un error de estimación ε en unidades de desviación estándar del estimador, es decir, si el parámetro θ tiene como estimador puntual a $\hat{\theta}$ con media μ_θ y desviación estándar σ_θ

$$\varepsilon = k\hat{\sigma} \tag{4.38}$$

Los límites de confianza son

$$LIC = \mu_{\hat{\theta}} - \varepsilon = \mu_{\hat{\theta}} - k\hat{\sigma}_\theta \tag{4.39}$$

$$LSC = \mu_{\hat{\theta}} + \varepsilon = \mu_{\hat{\theta}} + k\hat{\sigma}_\theta$$

En donde el valor que toma k depende de la función de probabilidad del estimador $f(\hat{\theta})$ y del nivel de confianza α que se fije (observe la figura 4.7).

4.6.1. Intervalo de Confianza para el número de elementos exitosos en una muestra de tamaño n o para la fracción o proporción de elementos exitosos *p* en una población

Se le llama obtener un elemento exitoso a extraer un elemento o artículo con cierta característica de interés. Por ejemplo, en una caja con n pelotas con la misma forma y el mismo diámetro, se tienen D pelotas rojas y $N–D$ pelotas negras. Generalmente se le llama éxito a extraer el elemento más raro o menos probable de los dos que existen; en este caso, si hay más pelotas negras que pelotas rojas, se le denomina éxito a extraer una pelota roja. Otra forma de verlo es que en un lote con n artículos, D son defectuosos y $N–D$ son no defectuosos; se le llama éxito a extraer un artículo defectuoso. Otro ejemplo, en un juego de baraja española donde existe un mazo de $N = 40$ cartas, de las cuales $D = 4$ son ases, se le llama éxito al extraer una carta y que esta sea as.

En el volumen III de Fundamentos de Probabilidad se definió a una variable hipergeométrica, donde x representa el número de artículos con cierta característica de interés, en una muestra de tamaño n obtenida aleatoriamente de una población de tamaño N, donde existen D elementos con esa característica de interés y cuya función de probabilidad es

$$p(x; n, D, N) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, \cdots, \textit{minimo}\ (n, D)$$

En donde su media y su desviación estándar están dadas por las expresiones

$$\mu_x = n\left(\frac{D}{N}\right) = np$$

$$\sigma_x^2 = n\left(\frac{D}{N}\right)\left(1 - \frac{D}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right)$$

También se dijo que para $np > 5$ y $p < 0.5$, esta función hipergeométrica se comporta como una normal.

$$x \sim N\left[np, \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}\right]$$

Nótese que la variable x que se puede suponer normal, si $np > 5$ para $p < 0.5$, es decir, si p es muy pequeña n debe ser muy grande para compensar; por ejemplo: si $p = 0.01$, $n > 500$; si $p < 0.001$, $n > 5000$, y así sucesivamente.

En tal caso, x se puede estandarizar, con el siguiente cambio de variable:

$$z = \frac{x - np}{\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}}$$

Lo cual implica que

$$-z_{\alpha/2} \leq \frac{x - np}{\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}} \leq z_{\alpha/2}$$

Despejando x, se obtiene un intervalo bilateral de confianza al $(1-\alpha)100\%$ de nivel de confianza del número de elementos exitosos en una muestra de tamaño n obtenida de una población de tamaño N, donde existen D elementos exitosos

$$np - z_{\alpha/2}\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)} \leq x \leq np + z_{\alpha/2}\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)} \tag{4.40}$$

De la misma forma, los intervalos unilaterales inferior y superior están dados por las expresiones

$$x \leq np + z_{\alpha}\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)} \tag{4.41}$$

$$np - z_{\alpha}\sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)} \leq x \tag{4.42}$$

Al factor

$$\left(\frac{N-n}{N-1}\right) \tag{4.43}$$

Se le denomina Factor de Corrección y se utiliza cuando el tamaño de la población N es finito. Note que

$$Lim_{N\to\infty}\left(\frac{N-n}{N-1}\right)=1$$

Por lo que si el tamaño de la población es muy grande, se puede considerar infinito y en tal caso las expresiones anteriores 4.40, 4.41 y 4.42 quedan de la siguiente forma:

$$np-z_{\alpha/2}\sqrt{np(1-p)}\le x\le np+z_{\alpha/2}\sqrt{np(1-p)} \tag{4.44}$$

$$x\le np+z_{\alpha}\sqrt{np(1-p)} \tag{4.45}$$

$$np-z_{\alpha}\sqrt{np(1-p)}\le x \tag{4.46}$$

En las expresiones 4.40, 4.41, 4.42, 4.44, 4.45 y 4.46 anteriores, si se dividen entre n, haciendo $\hat{p}=\frac{x}{n}$ e invirtiendo p con $\hat{p}=\frac{x}{n}$, se obtienen los intervalos bilateral, inferior y superior de confianza al 100(1–α)% de nivel de confianza para la fracción de elementos exitosos, es decir, con cierta característica de interés (defectuosos por ejemplo) en la población:

Para población finita

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}\le p\le\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.47}$$

$$p\le\hat{p}+z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.48}$$

$$\hat{p}-z_{\alpha}\sqrt{\frac{p(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}\le p \tag{4.49}$$

Para población infinita:

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\le p\le\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.50}$$

$$p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.51}$$

$$\hat{p} - z_{\alpha}\sqrt{\frac{p(1-\hat{p})}{n}} \leq p \tag{4.52}$$

El error de estimación o la imprecisión del estimador, en la expresión 4.47, para el intervalo bilateral de confianza está dado por la expresión

$$\varepsilon \leq z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}$$

Despejando n de esta última expresión se obtiene el tamaño de muestra adecuado para un intervalo bilateral de confianza al $(1-\alpha)100\%$ de nivel de confianza para un error de estimación ε dado

$$n \geq \frac{N\hat{p}(1-\hat{p})}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2(N-1)+\hat{p}(1-\hat{p})} \tag{4.53}$$

De la misma forma, para los intervalos unilaterales, n está dada por la expresión

$$n \geq \frac{N\hat{p}(1-\hat{p})}{\left(\frac{\varepsilon}{z_{\alpha}}\right)^2(N-1)+\hat{p}(1-\hat{p})} \tag{4.54}$$

Nótese que solo cambia en el valor de z a elegir.

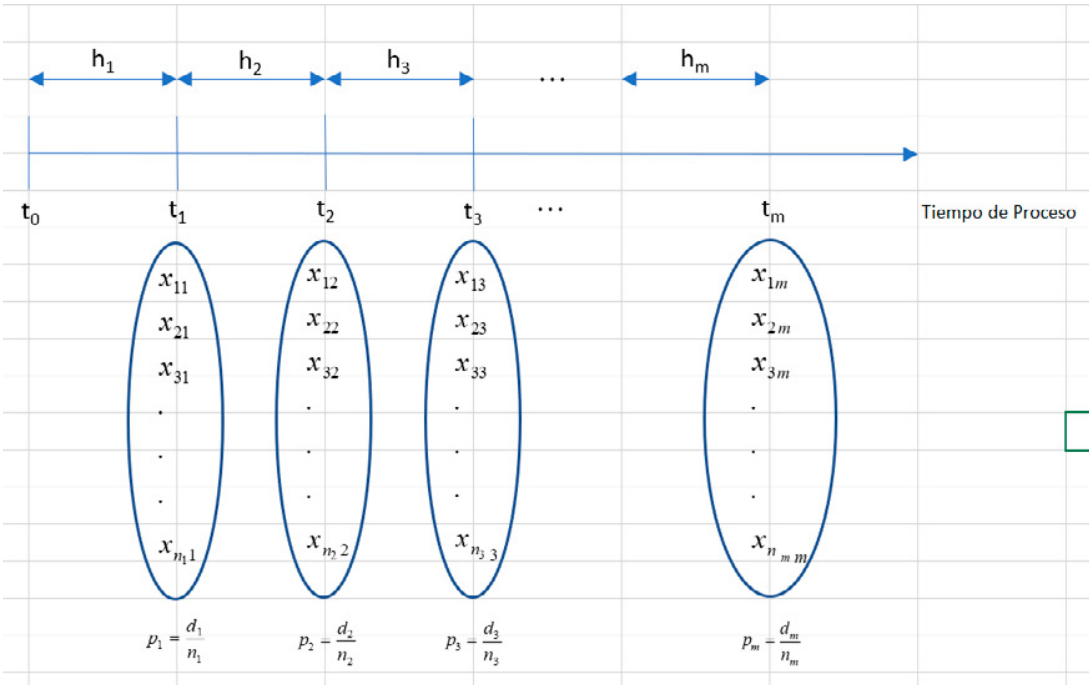
Si el tamaño de la población es infinito, la expresión 4.53 se convierte en

$$n \geq \frac{\hat{p}(1-\hat{p})}{\left(\frac{\varepsilon}{z_{\alpha/2}}\right)^2} \tag{4.55}$$

El estimador puntual $\hat{p}$ puede ser obtenido a partir de una muestra grande de elementos obtenidos aleatoriamente de la población, dividiendo el número de elementos que presenta la característica de interés x entre el número de elementos en la muestra n, es decir, $\hat{p} = {}^{x}/_{n}$.

Si se desea obtener un estimador por intervalos para p más preciso, el estimador puntual $\hat{p}$ puede ser estimado como la media de las fracciones exitosas obtenidas de un conjunto de muestras periódicas tomadas aleatoriamente de la población, como se muestra en la figura 4.8

Figura 4.8



Para lo cual se utiliza el concepto de media armónica para los pj, o en su defecto, usar la siguiente expresión:

$$\overline{p} = \frac{\sum_{j=1}^{m} d_j}{\sum_{j=1}^{m} n_j} \tag{4.56}$$

Ejercicio 4.16

Mediante estudios recientes se ha determinado que la probabilidad de morir por causa de cierta vacuna contra la gripe es de 0.00002. Obtenga los intervalos de confianza bilateral, superior e inferior al 95% de nivel de confianza, para la fracción de personas que pueden fallecer por causa de esta vacuna, si se administra a $n = 10000$ personas.

Primero deben obtenerse los valores de z al 95% de nivel de confianza, utilizando R:

$$z_{\frac{\alpha}{2}} = qnorm\left(0.025, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.959964$$

$$z_{\alpha} = qnorm\left(0.05, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.644854$$

Los intervalos de confianza al 95% de nivel de confianza serían:

$0 \leq p \leq 0.00011$	95%_ *nivel_de_confianza*
$p \geq 0$	95%_*nivel_de_confianza*
$p \leq 0.000094$	95%_*nivel_de_confianza*

En los cálculos anteriores el límite inferior del intervalo resultó ser negativo, pero dado que se trata de fracciones defectuosas, no puede ser negativo y se fija en su cota menor: cero.

Suponga que un ciudadano alarmista, de los que nunca faltan en nuestra sociedad, declara que el número de fallecimientos por la administración de la vacuna es de 5 por cada 10000 habitantes, ¿tiene usted argumentos para refutar su afirmación?, ¿qué le diría?

El límite superior de fallecimientos por la administración de la vacuna es de 9.4 por cada cien mil habitantes, con un nivel de confianza del 95%, de tal manera que el número de fallecimientos no llega a 10 por cada cien mil o a uno por cada diez mil, como se pudo apreciar anteriormente, lo cual desmiente tal afirmación con un nivel de certidumbre de 95%.

¿De qué tamaño tendría que ser la muestra para garantizar que el límite inferior del intervalo bilateral al 95% de nivel de confianza para p fuera mayor de cero?

Se desea lograr que

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{p\left(1-\hat{p}\right)}{n}} \geq 0$$

Al despejar n:

$$n \geq z^2{}_{\alpha/2}\left(\frac{1-\hat{p}}{\hat{p}}\right)$$

Sustituyendo valores $n > 192069$ el tamaño de muestra tendría que ser muy grande. Esto se debe a las dimensiones de la proporción exitosa que es de dos cienmilésimas.

Ejercicio 4.17

Un fabricante de componentes de audio para computadora tiene un proceso productivo al que le recopila datos tomados de la prueba final a que se somete el producto en el mes de abril, los cuales se muestran a continuación. Obtenga intervalos de confianza bilateral, superior e inferior al 95% del nivel de confianza para la fracción de artículos defectuosos que se envían a un cliente en lotes de tamaño $n = 2000$.

Día	Cantidad inspeccionada	Artículos defectuosos
1	2,450	42
2	1,997	39
5	2,168	52
6	1,941	47
7	1,962	34
8	2,244	29
9	1,238	53
12	2,289	45
13	1,464	26
14	2,061	47
15	1,667	34
16	2,350	31
19	2,354	38
20	1,509	28
21	2,190	30
22	2,678	113
23	2,252	58
26	1,641	34
27	1,782	19
28	1,993	30
29	2,382	17
30	2,132	46
Suma =	44,744	892

La estimación puntual de la fracción defectuosa sería la media de las fracciones defectuosas de cada una de las muestras tomadas, por lo que

$$\hat{p} = \frac{892}{44744} = 0.019936 = 1.9936\%$$

Primero deben obtenerse los valores de z al 95% de nivel de confianza, utilizando R:

$$z_{\frac{\alpha}{2}} = qnorm\left(0.025, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.959964$$

$$z_{\alpha} = qnorm\left(0.05, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.644854$$

El intervalo bilateral de confianza al 95% de nivel de confianza para p, en lotes de tamaño n = 2000, es:

$0.01381 \leq p \leq 0.026062$

El intervalo superior de confianza al 95% de nivel de confianza para p, en lotes de tamaño ^{n} = 2000, es:

$0.01479 \leq p$

El intervalo inferior de confianza al 95% de nivel de confianza para p, en lotes de tamaño n = 2000, es:

$p \leq 0.025077$

Suponga que el cliente que recibe los lotes de 2000 piezas afirma que se le está enviando un porcentaje defectuoso mayor a 3%, ¿tiene usted argumentos para rebatirle al cliente tal afirmación?

De acuerdo con los intervalos de confianza, en particular del intervalo inferior de confianza, el máximo porcentaje defectuoso que se le ha enviado al cliente es de 2.5%, con un grado de certeza del 95%, de tal manera que no llega a 3%.

Suponga que el dueño de la empresa declara que fabrica componentes para audio con menos del 1% de defectuosos, ¿puede usted apuntalar o sostener estadísticamente la afirmación del jefe?

De acuerdo con los intervalos de confianza, en particular del intervalo superior de confianza, el mínimo porcentaje defectuoso que se fabrica es de 1.48%, con un grado de certeza del 95%, de tal manera que no existe evidencia estadística para sostener la afirmación del dueño de la compañía.

Si el gerente de planta de la empresa declara que el porcentaje defectuoso es de alrededor del 2% de defectuosos, de acuerdo con el intervalo bilateral de confianza al 95% de nivel de confianza, no existe evidencia estadística para refutar o negar su afirmación, ya que el 2% cae dentro de dicho intervalo. Si cayera fuera se tendría evidencia estadística para refutar su afirmación.

Ejercicio 4.18

Se requiere estimar la fracción de tuercas defectuosas en un lote de N = 5000 piezas. Para ello, se toma una muestra de *n* piezas, las cuales se inspeccionan en cuanto a la forma de la tuerca (debe ser hexagonal, el diámetro (debe estar entre 0.490 y 0.510 pulgadas), el tipo de rosca (debe ser milimétrica), y las condiciones de la rosca (no debe estar barrida). Si la tuerca no cumple con una o más de estas características se clasifica como defectuosa; si cumple todas las características se clasifica como no defectuosa. Suponga que después de hacer la inspección se obtuvo una fracción defectuosa de 4%.

a. Si n = 500, obtenga un intervalo de confianza al 95% de nivel de confianza.

Primero deben obtenerse los valores de z al 95% de nivel de confianza, utilizando R:

$$z_{\frac{\alpha}{2}} = qnorm\left(0.025, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.959964$$

$$z_{\alpha} = qnorm\left(0.05, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.644854$$

No debe perderse de vista que en este caso el tamaño de la población es finito, N = 5000, por lo que debe considerarse el factor de corrección.

$$\left(\frac{N-n}{N-1}\right) = \frac{4500}{4999} = 0.90018$$

El intervalo bilateral de confianza al 95% de nivel de confianza para p, en una muestra de tamaño n = 500, es:

$$0.03593 \leq p \leq 0.044074$$

El intervalo superior de confianza al 95% de nivel de confianza para p, en una muestra de tamaño $n = 500$, es:

$0.03658 \leq p$

El intervalo inferior de confianza al 95% de nivel de confianza para p, en una muestra de tamaño $n = 500$, es:

$p \leq 0.043419$

b. ¿De qué tamaño debe ser la muestra para obtener un error de estimación menor a 0.003 en el intervalo bilateral de confianza?

Sustituyendo en la expresión 4.54, $n > 3831$.

4.6.2. Intervalo de confianza para el número de defectos, ocurrencias, éxitos o llegadas en *n* unidades, así como la fracción de defectos, ocurrencias, éxitos o llegadas por unidad

En el volumen III de Fundamentos de Probabilidad se definió a una variable tipo Poisson, donde x representa el número de defectos, ocurrencias, éxitos o llegadas, en n unidades. Sea c el número promedio de defectos, ocurrencias, éxitos o llegadas en promedio en n unidades. La función de probabilidad de x está definida por la expresión:

$$p(x;c) = e^{-c}\left(\frac{c^x}{x!}\right) \qquad x = 0, 1, 2, \cdots$$

Su media y su desviación estándar están dadas por las expresiones

$\mu_x = c$

$\sigma_x^2 = c$

También se demostró en el volumen III que dicha función de probabilidad de Poisson se deduce como una forma límite de la distribución binomial y que existe una aproximación de la Poisson a través de la normal cuando $c > 5$.

$$x \sim N\left[c, \sqrt{c}\right]$$

En tal caso, x se puede estandarizar, con el siguiente cambio de variable:

$$z = \frac{x - c}{\sqrt{c}}$$

Lo cual implica que

$$-z_{\alpha/2} \leq \frac{x - c}{\sqrt{c}} \leq z_{\alpha/2}$$

Despejando x

$$c - z_{\alpha/2}\sqrt{c} \leq x \leq c + z_{\alpha/2}\sqrt{c} \tag{4.57}$$

Se obtiene el intervalo bilateral de confianza al 100(1–α)% de nivel de confianza para el número de defectos, llegadas, éxitos u ocurrencias en n unidades:

$$\hat{c} - z_{\alpha/2}\sqrt{\hat{c}} \leq x \leq \hat{c} + z_{\alpha/2}\sqrt{\hat{c}} \tag{4.58}$$

El intervalo superior de confianza para c al 100(1–α)% de nivel de confianza queda como:

$$\hat{c} - z_{\alpha}\sqrt{\hat{c}} \leq c \tag{4.59}$$

El intervalo inferior de confianza para c al 100(1–α)% de nivel de confianza es:

$$c \leq \hat{c} + z_{\alpha}\sqrt{\hat{c}} \tag{4.60}$$

Si el tamaño de la población es finito N, los intervalos bilateral, inferior y superior de confianza al (1–α)100% de nivel de confianza para c están dados por las expresiones

$$\hat{c} - z_{\alpha/2}\sqrt{\hat{c}\left(\frac{N-n}{N-1}\right)} \leq c \leq \hat{c} + z_{\alpha/2}\sqrt{\hat{c}\left(\frac{N-n}{N-1}\right)} \tag{4.61}$$

$$\hat{c} - z_{\alpha}\sqrt{\hat{c}\left(\frac{N-n}{N-1}\right)} \leq c \tag{4.62}$$

$$c \leq \hat{c} + z_{\alpha}\sqrt{\hat{c}\left(\frac{N-n}{N-1}\right)} \tag{4.63}$$

Ejercicio 4.19

El número de clientes que llega a un banco es una variable aleatoria de Poisson. Suponga que por datos históricos se ha logrado determinar que el número de llegadas promedio es de 120 por hora. Obtenga intervalos de confianza bilateral, superior e inferior para el número de clientes que llegan al banco por hora al 95% de nivel de confianza.

Primero deben obtenerse los valores de z al 95% de nivel de confianza, utilizando R:

$$z_{\frac{\alpha}{2}} = qnorm\left(0.025, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.959964$$

$$z_{\alpha} = qnorm\left(0.05, mean = 0, sd = 1, lower.tail = FALSE, log.p = FALSE\right) = 1.644854$$

El intervalo bilateral de confianza al 95% de nivel de confianza para el número de clientes que llegan al banco, es:

$98.53 \leq c \leq 141.47$

El intervalo superior de confianza al 95% de nivel de confianza para c, es:

$101.98 \leq c$

El intervalo inferior de confianza al 95% de nivel de confianza para c, es:

$c \leq 138.02$

Suponga que el gerente de la sucursal afirma que llegan al banco más de 150 clientes por hora, ¿tiene usted evidencia estadística para sostener tal afirmación?

Como se puede apreciar, el máximo número de personas que llega al banco, según el intervalo inferior de confianza, es de 138 al 95% de nivel de confianza, de tal manera que la afirmación del gerente no es tan verídica y no se podría sostener.

Sea

$$u = \frac{c}{n} \tag{4.64}$$

La fracción de defectos, ocurrencias, llegadas o éxitos por unidad en promedio, entonces, dividiendo entre *n* a las expresiones 4.51, 4.52 y 4.53, se obtienen los intervalos bilateral, inferior y superior de confianza para la fracción de defectos, ocurrencias, llegadas o éxitos por unidad en una población de tamaño infinito:

$$\hat{u} - z_{\alpha/2}\sqrt{\frac{\hat{u}}{n}} \leq u \leq \hat{u} + z_{\alpha/2}\sqrt{\frac{\hat{u}}{n}} \tag{4.65}$$

$$u \leq \hat{u} + z_{\alpha}\sqrt{\frac{\hat{u}}{n}} \tag{4.66}$$

$$\hat{u} - z_{\alpha}\sqrt{\frac{\hat{u}}{n}} \leq u \tag{4.67}$$

Los intervalos bilateral, inferior y superior de confianza para la fracción de defectos, ocurrencias, llegadas o éxitos por unidad en una población de tamaño finito son:

$$\hat{u} - z_{\alpha/2}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)} \leq u \leq \hat{u} + z_{\alpha/2}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.68}$$

$$u \leq \hat{u} + z_{\alpha}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.69}$$

$$\hat{u} - z_{\alpha}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)} \leq u \tag{4.70}$$

El error de estimación o imprecisión del intervalo bilateral de confianza al $(1-\alpha)100\%$ de nivel de confianza, de acuerdo con la expresión 4.68, está dado por

$$\varepsilon \leq z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)}$$

Despejando n, se obtiene el tamaño de muestra adecuado para calcular un intervalo de confianza al $(1-\alpha)$ 100% de nivel de confianza de la fracción de defectos por unidad en una población finita

$$n \geq \frac{N\hat{u}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2 (N-1) + \hat{u}} \tag{4.71}$$

Si el tamaño de la población es infinito, la expresión anterior se convierte en

$$n \geq \frac{\hat{u}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2} \tag{4.72}$$

En las expresiones anteriores el estimador de u puede también obtenerse como una media de los valores estimados de c o de u de un conjunto de muestras obtenidas históricamente, es decir,

$$\overline{u} = \frac{\sum_{j=1}^{m} c_j}{\sum_{j=1}^{m} n_j} \tag{4.73}$$

Ejercicio 4.20

Se requiere estimar un intervalo de confianza al 95% de nivel de confianza, para el número de defectos por unidad que presenta una flotilla de $n = 200$ automóviles en una agencia que los renta, suponga que el número de autos que posee el corporativo es infinito. Por datos históricos se ha logrado contabilizar el número de defectos de los automóviles que se envían al taller en los últimos diez meses, como se muestra a continuación:

mes	No. Automóviles	No. Defectos
1	15	87
2	12	93
3	10	112
4	13	115
5	12	120
6	11	93
7	20	130
8	12	105
9	13	100
10	13	99
Suma =	131	1054

Calcule intervalos de confianza para el número de defectos por unidad de los 200 automóviles de esta compañía.

Primero deben obtenerse los valores de z al 95% de nivel de confianza, utilizando R:

$z_{\alpha/2} = qnorm(0.025, mean = 0, \text{sd} = 1, lower.tail = FALSE, log.p = FALSE) = 1.959964$

$z_{\alpha} = qnorm(0.05, mean = 0, \text{sd} = 1, lower.tail = FALSE, log.p = FALSE) = 1.644854$

Para este ejemplo, de la ecuación 4.73 y de los datos del cuadro anterior

$$u = \frac{\sum_{i=1}^{i=m} c_i}{\sum_{i=1}^{i=m} n_i} = \frac{1054}{131} = 8.0458$$

Por lo que el intervalo de confianza bilateral al 95% de nivel de confianza para el número de defectos por unidad u, con la expresión 4.65 es

$7.65 \leq u \leq 8.44$

El intervalo de confianza superior al 95% de nivel de confianza para el número de defectos por unidad u, con la expresión 4.66 es

$7.72 \leq u$

El intervalo de confianza inferior al 95% de nivel de confianza para el número de defectos por unidad u, con la expresión 4.67 es

$u \leq 8.38$

Suponga que el dueño de la empresa arrendadora de autos sostiene que sus automóviles no llegan a cinco defectos por unidad, ¿qué le diría?

Como se puede apreciar, el número de defectos por unidad, de conformidad con el intervalo superior de confianza, no es menor a siete, por lo cual, no es cierto, al 95% de nivel de confianza, que los automóviles tengan menos de cinco defectos por unidad.

4.6.3. Intervalos de confianza para la media poblacional con varianza poblacional conocida

En el volumen III de Fundamentos de Probabilidad se demostró que la media muestral presenta distribución normal de media la media poblacional y de varianza la varianza poblacional entre n, según la expresión definida en 1.64, es decir,

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

Esto se cumple bajo una de dos hipótesis iniciales:

i. La distribución de probabilidad de x es normal, $x \sim N(\mu_x, \sigma_x)$.

ii. El tamaño de la muestra es tan grande que puede considerarse infinito.

En este caso entonces, el intervalo de confianza de la media muestral $\bar{x}$ sería:

$$\mu_x - k\frac{\sigma_x}{\sqrt{n}} \leq \bar{x} \leq \mu_x + k\frac{\sigma_x}{\sqrt{n}} \tag{4.74}$$

Por otra parte, dado que presenta distribución normal, se puede estandarizar:

$$z = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

Lo que implica que

$$p\left(\mu_x - k\frac{\sigma_x}{\sqrt{n}} \leq \bar{x} \leq \mu_x + k\frac{\sigma_x}{\sqrt{n}}\right) = p\left(-z_{\frac{\alpha}{2}} \leq z \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Y conduce a deducir que para el caso de la media muestral $k = z_{\alpha/2}$

De esta forma, el intervalo de confianza para la media muestral está dado por la expresión

$$\mu_x - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \leq \bar{x} \leq \mu_x - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \tag{4.75}$$

Si se despeja mx en la expresión anterior, se obtiene un intervalo de confianza para la media poblacional al 100(1–α)% del nivel de confianza, en términos de la media de una muestra y de la varianza poblacional, la cual debe ser conocida.

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \tag{4.76}$$

Para el caso del intervalo unilateral inferior de confianza al 100(1–α)% del nivel de confianza:

$$\mu_x \leq \bar{x} + z_{\alpha} \frac{\sigma_x}{\sqrt{n}} \qquad \textit{Intervalo\_Inferior\_de\_Confianza} \tag{4.77}$$

Para el caso del intervalo unilateral superior de confianza al 100(1–α)% del nivel de confianza:

$$\bar{x} - z_{\alpha} \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \qquad \textit{Intervalo\_Superior\_de\_Confianza} \tag{4.78}$$

Si el tamaño de la población es finito, el intervalo bilateral de confianza está dado por la expresión

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu_x \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{4.79}$$

Si el tamaño de la población es finito, el intervalo unilateral inferior de confianza al 100(1–α)% del nivel de confianza es:

$$\mu_x \leq \bar{x} + z_{\alpha} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad \textit{Intervalo\_Inferior\_de\_Confianza} \tag{4.80}$$

Si el tamaño de la población es finito, el intervalo unilateral superior de confianza al 100(1–α)% del nivel de confianza es:

$$\bar{x} - z_{\alpha} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu_x \qquad \textit{Intervalo\_Superior\_de\_Confianza} \tag{4.81}$$

El error de estimación o imprecisión para un intervalo bilateral de confianza para la media poblacional conocida la varianza poblacional, de acuerdo con la expresión 4.79, está definido como

$$\varepsilon = z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Por lo que, despejando n

$$n \geq \frac{N\sigma_x^2}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2 (N-1) + \sigma_x^2} \tag{4.82}$$

Para N infinito

$$n \geq \frac{\sigma_x^2}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2} \tag{4.83}$$

Para un intervalo unilateral de confianza al 100(1–α)% de nivel de confianza, para *n* finito

$$n \geq \frac{N\sigma_x^2}{\left(\frac{\varepsilon}{z_{\alpha}}\right)^2 (N-1) + \sigma_x^2} \tag{4.84}$$

Para un intervalo unilateral de confianza al 100(1–α)% de nivel de confianza, para *n* infinito

$$n \geq \frac{\sigma_x^2}{\left(\frac{\varepsilon}{z_{\alpha}}\right)^2} \tag{4.85}$$

Ejercicio 4.21 En el ramo de la construcción, una característica de calidad importante es la resistencia del concreto a la compresión, la cual se mide fracturando probetas cilíndricas de concreto en una máquina de ensayos de compresión. Asimismo, la

resistencia a la compresión se calcula a partir de la carga de ruptura, dividida por el área de la sección que resiste a la carga en unidades de libra-fuerza por pulgada cuadrada (psi) en el Sistema Británico o en MegaPascales (MPa) en el Sistema Internacional de Unidades. Un ingeniero civil realiza la prueba a 50 probetas, obteniendo los siguientes datos en unidades de psi:

2243	2310	2281	2277	2272
2246	2271	2235	2261	2251
2320	2208	2268	2263	2295
2215	2270	2264	2241	2205
2234	2285	2256	2305	2223
2271	2244	2306	2281	2242
2287	2254	2212	2252	2258
2267	2268	2279	2304	2240
2257	2230	2263	2297	2270
2263	2290	2219	2224	2262

Suponga que la resistencia a la compresión de dichas probetas es normal y que se conoce su desviación estándar $\sigma_x = 30$ psi.

a. Construya un intervalo de confianza bilateral al 90, 95 y 99% de nivel de confianza de la resistencia promedio del concreto a la compresión.

b. Construya un intervalo inferior de confianza al 95% de la resistencia promedio del concreto a la compresión.

c. ¿De qué tamaño debe ser la muestra para suponer que el error de estimación sea menor a 7 psi, para el intervalo bilateral al 95% de nivel de confianza?

Se calcula la media aritmética de la muestra dada, usando excel:

$\overline{x}$ = promedio($A$1:$E$10) = 2260.78 psi

Se obtienen los valores de z para 90, 95 y 99% de nivel de confianza:

$$z_{0.05} = INV.NORM.ESTAND(0.05) = -1.644854; \quad z_{0.95} = INV.NORM.ESTAND(0.95) = 1.644854$$
$$z_{0.025} = INV.NORM.ESTAND(0.025) = -1.959964; \quad z_{0.975} = INV.NORM.ESTAND(0.975) = 1.959964$$
$$z_{0.005} = INV.NORM.ESTAND(0.005) = -2.575829; \quad z_{0.995} = INV.NORM.ESTAND(0.995) = 2.575829$$

Como se puede apreciar, si se teclea el valor de $\alpha/2$ se obtiene el valor negativo y si se teclea el valor de $1-\alpha/2$ se obtiene el positivo, esto se debe a la simetría de la distribución normal.

Se calculan los intervalos de confianza solicitados, usando la expresión 4.76:

$2253.8 \leq \mu_x \leq 2267.8$ *Al_90%_de_nivel_de_confianza*
$2252.5 \leq \mu_x \leq 2269.1$ *Al_95%_de_nivel_de_confianza*
$2249.9 \leq \mu_x \leq 2271.7$ *Al_99%_de_nivel_de_confianza*

Nótese que mientras el nivel de confianza es mayor, el intervalo de confianza se abre ofreciendo menos precisión.

Para el intervalo superior de confianza, al 95% de nivel de confianza:

$$z_{0.05} = INV.NORM.ESTAND\left(0.05\right) = -1.644854$$

El intervalo superior de confianza al 95% es:

$2253.8 \leq \mu_x$ *Al_95%_de_nivel_de_confianza*

El intervalo inferior de confianza al 95% es:

$\mu_x \leq 2267.76$ *Al_95%_de_nivel_de_confianza*

Sustituyendo en la expresión 4.83, para un intervalo bilateral de confianza, al 95% de nivel de confianza para una población de tamaño infinito, la muestra debe tener un tamaño mayor a 70.

$$n \geq \left(\frac{1.96 * 30}{7}\right)^2 = 70.56$$

4.6.4. Intervalos de confianza para la varianza poblacional de una población con distribución de probabilidad normal o para un tamaño de muestra grande

En el volumen III de Fundamentos de Probabilidad se definió a una variable ji cuadrada como la suma de k variables aleatorias normales estándar elevadas al cuadrado (ver expresión 1.65); con función de densidad dada por la expresión 1.66 y media y varianza dadas por las expresiones 1.69 y 1.70 respectivamente. También se remarcó que un caso especial y muy importante de variable aleatoria muestral con distribución t de Student está dado por la expresión 1.74, es decir,

$$\frac{(n-1)\,S^2{}_{n-1}}{\sigma_x^2} \sim \chi^2_{n-1}$$

Lo cual implica que, para un intervalo bilateral de confianza al 100(1–α)% de nivel de confianza:

$$\chi^2_{\frac{\alpha}{2},\,n-1} \leq \frac{(n-1)S^2_{n-1}}{\sigma_x^2} \leq \chi^2_{1-\frac{\alpha}{2},\,n-1}$$

Si se despeja la varianza poblacional de la expresión anterior, se obtiene un intervalo de confianza bilateral para la varianza al 100(1–α)% de nivel de confianza, es decir,

$$\frac{(n-1)S^2_{n-1}}{\chi^2_{1-\frac{\alpha}{2},\,n-1}} \leq \sigma_x^2 \leq \frac{(n-1)S^2_{n-1}}{\chi^2_{\frac{\alpha}{2},\,n-1}} \tag{4.86}$$

Para el caso del intervalo unilateral superior de confianza al 100(1–α)% del nivel de confianza para la varianza:

$$\frac{(n-1)S^2_{n-1}}{\chi^2_{1-\alpha,\,n-1}} \leq \sigma_x^2 \tag{4.87}$$

Para el caso del intervalo unilateral inferior de confianza al 100(1–α)% del nivel de confianza para la varianza:

$$\sigma_x^2 \leq \frac{(n-1)S^2_{n-1}}{\chi^2_{\alpha,\,n-1}} \tag{4.88}$$

Si la población fuera finita, los intervalos de confianza de las expresiones 4.86, 4.87 y 4.88 se convierten en

$$\frac{(n-1)S_{n-1}^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\left(\frac{N-n}{N-1}\right) \leq \sigma_x^2 \leq \frac{(n-1)S_{n-1}^2}{\chi^2_{\frac{\alpha}{2},n-1}}\left(\frac{N-n}{N-1}\right) \tag{4.89}$$

$$\sigma_x^2 \leq \frac{(n-1)S_{n-1}^2}{\chi^2_{\alpha,n-1}}\left(\frac{N-n}{N-1}\right) \tag{4.90}$$

$$\frac{(n-1)\,S_{n-1}^2}{\chi^2_{1-\alpha,n-1}}\left(\frac{N-n}{N-1}\right) \leq \sigma^2{}_x \tag{4.91}$$

Para el problema 4.16 calcule los intervalos de confianza bilateral, inferior y superior para la varianza poblacional al 95% de nivel de confianza, partiendo del hecho que no se conoce su valor.

Se calcula la varianza de la muestra dada, usando excel:

$S_{n-1}^2 = \text{VAR(\$A\$1:\$E\$10)} = 775.0731 \text{ psi}^2$

Se obtienen los valores de ji cuadrada para 95% de nivel de confianza:

$\chi^2_{\frac{\alpha}{2},n-1} = \text{INV.CHICUAD}\left(0.025,49\right) = 31.5549$

$\chi^2_{1-\frac{\alpha}{2},n-1} = \text{INV.CHICUAD}\left(0.975,49\right) = 70.2224$

$\chi^2_{\alpha,n-1} = \text{INV.CHICUAD}\left(0.05,49\right) = 33.9303$

$\chi^2_{1-\alpha,n-1} = \text{INV.CHICUAD}\left(0.95,49\right) = 66.3386$

El intervalo bilateral de confianza al 95% de nivel de confianza es

$540.833 \leq \sigma_x^2 \leq 1203.57$

El intervalo inferior de confianza al 95% de nivel de confianza es

$\sigma_x^2 \leq 1119.31$

El intervalo superior de confianza al 95% de nivel de confianza es

$$572.496 \leq \sigma_x^2$$

Cabe señalar que la distribución ji cuadrada es un caso especial de la distribución gamma. De hecho

$$\chi_k^2 \sim \Gamma\left(\frac{k}{2}, \theta = \frac{1}{2}\right) \tag{4.92}$$

Si k, el número de grados de libertad, es suficientemente grande, experimentalmente se puede comprobar que para $k > 100$, como consecuencia del límite central, puede aproximarse a través de una distribución normal:

$$\chi_k^2 \sim N\left(k, \sqrt{2k}\right) \tag{4.93}$$

Esto significa que se puede sustituir el valor de ji cuadrada por la expresión:

$$\frac{\chi^2_{\frac{\alpha}{2}, n-1}}{n-1} = 1 + z_{\frac{\alpha}{2}, n-1}\sqrt{\frac{2}{n-1}} \tag{4.94}$$

De tal manera que la expresión 4.89, para $n > 100$, se puede escribir de la siguiente forma:

$$\frac{S^2_{n-1}}{1 + z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-1}}} \leq \sigma_x^2 \leq \frac{S^2_{n-1}}{1 - z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-1}}} \tag{4.95}$$

4.6.5. Intervalo de confianza para la media poblacional de una población con media y varianza desconocida

En el volumen III de Fundamentos de Probabilidad se definió a una variable t de Student como el cociente entre una normal estándar y una variable aleatoria ji cuadrada (ver expresión 1.75); con función de densidad dada por la expresión 1.76 y media y varianza dadas por las expresiones 1.77 y 1.78 respectivamente. También se remarcó que un caso especial y muy importante de variable aleatoria muestral con distribución t de Student está dado por la expresión 1.81, es decir,

$$\frac{\bar{x}-\mu_x}{\frac{S_{n-1}}{\sqrt{n}}} \sim t_{n-1}$$

Lo cual implica que

$$-t_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu_x}{\frac{S_{n-1}}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}$$

Si se despeja μ_x de esta expresión, se obtiene un intervalo bilateral de confianza para la media poblacional en términos de los estimadores puntuales de media y desviación estándar:

$$\bar{x}-t_{\frac{\alpha}{2},n-1}\frac{S_{n-1}}{\sqrt{n}} \leq \mu_x \leq \bar{x}+t_{\frac{\alpha}{2},n-1}\frac{S_{n-1}}{\sqrt{n}} \tag{4.96}$$

Para el problema 4.16 calcule los intervalos de confianza bilateral, superior e inferior para la media poblacional al 95% de nivel de confianza, partiendo del hecho que no se conoce su varianza poblacional.

Se calcula la media y la desviación estándar de la muestra, usando excel:

$\bar{x}$ = promedio($A$1:$E$10) = 2260.78 psi

S_{n-1} = DESVEST($A$1:$E$10) = 27.8401 psi

Nótese que la desviación estándar muestral es menor que la supuestamente conocida de 30 psi, lo cual de entrada va a reducir la amplitud del intervalo de confianza.

Se obtiene el valor de t para 95% de nivel de confianza con 49 grados de libertad, usando Excel,

$$t_{\frac{\alpha}{2},n-1} = \text{INV.T}\left(0.975,49\right) = 2.009575$$

$$t_{\alpha,n-1} = \text{INV.T}\left(0.95,49\right) = 1.67655$$

Observe que en el caso bilateral el valor de t = 2.009575 es mayor que el valor de z = 1.959964, lo cual abre el intervalo bilateral de confianza. En el límite, cuando *n* tiende a infinito, t tiende a z. De hecho, para valores de *n* mayores a 30 se considera que el valor de t puede ser estimado a través de la normal estándar. En este caso, como n = 50 vea que t está muy cercano a z (con una imprecisión de 2.5% para el intervalo bilateral).

El intervalo bilateral de confianza al 95% de nivel de confianza es

$$2252.87 \leq \mu_x \leq 2268.69$$

El intervalo superior de confianza al 95% de nivel de confianza es

$$2254.18 \leq \mu_x$$

El intervalo inferior de confianza al 95% de nivel de confianza es

$$\mu_x \leq 2267.38$$

Figura 4.9. Intervalo Bilateral de Confianza
para un parámetro de una población con una muestra simple

Variable	Condiciones iniciales	Estimador Puntual	Intervalo Bilateral de Confianza al (1-α)100%	Error
μ_x	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$ σ_x^2 *conocida*	$\bar{x}$	$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$	$E = z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$
μ_x	$x \sim N(\mu_x, \sigma_x)$ N finita σ_x^2 *conocida*	$\bar{x}$	$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)} \le \mu_x \le \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$	$E = z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$
μ_x	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$\bar{x}$	$\bar{x} - t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}} \le \mu_x \le \bar{x} + t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}}$	$E = t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}}$
μ_x	$x \sim N(\mu_x, \sigma_x)$ N finita	$\bar{x}$	$\bar{x} - t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)} \le \mu_x \le \bar{x} + t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$	$E = t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{S_{(n-1)}}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$
σ_x^2	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$S_{(n-1)}^2$	$\frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(\frac{\alpha}{2}, n-1\right)}} \le \sigma_x^2 \le \frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(1-\frac{\alpha}{2}, n-1\right)}}$	$E = \frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(1-\frac{\alpha}{2}, n-1\right)}}$
σ_x^2	$x \sim N(\mu_x, \sigma_x)$ N finita	$S_{(n-1)}^2$	$\frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(\frac{\alpha}{2}, n-1\right)}} \left(\frac{N-n}{N-1}\right) \le \sigma_x^2 \le \frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(1-\frac{\alpha}{2}, n-1\right)}} \left(\frac{N-n}{N-1}\right)$	$E = \frac{(n-1)S_{(n-1)}^2}{\chi^2_{\left(1-\frac{\alpha}{2}, n-1\right)}} \left(\frac{N-n}{N-1}\right)$
p	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$\hat{p}$	$\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
p	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$\hat{p}$	$\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)} \le p \le \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$	$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$
u	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$\hat{u}$	$\hat{u} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}} \le u \le \hat{u} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}}$	$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}}$
u	$x \sim N(\mu_x, \sigma_x)$ $N \to \infty$	$\hat{u}$	$\hat{u} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)} \le u \le \hat{u} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$	$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}}{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$

Figura 4.10. Fórmulas para estimar el tamaño de muestra necesario para deducir un intervalo de confianza para un parámetro de una población con una muestra simple

Parámetro	Estimador	Varianza del Estimador	Límite Error de Estimación	Tamaño de Muestra
Media μ_x	$\bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$	$\hat{\sigma}^2_{\bar{x}} = \frac{S^2_{n-1}}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\hat{\sigma}_{\bar{x}} = z_{\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	$n \geq \frac{N\hat{\sigma}^2_x}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2(N-1)+\hat{\sigma}^2_x}$
Total Poblac $\tau = N\mu_x$	$\hat{\tau} = N\bar{x} = \frac{N}{n}\sum_{i=1}^{i=n} x_i$	$\hat{\sigma}^2_{\hat{\tau}} = \frac{N^2 S^2_{n-1}}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\tau}} = z_{\frac{\alpha}{2}}\frac{NS_{n-1}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	$n \geq \frac{N^3\hat{\sigma}^2_x}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2(N-1)+N^2\hat{\sigma}^2_x}$
Fracción Defectuosa p	$\hat{p} = \frac{\sum_{k=1}^{k=m} d_k}{\sum_{k=1}^{k=m} n_k}$ ó $\hat{p} = \frac{d}{n}$	$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N\hat{p}(1-\hat{p})}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2(N-1)+\hat{p}(1-\hat{p})}$
Fracción de Defectos por Unidad u	$\hat{u} = \frac{\sum_{k=1}^{k=m} c_k}{\sum_{k=1}^{k=m} n_k}$ ó $\hat{u} = \frac{c}{n}$	$\hat{\sigma}^2_{\hat{u}} = \frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)}$	$n \geq \frac{N\hat{u}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2(N-1)+\hat{u}}$

4.7. Estimación de un mismo parámetro poblacional para dos poblaciones

En este subtema se analiza la relación que guarda un parámetro estadístico como la media, varianza, fracción de elementos exitosos o fracción de éxitos en dos poblaciones diferentes. Este subtema tiene muchas aplicaciones, por ejemplo, en el caso de la gasolina magna o de la gasolina premium en México, ¿cuál es mejor de las dos?, si el criterio para compararlas es el precio, pues se elegiría la más barata; sin embargo, la gasolina premium contiene 92 octanos y la magna 87, lo cual hace más explosiva a la premium sobre la magna; el hecho de que sea más explosiva significa que el motor se desgasta menos y reduce la cantidad de contaminantes que se emiten a la atmósfera, pero la diferencia en octanos puede ser no significativa, dado lo cual la pregunta podría ser ¿cuál presenta mayor rendimiento en kilometraje recorrido por litro consumido, la premium o la magna?, esto implica comparar dos poblaciones de automóviles diferentes, los que usen la premium y los que usen la magna, de allí la necesidad de comparar la media en kilometraje recorrido por litro consumido de la gasolina premium contra la magna.

Por otra parte, citando otro ejemplo, la Organización Mundial de la Salud establece una norma para el Índice de Masa Corporal (IMC) que debiera cumplir una persona. Según dicha norma, el IMC debe estar entre 18.5 y 25 Kg/m^2 para considerarse como normal, pero queda la duda: ¿este índice depende del sexo?, ¿la norma es la misma para un hombre que para una mujer? De aquí la necesidad de comparar si el IMC del hombre es igual al de la mujer.

4.7.1. Intervalo de confianza del cociente entre varianzas para dos poblaciones normales

Suponga que se tienen dos variables aleatorias normales, estadísticamente independientes, $x_1 \sim N(\mu_1, \sigma_1)$ y $x_2 \sim N(\mu_2, \sigma_2)$.

En el volumen III de Fundamentos de Probabilidad se definió a una variable aleatoria F como el cociente entre dos variables aleatorias χ^2, entre sus grados de libertad, como se muestra en la expresión 1.82 de este volumen. La función de densidad de probabilidad para una distribución F se establece en la expresión 1.83; su media en la expresión 1.85 y su varianza en la expresión 1.86.

En la expresión 1.87 se estableció que la variable

$$\frac{\frac{S^2_{n_1-1}}{\sigma_1^2}}{\frac{S^2_{n_2-1}}{\sigma_2^2}} \sim F_{n_1-1,\, n_2-1}$$

Por lo cual

$$F_{\frac{\alpha}{2},\, n_1-1,\, n_2-1} \leq \frac{\frac{S^2_{n_1-1}}{\sigma_1^2}}{\frac{S^2_{n_2-1}}{\sigma_2^2}} \leq F_{1-\frac{\alpha}{2},\, n_1-1,\, n_2-1}$$

Despejando el cociente entre varianzas

$$\frac{S^2_{n_2-1}}{S^2_{n_1-1}} F_{\frac{\alpha}{2},\, n_1-1,\, n_2-1} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\frac{\alpha}{2},\, n_1-1,\, n_2-1} \frac{S^2_{n_2-1}}{S^2_{n_1-1}}$$

Invirtiendo el cociente, recordando que al invertir se invierte el símbolo de desigualdad y sabiendo que

$$F_{1-\frac{\alpha}{2},\, n_2-1,\, n_1-1} = \frac{1}{F_{\frac{\alpha}{2},\, n_1-1,\, n_2-1}}$$

Se obtiene un intervalo bilateral de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales:

$$\frac{S^2_{n_1-1}}{S^2_{n_2-1}} F_{1-\frac{\alpha}{2},\, n_2-1,\, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S^2_{n_1-1}}{S^2_{n_2-1}} F_{\frac{\alpha}{2},\, n_2-1,\, n_1-1} \tag{4.97}$$

El intervalo inferior de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales es:

$$\frac{\sigma_1^2}{\sigma_2^2} \leq F_{\alpha,\, n_2-1,\, n_1-1} \frac{S^2_{n_1-1}}{S^2_{n_2-1}} \tag{4.98}$$

El intervalo superior de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales es:

$$\frac{S^2_{n_1-1}}{S^2_{n_2-1}} F_{1-\alpha,\, n_2-1,\, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \tag{4.99}$$

Si las poblaciones de donde se extraen las muestras son finitas, el intervalo bilateral de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales es:

$$\frac{S^2_{n_1-1}}{S^2_{n_2-1}}\frac{\left(\frac{N_1-n_1}{N_1-1}\right)}{\left(\frac{N_2-n_2}{N_2-1}\right)}F_{1-\frac{\alpha}{2},\,n_2-1,\,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S^2_{n_1-1}}{S^2_{n_2-1}}\frac{\left(\frac{N_1-n_1}{N_1-1}\right)}{\left(\frac{N_2-n_2}{N_2-1}\right)}F_{\frac{\alpha}{2},\,n_2-1,\,n_1-1} \tag{4.100}$$

Si las poblaciones de donde se extraen las muestras son finitas, el intervalo inferior de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales es:

$$\frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S^2_{n_1-1}}{S^2_{n_2-1}}\frac{\left(\frac{N_1-n_1}{N_1-1}\right)}{\left(\frac{N_2-n_2}{N_2-1}\right)}F_{\alpha,\,n_2-1,\,n_1-1} \tag{4.101}$$

Si las poblaciones son finitas, el intervalo superior de confianza al 100(1–α)% de nivel de confianza del cociente de dos varianzas poblacionales es:

$$\frac{S^2_{n_1-1}}{S^2_{n_2-1}}\frac{\left(\frac{N_1-n_1}{N_1-1}\right)}{\left(\frac{N_2-n_2}{N_2-1}\right)}F_{1-\alpha,\,n_2-1,\,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \tag{4.102}$$

Ejercicio 4.22

Se desea determinar los intervalos bilateral, inferior y superior de confianza al 95% del nivel de confianza, para el cociente entre varianzas del índice de masa corporal de los estudiantes con género masculino, comparado con el índice de masa corporal de los estudiantes con género femenino, de la Facultad de Ingeniería de la UNAM. Suponga que los tamaños poblacionales son infinitos. Para ello, se eligió una muestra aleatoria de 120 estudiantes, a los cuales se les preguntó su estatura y peso. Los resultados se muestran en la siguiente tabla.

No.	Sexo	Peso	Estat	No.	Sexo	Peso	Estat	No.	Sexo	Peso	Estat	No.	Sexo	Peso	Estat	No.	Sexo	Peso	Estat
1	F	76	1.69	25	F	58	1.65	49	F	48	1.52	73	M	72	1.69	97	M	62	1.65
2	F	56	1.51	26	F	64	1.60	50	F	56	1.56	74	M	78	1.78	98	M	64	1.68
3	F	56	1.50	27	F	53	1.63	51	F	60	1.64	75	M	63	1.69	99	M	63	1.74
4	F	73	1.65	28	F	55	1.53	52	F	60	1.50	76	M	65	1.72	100	M	73	1.76
5	F	60	1.51	29	F	59	1.65	53	F	75	1.68	77	M	74	1.64	101	M	84	1.70
6	F	50	1.60	30	F	72	1.60	54	F	52	1.63	78	M	56	1.67	102	M	79	1.62
7	F	61	1.58	31	F	65	1.66	55	F	54	1.52	79	M	57	1.75	103	M	65	1.73
8	F	61	1.56	32	F	58	1.60	56	F	50	1.59	80	M	91	1.85	104	M	52	1.56
9	F	52	1.53	33	F	47	1.60	57	F	52	1.54	81	M	76	1.74	105	M	52	1.70
10	F	60	1.72	34	F	54	1.57	58	F	50	1.56	82	M	74	1.76	106	M	68	1.68
11	F	70	1.54	35	F	60	1.62	59	F	60	1.50	83	M	66	1.70	107	M	80	1.72
12	F	65	1.63	36	F	48	1.60	60	F	45	1.55	84	M	80	1.80	108	M	76	1.65
13	F	63	1.54	37	F	52	1.51	61	F	55	1.60	85	M	64	1.68	109	M	72	1.80
14	F	57	1.70	38	F	64	1.58	62	F	55	1.67	86	M	59	1.72	110	M	85	1.75
15	F	41	1.52	39	F	59	1.58	63	F	45	1.61	87	M	73	1.73	111	M	77	1.72
16	F	44	1.50	40	F	60	1.67	64	F	55	1.58	88	M	82	1.68	112	M	64	1.62
17	F	61	1.63	41	F	57	1.64	65	M	60	1.73	89	M	79	1.88	113	M	52	1.66
18	F	53	1.60	42	F	60	1.65	66	M	68	1.85	90	M	74	1.68	114	M	56	1.70
19	F	70	1.59	43	F	65	1.57	67	M	85	1.75	91	M	67	1.67	115	M	65	1.76
20	F	55	1.63	44	F	39	1.55	68	M	62	1.69	92	M	89	1.79	116	M	65	1.75
21	F	57	1.60	45	F	45	1.53	69	M	84	1.82	93	M	62	1.70	117	M	65	1.74
22	F	52	1.58	46	F	56	1.63	70	M	69	1.54	94	M	65	1.70	118	M	72	1.64
23	F	55	1.53	47	F	58	1.60	71	M	68	1.74	95	M	65	1.65	119	M	68	1.68
24	F	64	1.50	48	F	53	1.50	72	M	79	1.79	96	M	60	1.66	120	M	75	1.80

La tabla anterior debe separarse en dos muestras diferentes, una para 56 hombres (M) y la otra para 64 mujeres (F). Para cada muestra debe calcularse el Índice de Masa Corporal, el cual se define como:

$$IMC = \frac{Peso}{Estatura^2}$$

Para cada una de las muestras se calcula la varianza muestral, resultando lo siguiente:

$$S_M^2 = 8.773 \quad S_F^2 = 8.759$$

No.	Sexo	Peso	Estat	IMC
1	M	60	1.73	20.05
2	M	68	1.85	19.87
3	M	85	1.75	27.76
4	M	62	1.69	21.71
5	M	84	1.82	25.36
6	M	69	1.54	29.09
7	M	68	1.74	22.46
8	M	79	1.79	24.66
9	M	72	1.69	25.21
10	M	78	1.78	24.62
11	M	63	1.69	22.06
12	M	65	1.72	21.97
13	M	74	1.64	27.51
14	M	56	1.67	20.08
15	M	57	1.75	18.61
16	M	91	1.85	26.59
17	M	76	1.74	25.10
18	M	74	1.76	23.89
19	M	66	1.70	22.84
20	M	80	1.80	24.69
21	M	64	1.68	22.68
22	M	59	1.72	19.94
23	M	73	1.73	24.39
24	M	82	1.68	29.05
25	M	79	1.88	22.35
26	M	74	1.68	26.22
27	M	67	1.67	24.02
28	M	89	1.79	27.78
29	M	62	1.70	21.45
30	M	65	1.70	22.49
31	M	65	1.65	23.88
32	M	60	1.66	21.77
33	M	62	1.65	22.77
34	M	64	1.68	22.68
35	M	63	1.74	20.81
36	M	73	1.76	23.57
37	M	84	1.70	29.07
38	M	79	1.62	30.10
39	M	65	1.73	21.72
40	M	52	1.56	21.37
41	M	52	1.70	17.99
42	M	68	1.68	24.09
43	M	80	1.72	27.04
44	M	76	1.65	27.92
45	M	72	1.80	22.22
46	M	85	1.75	27.76
47	M	77	1.72	26.03
48	M	64	1.62	24.39
49	M	52	1.66	18.87
50	M	56	1.70	19.38
51	M	65	1.76	20.98
52	M	65	1.75	21.22
53	M	65	1.74	21.47
54	M	72	1.64	26.77
55	M	68	1.68	24.09
56	M	75	1.80	23.15
$Media_M$=				23.67
Var_M=				8.773

No.	Sexo	Peso	Estat	IMC
1	F	76	1.69	26.61
2	F	56	1.51	24.56
3	F	56	1.50	24.89
4	F	73	1.65	26.81
5	F	60	1.51	26.31
6	F	50	1.60	19.53
7	F	61	1.58	24.44
8	F	61	1.56	25.07
9	F	52	1.53	22.21
10	F	60	1.72	20.28
11	F	70	1.54	29.52
12	F	65	1.63	24.46
13	F	63	1.54	26.56
14	F	57	1.70	19.72
15	F	41	1.52	17.75
16	F	44	1.50	19.56
17	F	61	1.63	22.96
18	F	53	1.60	20.70
19	F	70	1.59	27.69
20	F	55	1.63	20.70
21	F	57	1.60	22.27
22	F	52	1.58	20.83
23	F	55	1.53	23.50
24	F	64	1.50	28.44
25	F	58	1.65	21.30
26	F	64	1.60	25.00
27	F	53	1.63	19.95
28	F	55	1.53	23.50
29	F	59	1.65	21.67
30	F	72	1.60	28.13
31	F	65	1.66	23.59
32	F	58	1.60	22.66
33	F	47	1.60	18.36
34	F	54	1.57	21.91
35	F	60	1.62	22.86
36	F	48	1.60	18.75
37	F	52	1.51	22.81
38	F	64	1.58	25.64
39	F	59	1.58	23.63
40	F	60	1.67	21.51
41	F	57	1.64	21.19
42	F	60	1.65	22.04
43	F	65	1.57	26.37
44	F	39	1.55	16.23
45	F	45	1.53	19.22
46	F	56	1.63	21.08
47	F	58	1.60	22.66
48	F	53	1.50	23.56
49	F	48	1.52	20.78
50	F	56	1.56	23.01
51	F	60	1.64	22.31
52	F	60	1.50	26.67
53	F	75	1.68	26.57
54	F	52	1.63	19.57
55	F	54	1.52	23.37
56	F	50	1.59	19.78
57	F	52	1.54	21.93
58	F	50	1.56	20.55
59	F	60	1.50	26.67
60	F	45	1.55	18.73
61	F	55	1.60	21.48
62	F	55	1.67	19.72
63	F	45	1.61	17.36
64	F	55	1.58	22.03
$Media_F$=				22.65
Var_F=				8.759

Ahora deben obtenerse los valores de F al 95% de nivel de confianza, utilizando Excel:

$$F_{\frac{\alpha}{2},\, n_2-1,n_1-1} = INV.F.CD\left(0.025,63,55\right) = 0.5993$$

$$F_{1-\frac{\alpha}{2},\, n_2-1,n_1-1} = INV.F.CD\left(0.975,63,55\right) = 1.6839$$

$$F_{\alpha,\, n_2-1,n_1-1} = INV.F.CD\left(0.025,63,55\right) = 0.65109$$

$$F_{1-\alpha,\, n_2-1,n_1-1} = INV.F.CD\left(0.975,63,55\right) = 1.54721$$

El intervalo bilateral de confianza al 95% de nivel de confianza es:

$$0.06003 \leq \frac{\sigma_M^2}{\sigma_F^2} \leq 1.6865$$

El intervalo inferior de confianza al 95% de nivel de confianza es:

$$\frac{\sigma_M^2}{\sigma_F^2} \leq 1.5496$$

El intervalo superior de confianza al 95% de nivel de confianza es:

$$0.65109 \leq \frac{\sigma_M^2}{\sigma_F^2}$$

Observe la expresión 4.79. ¿Qué ocurre si $\sigma_1^2 = \sigma_2^2$? Como se puede apreciar, si se supusiera que las varianzas poblacionales son iguales, entonces el intervalo bilateral de confianza debería contener en su interior al valor uno, lo cual para este caso se cumple con un nivel de confianza del 95%. Si el valor uno cayera fuera del intervalo bilateral, entonces se podría concluir que las varianzas son diferentes.

4.7.2. Intervalo de confianza de la diferencia entre medias para dos poblaciones normales estadísticamente independientes con varianzas conocidas

En el volumen III de Fundamentos de Probabilidad se demostró el Teorema de Aditividad o Reproducibilidad de la Distribución Normal. En la expresión 1.59 de este volumen se representa.

Suponga que se tienen dos variables aleatorias normales, estadísticamente independientes, con varianzas conocidas $\chi_1 \sim N(\mu_1,\sigma_1)$ y $\chi_2 \sim N(\mu_2,\sigma_2)$, ambas con población infinita. Por el Teorema de Aditividad de la Distribución Normal o por el Teorema del Límite Central, las medias muestrales de ambas poblaciones también presentan distribución normal, es decir,

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n^1}}\right) \text{ y } \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

Por los mismos teoremas anteriores, la diferencia de estas dos variables también es normal

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Por lo cual, estandarizando la variable anterior,

$$-z_{\frac{\alpha}{2}} \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\frac{\alpha}{2}}$$

Al despejar la diferencia entre medias poblacionales se obtiene un intervalo bilateral de confianza al 100(1–α)% de nivel de confianza de la diferencia entre dos medias poblacionales de dos poblaciones normales estadísticamente independientes con varianzas conocidas, así como sus intervalos unilaterales inferior y superior de confianza.

$$(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (4.103)$$

$$(\mu_1-\mu_2)\leq(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{4.104}$$

$$(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leq(\mu_1-\mu_2) \tag{4.105}$$

Si ambas poblaciones fueran finitas, los intervalos anteriores serían

$$(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\leq(\mu_1-\mu_2)\leq(\bar{x}_1+\bar{x}_2)+$$

$$+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.106}$$

$$(\mu_1-\mu_2)\leq(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.107}$$

$$(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\leq(\mu_1-\mu_2) \tag{4.108}$$

El error de estimación del intervalo de confianza al (1–α)100% de nivel de confianza para la diferencia entre medias de dos poblaciones normales finitas con varianzas conocidas está dado, de acuerdo con la expresión 4.102, por

$$\varepsilon=z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$$

Si se toman las muestras de tal forma que sus tamaños de muestra son iguales, es decir, $n=n_1=n_2$ y se despeja el valor de n, se obtiene una expresión para calcular el tamaño de muestra que se requiere tomar para estimar un intervalo de confianza de la diferencia entre medias, para poblaciones con varianzas conocidas y tamaño finito:

$$n\geq\frac{\frac{N_1\sigma_1^2}{N_1-1}+\frac{N_2\,\sigma_2^2}{N_2-1}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2+\left(\frac{\sigma_1^2}{N_1-1}+\frac{\sigma_2^2}{N_2-1}\right)} \tag{4.109}$$

Si el tamaño de ambas poblaciones es infinito la expresión es

$$n \geq \frac{\sigma_1^2 + \sigma_2^2}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2} \tag{4.110}$$

Ejercicio 4.23

Una ciudad es abastecida de energía eléctrica por dos termoeléctricas independientes una de la otra. En los últimos 40 días se observó que la termoeléctrica A generó como media 100000 kw/h y la termoeléctrica B generó como media 150000 kw/h, respectivamente. Suponga que sus desviaciones estándar son conocidas por datos históricos 8000 kw/h y 10000 kw/h respectivamente.

a. Con un nivel de confianza del 90%, estime la energía media que recibe la ciudad.

La energía media que recibe la ciudad es igual a la suma de las energías medias que generan las dos termoeléctricas, es decir,

$$\bar{x}_1 + \bar{x}_2 \sim N\left(\mu_1 + \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

El intervalo de confianza de la suma, suponiendo que las poblaciones son infinitas, está dado por la expresión

$$\left(\bar{x}_1 + \bar{x}_2\right) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \left(\mu_1 + \mu_2\right) \leq \left(\bar{x}_1 + \bar{x}_2\right) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

El valor de z al 90% de nivel de confianza, para un intervalo bilateral (el primero) y para uno unilateral (el segundo) es

$$z_{0.05} = INV.\ NORM.\ ESTAND\left(0.95\right) = 1.644854$$
$$z_{0.10} = INV.\ NORM.\ ESTAND\left(0.90\right) = 1.281552$$

Por lo cual el intervalo bilateral de confianza al 90% de nivel de confianza para la suma de las medias de dos poblaciones normales infinitas con varianzas conocidas es

$$246{,}669.425 \leq \mu_1 + \mu_2 \leq 253{,}330.575$$

Su intervalo inferior de confianza, al 90% de nivel de confianza es

$$\mu_1 + \mu_2 \leq 252{,}594.944$$

Su intervalo superior de confianza, al 90% de nivel de confianza es

$$247{,}405.056 \leq \mu_1 + \mu_2$$

b. Con un nivel de confianza del 90%, estime la diferencia en energía media de cada una de las dos termoeléctricas.

El intervalo de confianza para la diferencia de medias de dos poblaciones normales infinitas, con varianzas conocidas, está dado por la expresión 4.99. Por lo cual el intervalo bilateral de confianza al 90% de nivel de confianza para la diferencia de las medias de dos poblaciones finitas normales con varianzas conocidas es

$$-53{,}330.5747 \leq \mu_1 - \mu_2 \leq -46{,}669.4253$$

Su intervalo inferior de confianza, al 90% de nivel de confianza es

$$\mu_1 - \mu_2 \leq -47{,}405.0559$$

Su intervalo superior de confianza, al 90% de nivel de confianza es

$$-52{,}594.9441 \leq \mu_1 - \mu_2$$

¿Qué ocurriría si las medias poblacionales fueran iguales?, esto implicaría que el intervalo de confianza para la diferencia entre medias contendría al cero; por lo cual, si se pretendiera demostrar que las medias son diferentes bastaría con determinar si el intervalo bilateral de confianza no contiene al cero.

¿Cuál es la magnitud del error de estimación del intervalo bilateral calculado anteriormente?

$$\varepsilon = 3{,}330.5747$$

¿De qué tamaño tendría que ser la muestra para que el error aleatorio en el intervalo bilateral de confianza fuera menor de 3000?

De la expresión 4.110:

$$n \geq \frac{8000^2 + 10000^2}{\left(\frac{3000}{1.644854}\right)^2} = 49.301$$

Ejercicio 4.24

Una empresa del sector eléctrico usa dos tipos de materiales aislantes (*A* y *B*) en las piezas que fabrica. Suponga que se fabrican 3000 piezas con el material *A* y 2000 con el material *B* diariamente. Para determinar la resistencia media del material *A* se tomaron $n = 8$ lecturas y para el material *B*, nueve lecturas. Los datos obtenidos se muestran a continuación:

Material	1	2	3	4	5	6	7	8	9
A	1.25	1.16	1.33	1.15	1.23	1.2	1.32	1.28	
B	1.01	0.89	0.97	0.95	0.94	1.02	0.99	1.06	0.98

Suponga que la varianza en la resistencia del material *A* es 0.05 y la varianza en la resistencia del material *B* es 0.0025.

a. Determine intervalos de confianza bilateral, inferior y superior, al 98% de nivel de confianza, de la diferencia entre las resistencias medias de ambos materiales.

El valor de z al 98% de nivel de confianza, para un intervalo bilateral (el primero) y para uno unilateral (el segundo) es

$$z_{0.01} = INV.\ NORM.\ ESTAND(0.99) = 2.32635$$
$$z_{0.02} = INV.\ NORM.\ ESTAND(0.98) = 2.05375$$

Las medias de ambas muestras son:

$$\bar{x}_1 = 1.24$$
$$\bar{x}_2 = 0.97889$$

Por lo cual, dado que se trata de poblaciones finitas que se consideran como normales con varianzas conocidas, se usará la expresión 4.102. El intervalo bilateral de confianza al 98% de nivel de confianza para la diferencia entre las resistencias medias de ambos materiales con poblaciones normales con varianza conocida es

$$0.19131 \leq \mu_1 - \mu_2 \leq 0.33091$$

Como se puede apreciar, este intervalo no contiene al cero, por lo cual se puede asegurar que la resistencia de ambos materiales, al 98% de nivel de confianza, no es la misma. Se observa que el material A es más resistente que el B.

Su intervalo inferior de confianza, al 98% de nivel de confianza es

$$\mu_1 - \mu_2 \leq 0.32273$$

Su intervalo superior de confianza, al 98% de nivel de confianza es

$$0.19949 \leq \mu_1 - \mu_2$$

El fabricante del material B asegura que su material es tan resistente como el del proveedor del material A, ¿usted qué le diría?

El material de A es al menos casi dos décimas de unidad más resistente que el material B con un 98% de nivel de confianza.

b. Se desea tener un error de estimación menor a 0.05, ¿de qué tamaño tendría que ser la muestra en ambos materiales?

Utilizando la expresión 4.105

$$n \geq \frac{\frac{3000(0.005)}{2999} + \frac{2000(0.0025)}{1999}}{\left[\left(\frac{0.05}{2.32635}\right)^2 + \frac{0.005}{2999} + \frac{0.0025}{1999}\right]} = 16.1401$$

La muestra tendría que ser mayor a 16.

4.7.3. Intervalo de Confianza de la Diferencia entre Medias para dos poblaciones normales estadísticamente independientes con varianzas desconocidas pero iguales

Suponga que se tienen dos variables aleatorias normales, estadísticamente independientes $x_1 \sim N(\mu_1,\sigma_1)$ y $x_2 \sim N(\mu_2,\sigma_2)$. Por el Teorema de Aditividad de la Distribución Normal o por el Teorema del Límite Central, las medias muestrales de ambas poblaciones también presentan distribución normal, es decir,

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \quad \text{y} \quad \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

Por los mismos teoremas anteriores, la diferencia de estas dos variables también es normal

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Si se supone que $\sigma_2 = \sigma_1^2 = \sigma_2^2$ la diferencia anterior

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

SI se estandariza esta variable, se obtiene una normal estándar

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Por otra parte, las variables

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_2 - 1} \quad \text{y} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1}$$

La suma de dos variables aleatorias ji cuadrada, estadísticamente independientes, también da a su vez otra variable ji cuadrada con grados de libertad iguales a la suma de los grados de libertad de las dos variables aleatorias ji cuadrada que se estén sumando, es decir $(n_1-1) + (n_2-1) = n_1 + n_2-2$, o sea,

$$\frac{(n_1-1)S_1^2}{\sigma_2}+\frac{(n_2-1)S_2^2}{\sigma_2}\sim\chi^2_{n_1+n_2-1}$$

Recuerde que el cociente de una normal estándar entre la raíz cuadrada de una ji cuadrada entre sus grados de libertad, da como resultado una variable aleatoria t de Student, o sea,

$$\frac{Z}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{(n_1+n_2-2)}}}=\frac{\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}}\sim t_{n_1+n_2-2}$$

Lo cual implica que

$$-t_{\frac{\alpha}{2},\,n_1+n_2-2}\sim\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{1-\frac{\alpha}{2},\,n_1+n_2-2} \tag{4.111}$$

Si se hace

$$\hat{\sigma}=Sp=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \tag{4.112}$$

Expresión que representa un estimador ponderado puntual de la desviación estándar poblacional, de la diferencia entre medias de dos poblaciones normales infinitas con varianzas desconocidas pero iguales.

Si ambas poblaciones de donde se extraen las muestras son finitas, el estimador puntual de la desviación estándar poblacional de la diferencia entre medias de dos poblaciones normales finitas con varianzas desconocidas pero iguales es

$$\hat{\sigma}=Sp=\sqrt{\frac{(n_1-1)\left(\frac{N_1-n_1}{N_1-1}\right)S_1^2+(n_2-1)\left(\frac{N_2-n_2}{N_2-1}\right)S_2^2}{n_1+n_2-2}} \tag{4.113}$$

Despejando la diferencia entre las medias poblacionales, de la expresión 4.106, se obtiene el intervalo bilateral de confianza, al 100(1–α)% del nivel de confianza, para la diferencia entre las medias de dos poblaciones normales con varianzas desconocidas pero iguales.

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le (\mu_1 - \mu_2) \le (x_1 - x_2) + t_{\frac{\alpha}{2},n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4.114}$$

El intervalo inferior de confianza, al 100(1–α)% del nivel de confianza, para la diferencia entre las medias de dos poblaciones normales con varianzas desconocidas pero iguales.

$$(\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4.115}$$

El intervalo superior de confianza, al 100(1–α)% del nivel de confianza, para la diferencia entre las medias de dos poblaciones normales con varianzas desconocidas pero iguales.

$$(\bar{x}_1 - \bar{x}_2) + t_{\alpha,n_1+n_2-2}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le (\mu_1 - \mu_2) \tag{4.116}$$

Ejercicio 4.25 Determinar el intervalo bilateral, unilateral inferior y unilateral superior de confianza al 95% del nivel de confianza respectivamente, para la diferencia entre medias poblacionales del índice de masa corporal del género masculino, comparado con el índice de masa corporal del género femenino, de la Facultad de Ingeniería de la UNAM, cuyos datos fueron proporcionados en el ejercicio 4.22. Suponga que los tamaños poblacionales son infinitos.

De la tabla correspondiente al ejercicio 4.22 se sabe que:

$n_1 = 56 \qquad \bar{x}_1 = 23.67 \qquad S^2_{n_1-1} = 8.773$

$n_2 = 56 \qquad \bar{x}_2 = 22.65 \qquad S^2_{n_2-1} = 8.759$

También, en el ejercicio 4.22 se obtuvo un intervalo de confianza para el cociente entre varianzas

$$0.06003 \leq \frac{\sigma_M^2}{\sigma_F^2} \leq 1.6865$$

Observe el intervalo de confianza anterior del cociente entre varianzas, como se puede apreciar, este intervalo contiene al uno, lo que implica que se puede suponer que existe igualdad entre varianzas $\sigma_1^2 = \sigma_2^2$. Si el valor uno cayera fuera del intervalo bilateral entonces se podría concluir que existe evidencia estadística que demuestra que las varianzas son diferentes, lo cual no es el caso.

Se determina el valor de t al 95% de nivel de confianza de dos colas y de una cola, con $n_1 + n_2 - 2$ grados de libertad:

$$t_{\frac{\alpha}{2},n_1+n_2-2} = t_{0.025,118} = INV.T\left(0.025,118\right) = 1.980272$$

$$t_{\frac{\alpha}{2},n_1+n_2-2} = t_{0.05,118} = INV.T\left(0.025,118\right) = 1.657869$$

Nótese que:

$z_{0.025} = 1.959964$
$z_{0.05} = 1.644854$

El estimador puntual de la varianza poblacional sería

Sp = 2.960715

Por lo cual el intervalo bilateral de confianza para la diferencia de las medias del IMC para el género masculino y el género femenino de los estudiantes de la Facultad de Ingeniería, con varianzas desconocidas pero iguales, al 95% de nivel de confianza es

$$-0.038883 \leq \mu_M - \mu_F \leq 2.08476$$

Como se puede apreciar, el intervalo anterior contiene al cero, lo cual implica que se puede suponer que las medias del IMC de hombres y de mujeres es la misma, tal como lo contempla la Organización Mundial de Salud (OMC).

El intervalo inferior de confianza al 95% del nivel de confianza entre el IMC de los estudiantes hombres y mujeres en la Facultad de Ingeniería de la UNAM es

$$\mu_M - \mu_F \leq 1.914047$$

El intervalo superior de confianza al 95% del nivel de confianza entre el IMC de los estudiantes hombres y mujeres en la Facultad de Ingeniería de la UNAM es

$$0.13183 \leq \mu_M - \mu_F$$

Nótese que en este caso el intervalo superior no contiene al cero lo que implica que el IMC del hombre está por arriba del IMC de las mujeres al 95% de nivel de confianza en una prueba de una cola.

4.7.4. Intervalo de Confianza de la Diferencia entre Medias para dos poblaciones normales, estadísticamente independientes, con varianzas desconocidas, pero con tamaños de muestra grandes, es decir, mayor de 30 (para la t de student)

Con relación a los intervalos de confianza de la diferencia entre medias para dos poblaciones normales estadísticamente independientes con varianzas desconocidas establecidos en las expresiones 4.114, 4.115 y 4.116, si los tamaños de las muestras tomadas en cada una de las poblaciones son grandes, es decir, mayores a 30, el estadístico t puede ser aproximado a través del estadístico z, aún en el supuesto de que las varianzas poblacionales no sean conocidas y sean diferentes. En este último caso, los intervalos de confianza bilateral, inferior y superior, para poblaciones de tamaño infinito, estarían dados por las siguientes expresiones:

$$(\bar{x}_1-\bar{x}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\le(\mu_1-\mu_2)\le(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{4.117}$$

$$(\mu_1-\mu_2)\le(\bar{x}_1-\bar{x}_2)+z_{\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{4.118}$$

$$(\bar{x}_1-\bar{x}_2)+z_{\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\le(\mu_1-\mu_2) \tag{4.119}$$

Los intervalos de confianza bilateral, superior e inferior, para poblaciones de tamaño finito, estarían dados por las siguientes expresiones:

$$(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\le(\mu_1-\mu_2)\le(\bar{x}_1+\bar{x}_2)$$

$$+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.120}$$

$$(\mu_1-\mu_2)\le(\bar{x}_1-\bar{x}_2)+z_{\alpha}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.121}$$

$$(\bar{x}_1-\bar{x}_2)+z_{\alpha}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\le(\mu_1-\mu_2) \tag{4.122}$$

Ejercicio 4.26

Para establecer el límite máximo de velocidad de una carretera urbana, se considera que esta debe ser tal que solo el 15% de los vehículos que actualmente la usan, sin ningún límite de velocidad, la transiten a velocidad mayor del límite que se fije. En un día cualquiera se tomó una muestra aleatoria del número de vehículos, de diferentes clases, que circularon con las velocidades que se muestran en la siguiente tabla:

Velocidad (km/h)	50-60	60-70	70-80	80-90	90-100	100-110	110-120
Coches	1	6	27	16	5	4	1
Autobuses	7	18	5	3	1	0	0

a. ¿Recomendaría límites de velocidad diferentes para los coches y autobuses?

Velocidad (Km/h)	50-60	60-70	70-80	80-90	90-100	100-110	110-120			
Marca Clase	55	65	75	85	95	105	115	n	Media	Varianza
Coches	1	6	27	16	5	4	1	60	80.67	133.45
Autobuses	7	18	5	3	1	0	0	34	67.06	95.63

Nótese que ambos tamaños de muestra son mayores de 30, por lo cual se puede aproximar la t de Student a través de la normal estándar y usar las expresiones 4.94 para obtener el intervalo de confianza de la diferencia entre medias, si el intervalo contiene al cero entonces no se tiene evidencia estadística para rechazar la hipótesis estadística de que las medias son iguales.

$z_{0.975} = 1.959964$ $z_{0.95} = 1.644854$

Los intervalos de confianza son

$9.2 \leq x_1 - x_2 \leq 18$ al 95% de nivel de confianza
$x_1 - x_2 \leq 17.3$ al 95% de nivel de confianza
$9.9 \leq x_1 - x_2$ al 95% de nivel de confianza

Nótese que el intervalo bilateral no contiene al cero por lo que existe evidencia estadística para rechazar la igualdad entre las medias, la diferencia en velocidad en promedio entre los automóviles y los autobuses se encuentra entre 9 y 18 km/h.

b. Si se piensa adoptar un solo límite de velocidad para cualquier tipo de vehículo, ¿cuál sería la velocidad máxima permitida?

Velocidad (Km/h)	50-60	60-70	70-80	80-90	90-100	100-110	110-120				Lim Inf	Lim Sup	Int Inf	Int Sup
Marca Clase	55	65	75	85	95	105	115	n	Media	Varianza	Media	Media	Media	Media
Coches	1	6	27	16	5	4	1	60	80.67	133.45	77.74	83.59	83.12	78.21
Autobuses	7	18	5	3	1	0	0	34	67.06	95.63	63.77	70.35	69.82	64.30
Vehículos	8	24	32	19	6	4	1	94	75.74	161.81	73.17	78.32	77.90	73.59

Existen dos formas de obtener dicho límite, el primero sería seguir la recomendación que establece el encabezado del problema, que no más del 15% de los vehículos rebase el tope que se fije, al cual se le denominará LSE.

Suponiendo que los datos tienen distribución normal, de media 75.74 y varianza 161.81, se obtendría el valor de LSE tal que

$$p\left(x \leq LSE\right) \leq 0.15$$

Estandarizando y aplicando la inversa de la normal, se obtiene que:

LSE=INV.NORM(0.85, 75.74, raíz(161.81) = 88.92 km/h

Con lo cual el límite de velocidad sería 90 km/h

El otro método sería obtener un intervalo inferior de confianza para un solo vehículo y tomar el límite superior de este intervalo, en este caso estaría dado por

$$LSE\_vehiculo = \hat{\mu}_x + z_\alpha \, \hat{\sigma}_x = 75.74 + 1.645 * \sqrt{161.81} = 96.67 \text{ km/h}$$

Nótese que este intervalo proporciona un tope más alto, esto se debe básicamente a que se está calculando un límite superior para el 95% de los vehículos, esto significa que cuando mucho solo el 5% de ellos rebasaría el tope superior.

Si se hace para un nivel de confianza del 85%, el resultado sería

$$LSE\_vehiculo = \hat{\mu}_x + z_\alpha \, \hat{\sigma}_x = 75.74 + 1.036433 * \sqrt{161.81} = 88.92 \text{ km/h}$$

Un punto importante para tener en cuenta es que este tope superior no se obtuvo para la media, el cual estaría dado por

$$LSE\_vehiculo = \hat{\mu}_x + z_{\alpha}\frac{\hat{\sigma}_x}{\sqrt{n}} = 75.74 + 1.645 * \frac{\sqrt{161.81}}{\sqrt{94}} = 77.898$$

De lo anterior se desprende que los límites de confianza para la media son diferentes a los límites de confianza para las lecturas individuales

$$\hat{\mu}_x - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_x}{\sqrt{n}} \leq \mu_x \leq \mu_x + z_{\frac{\alpha}{2}}\frac{\hat{\sigma}_x}{\sqrt{n}}$$

$$\hat{\mu}_x - z_{\frac{\alpha}{2}}\hat{\sigma}_x \leq x \leq \hat{\mu}_x + z_{\frac{\alpha}{2}}\hat{\sigma}_x$$

4.7.5. Intervalo de Confianza de la Diferencia entre Medias para dos poblaciones normales estadísticamente independientes con varianzas desconocidas, diferentes y tamaños muestrales pequeños

De acuerdo a Welch, B. L. (1938), *The significance of the difference between two means when the population variances are unequal,* Biometrika, 29, 350-362, para el caso en que los tamaños muestrales son pequeños, $n < 30$, y las varianzas son desconocidas y distintas, los intervalos de confianza bilateral, superior e inferior, para poblaciones de tamaño infinito, estarían dados por las siguientes expresiones:

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{4.123}$$

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha,\upsilon}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \le (\mu_1 - \mu_2) \tag{4.124}$$

$$(\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + t_{\alpha,\upsilon}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{4.125}$$

Donde

$$\upsilon = \frac{\left(\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right)^2}{(n_1+1)}+\frac{\left(\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)\right)^2}{(n_2+1)}} - 2 \tag{4.126}$$

los intervalos de confianza bilateral, superior e inferior, para poblaciones de tamaño finito, estarían dados por las siguientes expresiones:

$$(\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \le (\mu_1 - \mu_2) \le (\bar{x}_1 + \bar{x}_2)$$

$$+ t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.127}$$

$$(\bar{x}_1 - \bar{x}_2) + t_{\alpha,\upsilon}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \le (\mu_1 - \mu_2) \tag{4.128}$$

$$(\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha,\upsilon}\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{S_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \tag{4.129}$$

Resuelva el ejercicio 4.24 partiendo del hecho que se desconocen las varianzas poblacionales.

Material	Media	Varianza
A	1.24	0.004629
B	0.978889	0.002461

$\upsilon = (0.004629/8 + 0.002461/9)\text{^}2/\left[(0.004629/8)\text{^}2/9 + (0.002461/9)\text{^}2/10\right] - 2 = 14.25$

$t_{\alpha/2,\upsilon} = \text{INV.T}(0.975, 14) = 2.144787$

$0.1985 \leq \mu_1 - \mu_2 \leq 0.3237$

4.7.6. Intervalo de Confianza de la Diferencia entre Medias para dos poblaciones normales con observaciones pareadas

Suponga que se desea saber qué gasolina, premium o magna, da mayor rendimiento en kilometraje por litro consumido; se decide hacer un experimento en el que se seleccionan n automóviles diferentes. El problema es que el rendimiento en kilometraje por litro de gasolina consumido depende de muchos y diversos factores, el número de cilindros de cada automóvil, las condiciones físicas de cada uno de ellos, las condiciones del terreno por donde se mueven, las condiciones del medio ambiente, la forma de manejo y el peso de cada conductor, el inflado de las llantas, el peso extra, el aire acondicionado, las ventanillas abiertas, etcétera. Para poder medir el kilometraje recorrido por litro consumido es necesario tratar de controlar cada uno de estos factores, lo cual puede resultar difícil. Para disminuir el efecto o variación de cada uno de estos factores, se decide hacer la prueba pareada, es decir, cada automóvil será probado con ambas gasolinas, el mismo día de la semana, a la misma hora, bajo las mismas condiciones de tránsito, bajo el mismo clima, con el mismo conductor y bajo las mismas condiciones de manejo, siguiendo la misma ruta, a 80 km/h, ventanillas cerradas, sin aire acondicionado, con un inflado de llantas de 32 libras.

Para ello, se decide escoger n automóviles, un sábado, a las 07:00 am, en la gasolinera "Que chula es Puebla" y allí vaciar y llenar el tanque de cada uno de los coches, con gasolina premium hasta el tope, poniendo el odómetro en cero kilómetros. Iniciar el recorrido procurando mantener una velocidad constante de 80 km/h rumbo a Puebla; llegar hasta la gasolinera que está a la entrada de Puebla en la Av. Aquiles Serdán y allí volver a llenar el tanque con gasolina Premium hasta el tope midiendo y registrando el número de litros que se le introdujeron y el número de kilómetros recorridos, con ello se puede calcular el kilometraje recorrido por litro consumido usando gasolina premium.

Otro sábado, con las mismas condiciones de tránsito, de clima, con el mismo conductor en cada automóvil, bajo las mismas condiciones de manejo, se repite el experimento, pero ahora utilizando gasolina magna. Al finalizar el segundo experimento se tienen parejas ordenadas (x_{11}, x_{21}), $(x_{12}, x_{22}), \cdots, (x_{1n}, x_{2n})$. Se partirá del supuesto que, tanto el kilometraje recorrido por litro consumido de gasolina premium como el kilometraje recorrido por litro consumido de gasolina magna, presentan distribución normal; cabe señalar que como el experimento se realizó con observaciones pareadas, ambas poblaciones no son necesariamente estadísticamente independientes.

Suponga que se obtiene la diferencia entre cada una de las observaciones pareadas, es decir, $D_1 = x_{11} - x_{21}$, $D_2 = x_{12} - x_{22}, \cdots, D_n = x_{1n} - x_{2n}$.

La media de la población de diferencias, dado que se supone que las variables x_1 y x_2 son normales, también será normal.

Ahora, si se aplica un procedimiento similar al que se siguió en el subtema 4.3.5, un intervalo de confianza bilateral, inferior y superior, de la variable aleatoria D, para la diferencia de observaciones pareadas, al $(1-\alpha)100\%$ de nivel de confianza, está dado por las expresiones:

$$\overline{D} - t_{\frac{\alpha}{2},n-1}\frac{S_D}{\sqrt{n}} \leq D \leq D + t_{\frac{\alpha}{2}n-1}\frac{S_D}{\sqrt{n}} \tag{4.130}$$

$$D \leq \overline{D} + t_{\alpha,n-1}\frac{S_D}{\sqrt{n}} \tag{4.131}$$

$$\overline{D} - t_{\alpha,n-1}\frac{S_D}{\sqrt{n}} \leq D \tag{4.132}$$

Si las poblaciones tienen un tamaño finito N, entonces, un intervalo de confianza bilateral, inferior y superior, de la variable aleatoria D, para la diferencia de observaciones pareadas, al $(1-\alpha)100\%$ de nivel de confianza, está dado por las expresiones:

$$\overline{D} - t_{\frac{\alpha}{2},n-1}\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right) \leq D \leq \overline{D} + t_{\frac{\alpha}{2},n-1}\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right) \tag{4.133}$$

$$D \leq \overline{D} + t_{\alpha,n-1}\frac{S_D}{\sqrt{n}} \tag{4.134}$$

$$\overline{D} - t_{\alpha,n-1}\frac{S_D}{\sqrt{n}} \leq D \tag{4.135}$$

Ejercicio 4.27

Suponga que se decide llevar a efecto el experimento antes mencionado de verificar cuál gasolina da mayor rendimiento en kilometraje por litro consumido, por lo que se seleccionan ocho automóviles de diferentes marcas y se hacen las pruebas obteniendo los siguientes resultados:

Automóvil	Rendimiento km/l	
	Magna	Premium
Fiat 500	18.08	20.00
Civic Honda	13.09	17.23
Mazda 3	22.98	23.55
Mercedes C200	16.07	25.55
Prius C	26.17	26.59
Audi A3	22.77	24.65
BMW320iA	25.00	25.11
Ford Focus	19.33	21.56

Cabe señalar que los datos fueron generados artificialmente con distribución normal, tomando como referencia parámetros reales de la PROFECO.

Determine si el rendimiento en km/l depende de la gasolina y en tal caso cuál es mejor.

Automóvil	Rendimiento km/l		
	Magna	Premium	D
Fiat 500	18.08	20.00	1.92
Civic Honda	13.09	17.23	4.14
Mazda 3	22.98	23.55	0.57
Mercedes C200	16.07	25.55	9.48
Prius C	26.17	26.59	0.42
Audi A3	22.77	24.65	1.88
BMW320iA	25.00	25.11	0.11
Ford Focus	19.33	21.56	2.23
Media =	20.44	23.03	2.59
DesvEst =	4.56	3.19	3.07

$t_{\alpha/2,n-1} =$ 2.26216

$t_{\alpha,n-1} =$ 1.83311

Con los datos dados, se obtiene el siguiente intervalo de confianza:

$$0.39889 \leq D \leq 4.78861$$

Se restó el rendimiento de la gasolina premium menos el rendimiento de la gasolina magna; se puede apreciar que el intervalo de confianza de la diferencia es positivo, no contiene al cero, eso implica que la gasolina premium es un poco mejor que la magna, les proporciona entre 0.4 y 4.8 km de más por cada litro consumido.

4.7.7. Intervalo de confianza de la diferencia entre dos proporciones de poblaciones normales

Suponga que se tiene una población o lote de artículos de tamaño N, de los cuales se establece que existen D de ellos con cierta característica de interés (por ejemplo ser defectuoso); de esta población se extrae aleatoriamente una muestra representativa de tamaño n, en la cual se establece que hay x artículos con la característica de interés ya mencionada. La función de probabilidad de x es hipergeométrica, con media y varianza dadas por

$$\mu_x = n\left(\frac{D}{N}\right) = np$$

$$\sigma_x^2 = n\left(\frac{D}{N}\right)\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right)$$

También se ha analizado que cuando $np > 5$ para $p < 0.5$, la distribución hipergeométrica se puede aproximar bastante bien por medio de una distribución normal

$$x \sim N\left(\mu_x = np,\ \sigma_x = \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}\right)$$

Suponga ahora que se tienen dos líneas de producción, cada una de ellas con cierta fracción de artículos defectuosos p1 y p2

$$x_1 \sim N\left(\mu_x = n_1p_1,\ \sigma_x = \sqrt{n_1p_1(1-p_1)\left(\frac{N_1-n_1}{N_1-1}\right)}\right)$$

$$x_2 \sim N\left(\mu_x = n_2p_2,\ \sigma_x = \sqrt{n_2p_2(1-p_2)\left(\frac{N_2-n_2}{N_2-1}\right)}\right)$$

Se desea darle mantenimiento a ambas líneas de producción, pero conviene empezar por aquella que presente mayor fracción de artículos defectuosos, por lo cual es importante saber si tienen la misma fracción de defectuosos o una de ellas está peor, ante lo cual, conviene calcular un intervalo de confianza de la diferencia entre fracciones defectuosas. Partiendo de la hipótesis inicial de que ambas líneas presentan un número de defectuosos con distribución normal, se reitera que la diferencia entre dos normales da a su vez una distribución normal.

Dado que ambas son normales

$$p_1 \sim N\left(\hat{p}_1, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}\right)$$

$$p_2 \sim N\left(\hat{p}_2, \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\right)$$

$$p_1 - p_2 \sim N\left(\mu_{p1-p2} = \hat{p}_1 - \hat{p}_2, \sigma_{p1-p2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\right)$$

Por lo cual, un intervalo de confianza bilateral, superior e inferior para la diferencia de proporción de eventos exitosos de dos poblaciones finitas con distribución normal está dado por la expresión:

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2)$$

$$+ z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.136}$$

$$p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.137}$$

$$(\hat{p}_1 - \hat{p}_2) + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)} \leq p_1 - p_2 \tag{4.138}$$

Si ambas poblaciones son infinitas:

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2)$$

$$+ z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{4.139}$$

$$p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{4.140}$$

$$(\hat{p}_1-\hat{p}_2)+z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\leq p_1-p_2 \tag{4.141}$$

Si en la expresión 4.122 se define el error de estimación como:

$$\varepsilon\leq z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$$

Suponiendo que se elige el muestreo de tal forma que los tamaños de muestra en ambas poblaciones son iguales $n_1 = n_2 = n$, y despejando el valor de n, se obtiene una expresión para determinar de qué tamaño debe ser la muestra para obtener un intervalo de confianza bilateral, para la diferencia entre proporciones de dos poblaciones finitas con distribución normal:

$$n\geq\frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)N_1}{(N_1-1)}+\frac{\hat{p}_2(1-\hat{p}_2)N_2}{(N_2-1)}\right]}{\left[\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2+\frac{\hat{p}_1(1-\hat{p}_1)}{(N_1-1)}+\frac{\hat{p}_2(1-\hat{p}_2)}{(N_2-1)}\right]} \tag{4.142}$$

Si las poblaciones fueran infinitas

$$n\geq\frac{p_1(1-p_1)+p_2(1-p_2)}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2} \tag{4.143}$$

Ejercicio 4.28

Suponga que en dos líneas de producción de rondanas se inspeccionaron a lo largo de un mes el número de rondanas defectuosas que presentaba cada línea de producción, resultando lo siguiente:

Línea	N	n	D
1	36000	500	65
2	40000	600	70

Estime un intervalo de confianza bilateral, superior e inferior de la diferencia entre las fracciones defectuosas de cada una de las líneas de producción, al 95% del nivel de confianza.

Utilizando las expresiones 4.132, 4.133 y 4.134 se obtienen los siguientes intervalos

$-0.0255 \leq p_1 - p_2 \leq 0.05215$ intervalo bilateral al 95% de nivel de confianza

$p_1 - p_2 \leq 0.04591$ intervalo inferior al 95% de nivel de confianza

$-0.0255 \leq p_1 - p_2$ intervalo superior al 95% de nivel de confianza

Como se puede apreciar el intervalo bilateral contiene al cero, por lo que no existe evidencia estadística que demuestre que una línea de producción está peor que otra; si se pretende darle mantenimiento a ambas no importa dónde se empiece si en la uno o en la dos.

En el anterior ejercicio, suponga que se desea tener un error de estimación menor a 0.02 en el intervalo bilateral de confianza; suponiendo que se toman tamaños de muestras iguales en ambas líneas de producción, ¿de qué tamaño tendría que ser la muestra?

De la expresión 4.138

$n > 1967.87$

4.7.8. Intervalo de confianza de la diferencia entre las fracciones de defectos, éxitos, ocurrencias o llegadas por unidad de dos poblaciones normales

Suponga que dos empresas fabrican el mismo tipo y modelo de automóvil. Un automóvil tiene demasiadas componentes y cada componente puede presentar uno o más defectos, de tal manera que una flota de n automóviles puede presentar x defectos. Sean x_1 y x_2 el número de defectos que presentan n automóviles de la compañía uno y de la dos respectivamente. La función de probabilidad del número de defectos por unidad presenta función de probabilidad de Poisson.

Si x representa el promedio de defectos por cada n automóviles, entonces $u = x/n$ representa la fracción de defectos por cada unidad. También se ha afirmado que, si c representa el promedio de defectos en n unidades, para $c > 5$ la función de Poisson se aproxima a una normal, por lo cual, las variables x_1 y x_2 pueden suponerse normales y por lo mismo, las variables c_1 y c_2 que representan el número de defectos en n unidades también se pueden suponer normales

$$x_1 \sim N\left[\hat{c}_1, \sqrt{\hat{c}_1}\right]$$

$$x_2 \sim N\left[\hat{c}_2, \sqrt{\hat{c}_2}\right]$$

De tal manera que la diferencia entre ellas también es normal y por lo mismo

$$-z_{\frac{\alpha}{2}} \leq \frac{(x_1 - x_2) - (\hat{c}_1 - \hat{c}_2)}{\sqrt{\hat{c}_1 + \hat{c}_2}} \leq z_{\frac{\alpha}{2}}$$

Lo cual implica que los intervalos bilateral, inferior y superior de confianza al $(1-\alpha)100\%$ de nivel de confianza para la diferencia de defectos en n unidades entre ambas compañías con poblaciones infinitas normales está dado por

$$(\hat{c}_1 - \hat{c}_2) - z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1 + \hat{c}_2} \leq (c_1 - c_2) \leq (\hat{c}_1 - \hat{c}_2) + z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1 + \hat{c}_2} \tag{4.144}$$

$$(c_1 - c_2) \leq (\hat{c}_1 - \hat{c}_2) + z_{\alpha}\sqrt{\hat{c}_1 + \hat{c}_2} \tag{4.145}$$

$$(\hat{c}_1 - \hat{c}_2) - z_{\alpha}\sqrt{\hat{c}_1 + \hat{c}_2} \leq (c_1 - c_2) \tag{4.146}$$

Si las poblaciones fueran finitas

$$\left(\hat{c}_1-\hat{c}_2\right)-z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1\left(\frac{N_1-n_1}{N_1-1}\right)+\hat{c}_2\left(\frac{N_2-n_2}{N_2-1}\right)}\leq\left(c_1-c_2\right)\leq\left(\hat{c}_1-\hat{c}_2\right)+z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1\left(\frac{N_1-n_1}{N_1-1}\right)+\hat{c}_2\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.147}$$

$$\left(c_1-c_2\right)\leq\left(\hat{c}_1-\hat{c}_2\right)+z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1\left(\frac{N_1-n_1}{N_1-1}\right)+\hat{c}_2\left(\frac{N_2-n_2}{N_2-1}\right)} \tag{4.148}$$

$$\left(\hat{c}_1-\hat{c}_2\right)-z_{\frac{\alpha}{2}}\sqrt{\hat{c}_1\left(\frac{N_1-n_1}{N_1-1}\right)+\hat{c}_2\left(\frac{N_2-n_2}{N_2-1}\right)}\leq\left(c_1-c_2\right) \tag{4.149}$$

Para el caso de los defectos, ocurrencias, llegadas o éxitos por unidad

$$u_1\sim N\left(\hat{u}_1,\sqrt{\frac{\hat{u}_1}{n_1}}\right)$$

$$u_2\sim N\left(\hat{u}_2,\sqrt{\frac{\hat{u}_2}{n_2}}\right)$$

De tal manera que la diferencia entre ellas, u1-u2, también es normal y por lo mismo

$$-z_{\frac{\alpha}{2}}\leq\frac{\left(u_1-u_2\right)-\left(\hat{u}_1-\hat{u}_2\right)}{\sqrt{\frac{\hat{u}_1}{n_1}+\frac{\hat{u}_2}{n_2}}}\leq z_{\frac{\alpha}{2}}$$

Lo cual implica que un intervalo bilateral de confianza para la diferencia de defectos por unidad entre ambas compañías con poblaciones infinitas normales está dado por

$$\left(\hat{u}_1-\hat{u}_2\right)-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{u}_1}{n_1}+\frac{\hat{u}_2}{n_2}}\leq\left(u_1-u_2\right)\leq\left(\hat{u}_1-\hat{u}_2\right)+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{u}_1}{n_1}+\frac{\hat{u}_2}{n_2}} \tag{4.150}$$

El intervalo inferior de confianza para la diferencia de defectos por unidad entre ambas compañías con poblaciones normales está dado por:

$$\left(u_1-u_2\right)\leq\left(\hat{u}_1-\hat{u}_2\right)+z_{\alpha}\sqrt{\frac{\hat{u}_1}{n_1}+\frac{\hat{u}_2}{n_2}} \tag{4.151}$$

El intervalo superior de confianza para la diferencia de defectos por unidad entre ambas compañías con poblaciones normales está dado por:

$$(\hat{u}_1 - \hat{u}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1} + \frac{\hat{u}_2}{n_2}} \leq (u_1 - u_2) \tag{4.152}$$

Si las poblaciones fueran finitas

$$(\hat{u}_1 - \hat{u}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \leq (u_1 - u_2) \leq (\hat{u}_1 - \hat{u}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \tag{4.153}$$

$$(u_1 - u_2) \leq (\hat{u}_1 - \hat{u}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \tag{4.154}$$

$$(\hat{u}_1 - \hat{u}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \leq (u_1 - u_2) \tag{4.155}$$

Si se define el error de estimación del intervalo bilateral como:

$$\varepsilon \leq z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

Al pretender diseñar un plan de muestreo para obtener un intervalo de confianza para la diferencia entre el número o fracción de defectos, ocurrencias, llegadas o éxitos por unidad en dos poblaciones normales finitas, se puede establecer que $n_1 = n_2 = n$, por lo que despejando n:

$$n \geq \frac{\frac{N_1 u_1}{N_1 - 1} + \frac{N_2 u_2}{N_2 - 1}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2 + \frac{u_1}{N_1 - 1} + \frac{u_2}{N_2 - 1}} \tag{4.156}$$

Si las poblaciones son infinitas:

$$n \geq \frac{u_1 + u_2}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2} \tag{4.157}$$

Ejercicio 4.29

El control de calidad de dos líneas de manufactura de circuitos impresos se lleva a cabo inspeccionando el número de defectos por unidad de superficie (arañazos, bandas incorrectas, grosor no uniforme, etcétera). Suponga que la producción diaria de cada línea es de 6000 placas la primera y 7500 placas la segunda. Para llevar a cabo la inspección de los circuitos se utiliza una plantilla o retícula de diversos tamaños. Para cada línea de producción se registra la superficie de la retícula usada y el número de defectos que se observan. Tras inspeccionar 12 placas de la primera línea y 15 de la segunda, se obtienen los datos de la tabla que se muestra a continuación.

Línea	1		2	
Placa	Superficie cm^2	Nº de defectos	Superficie cm^2	Nº de defectos
1	50	4	34	2
2	50	3	38	4
3	34	4	25	3
4	38	4	44	5
5	54	4	38	2
6	22	3	44	2
7	22	5	44	4
8	25	3	38	4
9	50	4	22	4
10	34	2	54	4
11	34	2	50	3
12	38	4	38	4
13			44	4
14			32	3
15			25	2

a. Calcular los intervalos de confianza bilateral, inferior y superior al 95% de nivel de confianza de la diferencia entre las fracciones de defectos por cm^2, de cada línea de producción de circuitos impresos y determine cuál de ellas presenta mayor número de defectos por cm^2.

Línea	u^	n	N
1	0.093126	12	6000
2	0.087719	15	7500

Utilizando la expresión 4.149, el intervalo bilateral de confianza al 95% de nivel de confianza es

$$-0.2230 \leq \hat{u}_1 - \hat{u}_2 \leq 0.2338$$

Nótese que este intervalo contiene al cero, lo que implica que al 95% de nivel de confianza se puede considerar que ambas líneas de producción presentan los mismos defectos por cm^2.

Con la expresión 4.150 se obtiene el intervalo superior de confianza al 95% de nivel de confianza

$$-0.1863 \leq \hat{u}_1 - \hat{u}_2$$

Usando la expresión 4.151 se determina el intervalo inferior de confianza al 95% de nivel de confianza

$$\hat{u}_1 - \hat{u}_2 \leq 0.1971$$

b. Suponiendo que se llegaran a inspeccionar el mismo número de placas en cada línea, ¿de qué tamaño debe ser la muestra de placas para que el error de estimación sea menor a 0.2?

Con la expresión 4.152 se calcula el tamaño que debe tener la muestra:

$$n \geq 17.325$$

Figura 4.11. Intervalo Bilateral de Confianza para Dos Poblaciones

Variable	Condiciones iniciales	Estimador Puntual	Intervalo Bilateral de Confianza al (1-α)100%	Error	Toma de Decisiones
$\frac{\sigma_1^2}{\sigma_2^2}$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$	$\frac{S_1^2}{S_2^2}$	$\frac{S_1^2}{S_2^2}F_{\left(\frac{\alpha}{2},n_2-1,n_1-1\right)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\left(1-\frac{\alpha}{2},n_2-1,n_1-1\right)}$	$E = \frac{S_1^2}{S_2^2}F_{\left(1-\frac{\alpha}{2},n_2-1,n_1-1\right)}$	Si el intervalo contiene al uno entonces las varianzas poblacionales son iguales, de lo contrario son diferentes
$\frac{\sigma_1^2}{\sigma_2^2}$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas*	$\frac{S_1^2}{S_2^2}$	$\frac{S_1^2}{S_2^2}\left(\frac{\frac{N_1-n_1}{N_1-1}}{\frac{N_2-n_2}{N_2-1}}\right)F_{\left(\frac{\alpha}{2},n_2-1,n_1-1\right)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\left(\frac{\frac{N_1-n_1}{N_1-1}}{\frac{N_2-n_2}{N_2-1}}\right)F_{\left(1-\frac{\alpha}{2},n_2-1,n_1-1\right)}$	$E = \frac{S_1^2}{S_2^2}F_{\left(1-\frac{\alpha}{2},n_2-1,n_1-1\right)}$	Si el intervalo contiene al uno entonces las varianzas poblacionales son iguales, de lo contrario son diferentes
$p_2 - p_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$	$\hat{p}_2 - \hat{p}_1$	$(\hat{p}_2-\hat{p}_1) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \le p_2 - p_1 \le (\hat{p}_2-\hat{p}_1) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$	$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$	Si el intervalo contiene al cero entonces las fracciones defectuosas poblacionales son iguales, de lo contrario son diferentes
$p_2 - p_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas*	$\hat{p}_2 - \hat{p}_1$	$(\hat{p}_2-\hat{p}_1) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)} \le p_2 - p_1 \le (\hat{p}_2-\hat{p}_1) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}$	$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}$	Si el intervalo contiene al cero entonces las fracciones defectuosas poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$ σ_1^2,σ_2^2 *conocidas*	$\bar{x}_2 - \bar{x}_1$	$(\bar{x}_2-\bar{x}_1) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_2^2}{n_2}+\frac{\sigma_1^2}{n_1}} \le (\mu_2-\mu_1) \le (\bar{x}_2-\bar{x}_1) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_2^2}{n_2}+\frac{\sigma_1^2}{n_1}}$	$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_2^2}{n_2}+\frac{\sigma_1^2}{n_1}}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas* σ_1^2,σ_2^2 *conocidas*	$\bar{x}_2 - \bar{x}_1$	$(\bar{x}_2-\bar{x}_1) - z_{\frac{\alpha}{2}}\sqrt{\left[\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right]} \le (\mu_2-\mu_1) \le (\bar{x}_2-\bar{x}_1) + z_{\frac{\alpha}{2}}\sqrt{\left[\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right]}$	$E = z_{\frac{\alpha}{2}}\sqrt{\left[\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right]}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$ $\sigma_1^2 = \sigma_2^2$ *desconocidas*	$\bar{x}_2 - \bar{x}_1$	$(\bar{x}_2-\bar{x}_1) - t_{\left(\frac{\alpha}{2},n_2+n_1-2\right)}\sqrt{\frac{(n_2-1)S_2^2+(n_1-1)S_1^2}{n_2+n_1-2}}\sqrt{\frac{1}{n_2}+\frac{1}{n_1}} \le (\mu_2-\mu_1) \le (\bar{x}_2-\bar{x}_1) + t_{\left(\frac{\alpha}{2},n_2+n_1-2\right)}\sqrt{\frac{(n_2-1)S_2^2+(n_1-1)S_1^2}{n_2+n_1-2}}\sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$	$E = t_{\left(\frac{\alpha}{2},n_2+n_1-2\right)}\sqrt{\frac{(n_2-1)S_2^2+(n_1-1)S_1^2}{n_2+n_1-2}}\sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas* $\sigma_1^2 = \sigma_2^2$ *desconocidas*	$\bar{x}_2 - \bar{x}_1$	$(\mu_2-\mu_1) = (\bar{x}_2-\bar{x}_1) \pm t_{\left(\frac{\alpha}{2},n_2+n_1-2\right)}\sqrt{\frac{(n_2-1)\left(\frac{N_2-n_2}{N_2-1}\right)S_2^2+(n_1-1)\left(\frac{N_1-n_1}{N_1-1}\right)S_1^2}{n_2+n_1-2}}\sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$	$E = t_{\left(\frac{\alpha}{2},n_2+n_1-2\right)}\sqrt{\frac{(n_2-1)\left(\frac{N_2-n_2}{N_2-}\right)S_2^2+(n_1-1)\left(\frac{N_2-n_2}{N_2-}\right)S_1^2}{n_2+n_1-2}}\sqrt{\frac{1}{n_2}+\frac{1}{n_1}}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$ $\sigma_1^2 \ne \sigma_2^2$ *desconocidas*	$\bar{x}_2 - \bar{x}_1$	$(\bar{x}_2-\bar{x}_1) - t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}+\frac{S_1^2}{n_1}} \le (\mu_2-\mu_1) \le (\bar{x}_2-\bar{x}_1) + t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}+\frac{S_1^2}{n_1}}$ $\nu = \frac{\left(\frac{S_2^2}{n_2}+\frac{S_1^2}{n_1}\right)^2}{\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}+\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1}} - 2$	$E = (\bar{x}_2-\bar{x}_1) + t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}+\frac{S_1^2}{n_1}}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas* $\sigma_1^2 \ne \sigma_2^2$ *desconocidas*	$\bar{x}_2 - \bar{x}_1$	$(\bar{x}_2-\bar{x}_1) - t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)} \le (\mu_2-\mu_1) \le (\bar{x}_2-\bar{x}_1) + t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}$ $\nu = \frac{\left(\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right)^2}{\frac{\left(\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)\right)^2}{n_2+1}+\frac{\left(\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N1-1}\right)\right)^2}{n_1+1}} - 2$	$E = (\bar{x}_2-\bar{x}_1) + t_{\left(\frac{\alpha}{2},\nu\right)}\sqrt{\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)+\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_0 = \mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ $N_1,N_2 \to \infty$ Muestras Apareadas	$D_i = x_{2i} - x_{1i} \quad i = 1,2,\cdots,n$	$\overline{D} - t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}} \le \mu_D \le \overline{D} + t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}}$	$E = \overline{D} + t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}}$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes
$\mu_0 = \mu_2 - \mu_1$	$x_1 \sim N(\mu_1,\sigma_1)$ $x_2 \sim N(\mu_2,\sigma_2)$ N_1,N_2 *Finitas* Muestras Apareadas	$D_i = x_{2i} - x_{1i} \quad i = 1,2,\cdots,n$	$\overline{D} - t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right) \le \mu_D \le \overline{D} + t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)$	$E = \overline{D} + t_{\left(\frac{\alpha}{2},n-1\right)}\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)$	Si el intervalo contiene al cero entonces las medias poblacionales son iguales, de lo contrario son diferentes

Figura 4.12. Fórmulas para estimar el tamaño de muestra necesario para deducir un intervalo de confianza para la diferencia de un parámetro de dos poblaciones con una muestra simple

Parámetros	Varianza del Estimador	Límite Error de Estimación	Tamaño de Muestra
Diferencia de Medias μ_1-μ_2	$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$	$n \geq \frac{\frac{N_1\sigma_1^2}{N_1-1}+\frac{N_2\sigma_2^2}{N_2-1}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2+\left(\frac{\sigma_1^2}{N_1-1}+\frac{\sigma_2^2}{N_2-1}\right)}$
Diferencia en la Fracciones Defectuosas p_1-p_2	$\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)$	$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$	$n \geq \frac{\left[\frac{\hat{p}_1(1-\hat{p}_1)N_1}{(N_1-1)}+\frac{\hat{p}_2(1-\hat{p}_2)N_2}{(N_2-1)}\right]}{\left[\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2+\frac{\hat{p}_1(1-\hat{p}_1)}{(N_1-1)}+\frac{\hat{p}_2(1-\hat{p}_2)}{(N_2-1)}\right]}$
Diferencia en las Fracciones de Defectos por Unidad u_1-u_2	$\sigma^2_{\hat{u}_1-\hat{u}_2} = \frac{\hat{u}_1}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{u}_2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)$	$\varepsilon \leq z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{u}_2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$	$n \geq \frac{\frac{N_1u_1}{N_1-1}+\frac{N_2u_2}{N_2-1}}{\left(\frac{\varepsilon}{z_{\frac{\alpha}{2}}}\right)^2+\frac{u_1}{N_1-1}+\frac{u_2}{N_2-1}}$

Ejercicios del capítulo 4

1. Sean los siguientes estadísticos:

$\hat{\mu}_1 = \bar{x}$ $\hat{\mu}_2 = mediana$

$\hat{\mu}_3 = moda$ $\hat{\mu}_4 = semirango$

$\hat{\mu}_5 = \frac{x_1 + 2x_2 + 3x_3}{6}$ $\hat{\mu}_6 = \frac{x_1 - 4x_2}{-3}$

$\hat{\mu}_7 = \frac{x_{1+}x_n + n}{n}$

   a. Obtenga cuáles de ellos son insesgados y cuáles son sesgados.
   b. Determine cuál de los siete es el menos eficiente y cuál es el óptimo.
   c. Explique el razonamiento a seguir para determinar cuáles son consistentes y cuáles no lo son.

2. Se recibe un lote de $N = 9000$ resortes de tracción utilizados para interruptores de flotador, los cuáles deben presentar una carga especificada de 6.5 ± 0.25 libras. Para proceder a aceptar el lote se realiza un plan de muestreo aleatorio simple, con un tamaño de muestra de n = 90 resortes, los cuales son probados con un medidor de cargas, arrojando los siguientes resultados.

6.87	6.82	6.76	6.79	6.52	6.90	6.70	7.07	6.63
6.82	6.51	6.45	6.59	6.71	6.65	6.88	6.81	6.52
6.74	6.78	6.70	6.82	6.85	6.80	6.79	6.82	6.58
7.12	6.74	6.80	6.60	6.82	7.18	6.89	6.60	6.77
6.82	6.67	7.06	6.69	6.42	6.51	6.72	6.96	6.96
6.60	7.16	6.99	6.78	6.86	7.00	6.64	6.78	6.57
6.53	6.84	6.58	6.63	6.72	6.47	6.75	6.69	6.48
6.99	6.48	6.90	6.67	6.77	6.54	6.82	6.75	6.63
6.68	6.70	6.47	6.98	6.69	6.94	6.71	6.49	6.94
6.73	6.71	6.87	6.37	7.05	6.79	6.97	6.72	6.88

a. Obtenga un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para el promedio en la carga del lote de resortes.
b. Calcule un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para la varianza en la carga del lote de resortes.
c. Obtenga un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para la fracción de resortes en el lote que presenta una carga por arriba de 6.75 libras.

3. Una oficina expendedora de pasaportes en la Ciudad de México establece un control de los errores que cometen al tramitar 500 pasaportes diariamente, los cuales operan por cita. Para analizar su desempeño decide recopilar una muestra de $n = 20$ días. Cabe señalar que los errores que cometen son muy diversos, foto borrosa, foto equivocada, errores de tipografía, fecha de nacimiento equivocada, fecha de emisión equivocada, nombre mal escrito, tiempo de espera excedido, tiempo de emisión excedido, poca cortesía al recibir a los ciudadanos, trato descortés, personas que se cuelan a la fila, etcétera.

   La oficina diariamente levanta una encuesta y le solicita a cada cliente le indique los errores cometidos, los cuales pueden ser críticos, mayores, menores, etcétera, nótese que el número de errores puede ser mayor que el número de pasaportes emitidos, ya que cada pasaporte puede presentar desde cero hasta una cantidad incontable de errores. Suponga que los errores o defectos durante esos $n = 20$ días fueron los siguientes:149, 150, 154, 139, 145, 152, 142, 142, 154, 138, 148, 130, 149, 143, 151, 131, 130, 115, 159, 149. Obtenga un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, del número de errores que se cometen por día en dicha oficina.

4. En un puerto marítimo el tiempo de descarga en horas de un barco carguero es crítico. Existe una cláusula en el contrato de la compañía naviera con el puerto, de no tardarse más de 20 horas en la descarga. Por cada hora de retraso, un barco le cobrará al puerto $50,000.00. El puerto decide realizar un análisis del tiempo que le lleva descargar cada barco y recolecta una muestra aleatoria de $n = 60$ barcos a lo largo de un mes. Los resultados en horas se muestran a continuación:

6.2	2.5	13.8	2.3	20.6	0
8	10.7	9.7	16.1	4.8	4.1
4.8	4.4	14	1.1	1	32.1
15.1	0.6	4	25	4.4	7.7
15.6	7.9	10.7	29.4	11.7	0.9
4.3	14.5	1.7	6.5	6.1	7.3
1.7	39	2.8	14.9	11.8	2.2
25.7	5.9	4.7	3.9	3.2	2.5
12.6	8.2	0.9	6.5	0.5	3.8
12.6	12.5	7.1	0.4	3.3	23.3

a. Obtenga un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para el promedio de horas de descarga.
b. Calcule un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para la varianza en el número de horas de descarga de cada barco.
c. Obtenga un intervalo de confianza bilateral, unilateral inferior y unilateral superior, al 95% y al 99% de nivel de confianza, para la fracción de barcos a los que se les pagará.
d. ¿Cuánto se pagará en promedio?
e. Revise las hipótesis básicas para el intervalo de confianza de los incisos anteriores, ¿considera usted que se cumplen?

5. La durabilidad de una llanta se mide con base en el número promedio de kilómetros recorridos bajo condiciones controladas, hasta que aparezcan bandas de desgaste con una profundidad de rodadura de 1.6 mm de espesor o menos. Los compuestos de una llanta están diseñados para funcionar idealmente por un máximo de 5 años. El siguiente ejemplo es hipotético y se utilizan datos generados aleatoriamente que no son reales de las marcas que se mencionan, pero es ilustrativo del empleo que debe darse a los intervalos de confianza para dos poblaciones. Suponga que se desea adquirir un lote de $N = 1600$ llantas 245/40r20 Run Flat 99y. Para ello, existen dos marcas diferentes Pirelli y Michelin. Se realiza una prueba de durabilidad, usando ocho coches, Mercedes Benz C200, del mismo modelo y año, con las mismas condiciones de uso, con la misma carga, recién alineados y balanceados bajo las mismas especificaciones, al primer equipo A de cuatro coches se les montan cuatro llantas Pirelli y al otro equipo B de cuatro coches, llantas

Michelin. Se ponen a rodar bajo las mismas condiciones hasta que aparecen bandas de desgaste con una profundidad de rodadura de 1.6 mm de espesor o menos. Los resultados se muestran a continuación.

a. Obtenga un intervalo de confianza para la diferencia entre medias, de las marcas citadas, al 95% y al 99% de nivel de confianza, suponiendo que las varianzas poblacionales son conocidas s12=5600, s22=5400.
b. Determine un intervalo de confianza para la diferencia entre medias de las marcas citadas, al 95% y al 99% de nivel de confianza, suponiendo varianzas no conocidas pero iguales. Primero, obtenga un intervalo de confianza para el cociente entre varianzas y posteriormente un intervalo de confianza para la diferencia entre las medias de las marcas citadas.
c. Calcule un intervalo de confianza para la diferencia entre medias de las marcas citadas, al 95% y al 99% de nivel de confianza, suponiendo varianzas no conocidas y diferentes.
d. Obtenga un intervalo de confianza para la diferencia entre medias de las marcas citadas, al 95% y al 99% de nivel de confianza, suponiendo que las mediciones fueron apareadas, usando los mismos coches de una marca de llanta a otra y suponiendo las mismas condiciones de prueba.
e. Obtenga un intervalo de confianza para la diferencia entre fracciones de llantas que duran más de 52,000 Km. Conceptualmente qué hipótesis básica no cumple este intervalo de confianza.

Michelin	Pirelli
50,133.93	60,014.73
49,954.33	54,336.33
49,164.83	54,940.62
51,300.54	52,895.42
48,249.71	57,123.37
48,494.13	57,535.76
51,038.79	51,359.57
53,306.49	58,214.95
52,109.23	53,801.49
48,483.41	57,228.75
46,079.35	54,037.71
47,043.97	55,067.46
44,855.49	55,990.88
52,853.32	54,487.21
49,679.16	52,606.81
47,965.69	53,523.96

6. Una 'caricia' en este texto, es una unidad de reconocimiento, es una forma o manera de medir el reconocimiento, aprecio o afecto que se le tiene a una persona. Hay caricias, sinestésicas o físicas, verbales y gestuales. Otra clasificación de caricias es que existen caricias positivas, negativas y con descuento, estas últimas son aquellas a las que después de decirles lo positivo de algo, van acompañadas del término 'pero'.

   Suponga que una joven tiene dos pretendientes; ella trata de decidir cuál es el pretendiente más adecuado y para ello mide el afecto que cada uno de ellos le tiene, con base en el número de caricias que le prodiga por día, las caricias positivas las cuenta con números positivos, las negativas con números negativos y las con descuento les asigna el valor de cero, posteriormente las suma y obtiene un número que puede ser positivo, negativo o cero. Suponga que pone a prueba a sus pretendientes y contabiliza el número de caricias por día a lo largo de un mes, obteniendo los siguientes resultados.

   Obtenga un intervalo de confianza de la diferencia en el número de caricias promedio por cada punto de contacto que cada uno de los pretendientes le prodiga.

Pretendiente A		Pretendiente B	
Puntos de Contacto	No. Caricias	Puntos de Contacto	No. Caricias
4	5	1	3
7	5	4	4
6	5	3	2
8	9	3	2
10	6	7	3
7	6	5	3
7	3	2	5
1	5	5	1
11	2	8	0
4	4	4	5
8	1	3	1
5	7	7	1
10	5	8	3
7	4	6	0
3	3	9	3
14	2	2	3
5	7	5	4
3	4	3	4
3	7	3	3
6	4	3	4
4	6	3	1
5	5	3	4
6	3	2	1
6	3	6	1
9	3	8	4
9	5	7	4
6	5	5	2
10	2	4	5
8	3	6	2
5	5	5	3

5. Pruebas de hipótesis estadística

5.1. Hipótesis estadística

Una hipótesis, del griego hipo, 'subordinación' o 'por debajo', y tesis, 'conclusión que se mantiene con un razonamiento', es un enunciado no verificado, una conjetura científica, una afirmación o supuesto que requiere una contrastación y validación con la realidad a través de la recolección de información y datos. El nivel de veracidad que se otorga a una hipótesis dependerá de la medida en que los datos empíricos apoyan lo afirmado en la hipótesis. Esto es lo que se conoce como contrastación empírica de la hipótesis o bien el proceso de validación de la hipótesis.

Para formular una hipótesis se requiere observar un fenómeno, reunir información, comparar la información con el fenómeno, establecer posibles explicaciones, escoger la explicación más probable y enunciar una o más hipótesis. Posteriormente, se debe comprobar dicha hipótesis a través de la experimentación, en la que se confirma la hipótesis (si es verdadera) o no (si es falsa).

Una hipótesis estadística es una afirmación sobre el valor que toma un parámetro poblacional como puede ser su tendencia central, su dispersión, o la forma que toma una variable aleatoria, es decir, la distribución de probabilidad a la que obedece.

Una Prueba de Hipótesis es un procedimiento estadístico que permite aceptar o rechazar una afirmación hecha con respecto a un fenómeno o suceso y consta de los siguientes pasos:

a. Se plantean las hipótesis nula y alternativa
b. Se selecciona el nivel de significancia
c. Se identifica el estadístico de prueba

d. Se formula la regla de decisión
e. Se toma una muestra y se decide si se rechaza Ho o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

a. Se plantean las hipótesis nula H_0 y alternativa H_1

En una prueba de hipótesis generalmente se plantean dos hipótesis mutuamente excluyentes: la hipótesis nula o hipótesis de nulidad H_0 y la hipótesis alternativa H_1. La hipótesis nula es un punto de partida para la investigación que no se rechaza a menos que los datos de la muestra proporcionen evidencia estadística suficiente de que es falsa.

La hipótesis nula es una aplicación a la estadística del método de reducción al absurdo, por el cual se supone, en principio, lo contrario de lo que se desea probar, hasta que la evidencia o las conclusiones obtenidas demuestran que el punto de partida fue falso o absurdo y, por tanto, se rechaza y se concluye lo contrario aceptando la hipótesis alternativa (lo que se quería probar). La hipótesis alternativa es una afirmación especial cuya validez se pretende demostrar, y si las pruebas empíricas no apoyan decididamente la hipótesis alternativa, entonces se dice que no existe suficiente evidencia estadística para demostrarla y se termina por creer la hipótesis nula, aunque no se esté convencido de ello.

Ejercicio 5.1

Una empresa fabricante de rodamientos requiere para uno de sus procesos balines para balero de media pulgada de diámetro. Un proveedor de dicha empresa afirma que sus balines son de media pulgada en promedio. En este caso, la prueba de hipótesis contiene las siguientes:

$$H_0 : \mu_x = 0.5$$

$$H_1 : \mu_x \neq 0.5$$

A este tipo de pruebas se les conoce como bilaterales o de dos colas.

Ejercicio 5.2

Una empresa fabricante de elevadores o ascensores requiere cable o malacate de acero con una resistencia mínima de 3 toneladas. Un proveedor afirma que cumple la especificación. La empresa cliente decide hacer una prueba de hipótesis para ello y establece las siguientes:

$H_0: \mu_x = 3$

$H_1: \mu_x < 3$

A este tipo de pruebas se les conoce como unilaterales o de una cola y en particular a esta se le conoce como prueba unilateral inferior.

Ejercicio 5.3

Una empresa fabricante de instrumentos musicales requiere mandar sus instrumentos a mantenimiento, calibración y afinación. La empresa cliente requiere de un máximo tiempo de reparación de tres días. Un proveedor afirma que cumple la especificación. La empresa cliente decide hacer una prueba de hipótesis para ello y establece las siguientes:

$H_0: \mu_t = 3$

$H_1: \mu_t > 3$

A este tipo de pruebas se les conoce como unilaterales o de una cola y en particular a esta se le conoce como prueba unilateral inferior.

Ejercicio 5.4

Una empresa metalmecánica necesita adquirir botas de seguridad para sus empleados. Existen dos marcas de botas diferentes A y B. Dado que la marca A es más barata que la B, en caso de que la A tenga el mismo nivel de protección que la B, se propone comprar la A. Si la marca B ofrece un mayor nivel de seguridad que la A, la empresa se inclinará por la marca B. Se requiere determinar qué marca de bota comprar, considerando el nivel de protección que cada una de ellas ofrece. La empresa cliente decide hacer una prueba de hipótesis para ello y establece las siguientes:

$H_0: \mu_A = \mu_B$

$H_1: \mu_A < \mu_B$

Ejercicio 5.5

Se sospecha que el tiempo de vida t de un tipo particular de lámpara presenta un comportamiento probabilístico exponencial negativo. Se decide realizar una prueba de hipótesis para ello y se establecen las siguientes:

$H_0: t\_es\_exponencial\_negativa$

$H_1: t\_no\_es\_exponencial\_negativa$

b. Se selecciona el nivel de significancia α

La decisión para rechazar la hipótesis nula se basa en un estadístico de prueba a partir de los datos obtenidos de una muestra aleatoria. Siempre que se toma una decisión a partir de los datos obtenidos de una muestra, pueden aparecer dos tipos de error:

Error tipo I o Falso Positivo: Rechazar una afirmación que debió aceptarse, es decir en este caso, rechazar la hipótesis nula dado que es verdadera.

Error tipo II o Falso Negativo: Aceptar una afirmación que debió rechazarse, es decir en este caso, aceptar la hipótesis nula dado que es falsa.

Las probabilidades de ocurrencia de los errores tipo I y II, se representan por dos letras griegas:

$$\alpha = p\left(error\_tipo\_I\right) = p\left(rechazar\_H_0 \mid H_0\_es\_verdadera\right) \tag{5.1}$$

$$\beta = p\left(error\_tipo\_II\right) = p\left(aceptar\_H_0 \mid H_0\_es\_falsa\right) \tag{5.2}$$

A la probabilidad del error tipo I se le conoce como Nivel de Significancia a de una prueba estadística.

Son comunes los niveles de significancia del 0.10, 0.05, 0.01 y 0.0027.

En algunas situaciones es conveniente expresar la significancia estadística como su complemento 1 – a, a la cual se le conoce como nivel de confianza.

$$1 - \alpha = \mathrm{p}\left(\mathrm{aceptar}\_H_0 \mid H_0\_\mathrm{es}\_\mathrm{verdadera}\right) \tag{5.3}$$

De la misma forma, en ciertas ocasiones es mejor trabajar con el complemento de b, conocido como la Potencia de la Prueba:

$$\text{Potencia de la Prueba} = 1-\beta = \mathrm{p}(\text{rechazar } H_0 \mid H_0 \text{ es falsa}) \tag{5.4}$$

En la figura 5.1 se ilustra el nivel de significancia para una prueba de dos colas, para una distribución normal con media 50 y desviación estándar 5. En esta figura, el nivel de confianza representa el área bajo la curva 1–α situada entre

los límites de confianza, que en este caso en particular es 0.90 o 90%; el valor de α es 0.10 (0.05 en la cola izquierda y 0.05 en la cola derecha). Existen varios criterios para decidir aceptar o rechazar la hipótesis nula. En este caso particular, el primer criterio que podría usarse sería si el valor de x_0 cae entre los límites de confianza calculados, entonces se dice que se cumple la hipótesis nula. Si el valor de x_0 cayera antes del límite inferior de confianza LIC = 41.78 o si cayera después del límite superior de confianza LSC = 58.22, entonces se tendría que rechazar la hipótesis nula.

Otro criterio que podría usarse, en pruebas de límite superior sería calcular el área bajo la curva a la derecha de x_0, a la cual se le denominará p (*p-value*):

$$p = p(x > x_0) \tag{5.5}$$

En pruebas de límite inferior, sería calcular el área bajo la curva a la izquierda de x_0:

$$p = p(x < x_0) \tag{5.6}$$

Si la prueba es de dos colas como la que se muestra en la figura 5.1, para $p > \alpha/2$ entonces se tendría que dar por cierta la hipótesis nula y si $p < \alpha/2$ se rechazaría la hipótesis nula.

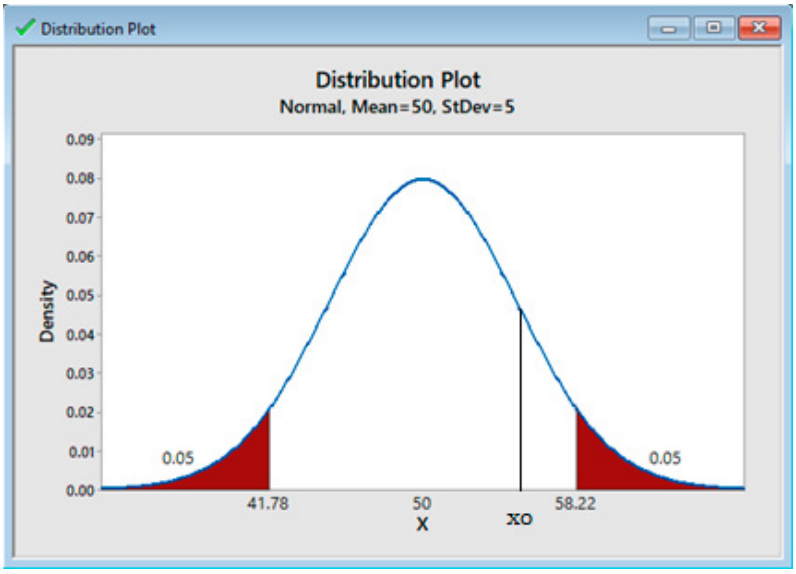


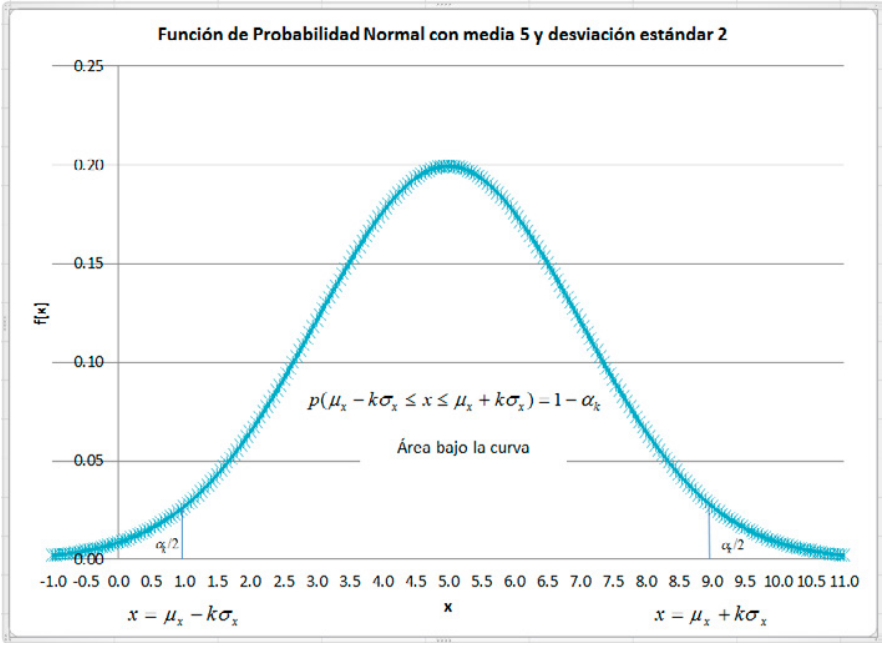
Figura 5.1

Para una prueba de una cola la comparación sería contra el valor de α. Si el valor *p* es inferior al nivel de significancia α, entonces la hipótesis nula es rechazada. Cuanto menor sea el valor p, más significativo será el resultado.

Valores pequeños de α otorgan mayor confianza en la determinación de la significancia, pero hacen correr mayores riesgos de equivocarse al aceptar una hipótesis nula falsa (error de tipo II o "falso negativo"), con lo cual se pierde potencia de estudio. La elección de un nivel de a inevitablemente envuelve un compromiso entre significancia y potencia, y consecuentemente entre errores de tipo I y de tipo II. En algunas disciplinas o áreas es común expresar la significancia estadística en k unidades de desviación estándar σ de una distribución normal, como se ilustra en la figura 5.2. A partir de esta figura y usando Excel se obtienen los siguientes valores para pruebas bilaterales o de dos colas:

α/2	k
0.308538	0.5
0.158655	1.0
0.066807	1.5
0.022750	2.0
0.006210	2.5
0.001350	3.0
0.000233	3.5
0.000032	4.0
0.000003	4.5

Figura 5.2



c. Se identifica el estadístico de prueba

El estadístico de prueba depende de la distribución de probabilidad que presente la variable sujeta a estudio. Por ejemplo, para el caso de una prueba de hipótesis para la media poblacional, el estadístico de prueba es z la variable aleatoria normal estándar para el caso de que la población sea normal o el tamaño de muestra sea grande o en su defecto, el estadístico t la variable aleatoria t de Student. Para el caso de la varianza poblacional, el estadístico de prueba es ji cuadrada. Para el caso del cociente entre varianzas el estadístico de prueba es F de Fisher-Snedeccor, para el caso de igualdad entre medias el estadístico de prueba es z la variable aleatoria normal estándar para el caso de que las poblaciones sean normales o los tamaños de muestra sean grandes o en su defecto, el estadístico t la variable aleatoria t de Student. Para los casos de pruebas de hipótesis de proporciones de defectuosos o de defectos el estadístico de prueba puede ser normal, binomial o Poisson. Para el caso de pruebas de hipótesis de bondad de ajuste, es decir, que el conjunto de datos recolectados se comporte con una distribución de probabilidad conocida, el estadístico de prueba depende de la distribución de probabilidad que presente la variable sujeta a estudio.

d. Se formula la regla de decisión

La regla de decisión es un enunciado que se emite para determinar si se rechaza o no la hipótesis nula. Específica el valor crítico de los resultados muestrales.

e. Se toma una muestra aleatoria

Se calcula el valor del estadístico de prueba, se compara contra el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

5.2. Pruebas de hipótesis de un parámetro para una población

5.2.1. Pruebas de hipótesis sobre la media de una población normal o tamaño de muestra muy grande, con varianza conocida

Suponga que x es normal $x \sim N(\mu_x, \sigma_x)$ o n es suficientemente grande para suponer que se cumple el teorema del límite central, es decir, $\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al $100(1-\alpha)\%$ de nivel de confianza, para la media poblacional; si dicho intervalo no contiene al valor μ_0 entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor μ_0 entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma_x / \sqrt{n}} \tag{5.7}$$

Para una prueba bilateral o de dos colas, donde H_1: $\mu_x \neq \mu_0$

$Si \mid z_0 \mid > z_{\frac{\alpha}{2}}$

Es decir:

$$z_0 > z_{\frac{\alpha}{2}} \quad \text{o} \quad z_0 < -z_{\frac{\alpha}{2}} \tag{5.8}$$

Se rechaza la hipótesis nula

De lo contrario, si

$$Si \qquad -z_{\frac{\alpha}{2}} < z_0 < z_{\frac{\alpha}{2}} \tag{5.9}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x < \mu_0$

Si $z_0 < z_\alpha$ se rechaza la hipótesis nula (5.10)

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x > \mu_0$

Si $z_0 < z_\alpha$ se rechaza la hipótesis nula (5.11)

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(z > z_0)$ o $p = p(z < -z_0)$.

Para una prueba bilateral o de dos colas, donde H_1: $\mu_x \neq \mu_0$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x < \mu_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x < \mu_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x \neq \mu_0$

Unilateral_Inferior (5.12)
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x \neq \mu_0$

Unilateral_Superior
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x > \mu_0$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, la prueba puede ser de dos colas $z_{\alpha/2}$, unilateral inferior za o unilateral superior $-z_{\alpha}$

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.6

Del ejercicio 4.21. Un proveedor afirma que la resistencia del concreto hidráulico que él fabrica es mayor a 2475 psi. En el área de recibo de una empresa constructora un ingeniero civil realiza la prueba a 50 probetas, obteniendo los siguientes datos en unidades de psi:

2243	2310	2281	2277	2272
2246	2271	2235	2261	2251
2320	2208	2268	2263	2295
2215	2270	2264	2241	2205
2234	2285	2256	2305	2223
2271	2244	2306	2281	2242
2287	2254	2212	2252	2258
2267	2268	2279	2304	2240
2257	2230	2263	2297	2270
2263	2290	2219	2224	2262

Suponga que la resistencia a la compresión de dichas probetas es normal y que se conoce su desviación estándar $\sigma_x = 30$ psi.

a. Realice una prueba de hipótesis estadística al 95 y 99% de nivel de confianza de la resistencia promedio del concreto a la compresión.

H_0: $\mu_x = 2475$
H_1: $\mu_x < 2475$

Nótese que la prueba es de una cola. Los valores críticos son:

$z_{0.05} = 1.64485$
$z_{0.01} = 2.32635$

Con el primer criterio, se calcula un intervalo inferior de confianza:

$\mu_x < 2267.76$ al 95% de nivel de confianza
$\mu_x < 2270.65$ al 99% de nivel de confianza

En ambos casos no contiene al valor que establece el proveedor, por lo que se rechaza contundentemente su afirmación.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$z_0 = (2260.78\text{-}2475)/(30/\text{raíz}(50)) = -50.49$

Nótese que, en ambos casos, $z_0 < -z_\alpha$, por lo que se rechaza la hipótesis nula del proveedor.

Con el tercer criterio, se calcula la probabilidad de que $z < -z_0$

$p = p(z < -z_0) = 0 < 0.01 < 0.05$, por lo tanto, se rechaza la hipótesis nula del proveedor.

b. ¿De qué tamaño debe ser la muestra para suponer que el error de estimación sea menor a 5 psi, para el intervalo inferior al 95% de nivel de confianza?

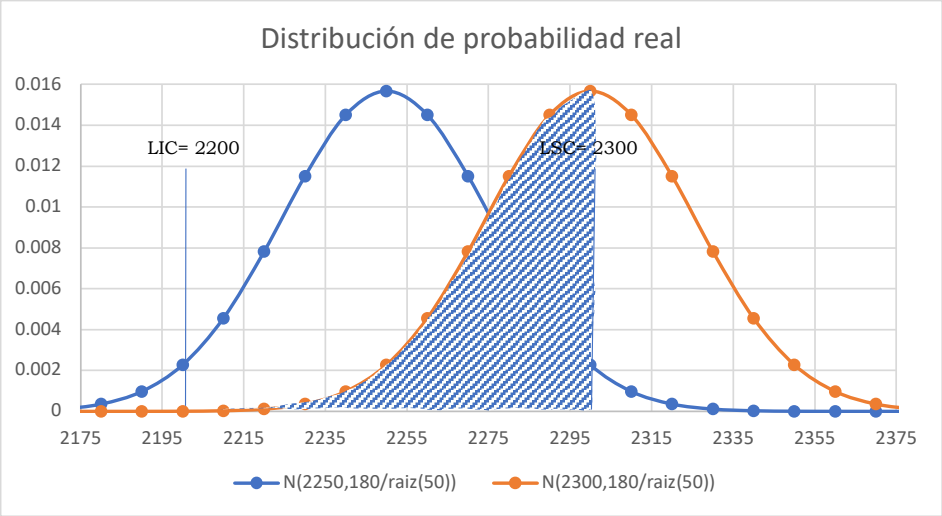
$$n \geq \frac{\hat{\sigma}_x^2}{\left(\frac{\varepsilon}{z_\alpha}\right)^2} = \frac{(30)^2}{\left(\frac{5}{1.64485}\right)^2} = 97.3991$$

Ejercicio 5.7

Una empresa constructora adquiere concreto, el cual debe tener una media de resistencia a la compresión de 2250 psi. La desviación estándar del concreto históricamente es de 180 psi. Para aprobar cada lote toma una muestra de $n = 50$ cilindros de concreto. Se acepta cada lote si la media estimada se encuentra entre 2200 y 2300 psi.

a. Formule la expresión matemática para aceptar que la media es de 2250 psi, dado que la media real se encuentra en el intervalo $2000 \leq \mu_x \leq 2350$ psi.

Figura 5.3



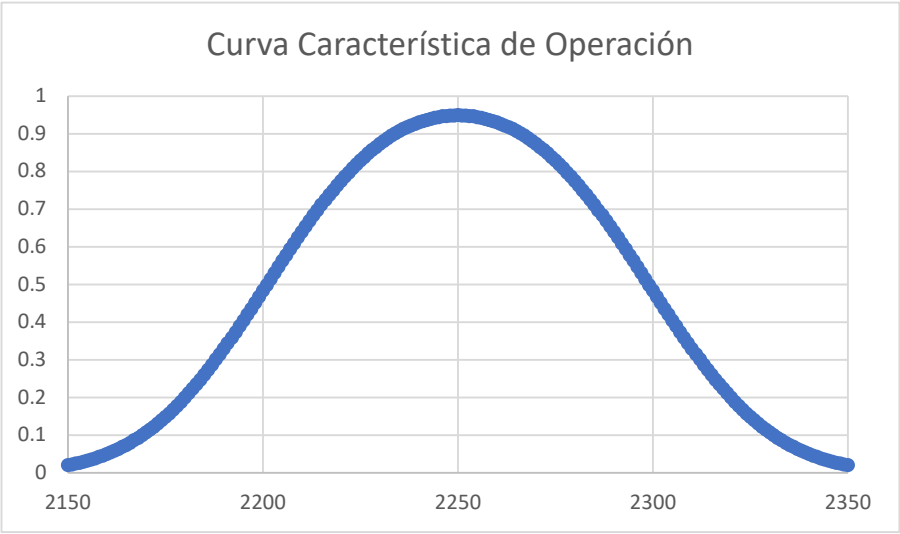
Tal como se muestra en la figura 5.3, la probabilidad de aceptar que la media es de 2250 psi es la parte que se muestra entramada, es decir,

$$\beta = p(2000 \leq \mu_x \leq 2300 \mid \mu_x)$$

μ_x	β	μ_x	β	μ_x	β	μ_x	β	μ_x	β
2150	0.0206752	2190	0.3299621	2230	0.8740855	2270	0.8740855	2310	0.3299621
2151	0.0227501	2191	0.3445705	2231	0.8816662	2271	0.866088	2311	0.3156083
2152	0.0249979	2192	0.3594142	2232	0.8888312	2272	0.8576732	2312	0.3015273
2153	0.0274289	2193	0.374473	2233	0.8955821	2273	0.8488417	2313	0.287736
2154	0.030054	2194	0.3897254	2234	0.9019213	2274	0.8395946	2314	0.27425
2155	0.0328841	2195	0.4051492	2235	0.9078514	2275	0.8299342	2315	0.2610837
2156	0.0359303	2196	0.4207213	2236	0.9133756	2276	0.8198637	2316	0.2482501
2157	0.0392039	2197	0.436418	2237	0.9184972	2277	0.8093875	2317	0.2357608
2158	0.0427162	2198	0.4522148	2238	0.9232197	2278	0.7985108	2318	0.2236259
2159	0.0464786	2199	0.468087	2239	0.927547	2279	0.7872403	2319	0.2118542
2160	0.0505026	2200	0.4840091	2240	0.9314826	2280	0.7755839	2320	0.2004532
2161	0.0547993	2201	0.4999557	2241	0.9350303	2281	0.7635504	2321	0.1894289
2162	0.0593799	2202	0.5159012	2242	0.9381936	2282	0.7511501	2322	0.1787857
2163	0.0642555	2203	0.5318199	2243	0.9409759	2283	0.7383947	2323	0.1685271
2164	0.0694366	2204	0.5476861	2244	0.9433803	2284	0.7252968	2324	0.1586548
2165	0.0749337	2205	0.5634745	2245	0.9454098	2285	0.7118706	2325	0.1491696
2166	0.0807566	2206	0.5791601	2246	0.9470667	2286	0.6981313	2326	0.1400708
2167	0.0869149	2207	0.5947183	2247	0.9483531	2287	0.6840954	2327	0.1313566
2168	0.0934174	2208	0.6101249	2248	0.9492708	2288	0.6697807	2328	0.1230242
2169	0.1002725	2209	0.6253567	2249	0.9498209	2289	0.655206	2329	0.1150695
2170	0.1074876	2210	0.640391	2250	0.9500042	2290	0.640391	2330	0.1074876
2171	0.1150695	2211	0.655206	2251	0.9498209	2291	0.6253567	2331	0.1002725
2172	0.1230242	2212	0.6697807	2252	0.9492708	2292	0.6101249	2332	0.0934174
2173	0.1313566	2213	0.6840954	2253	0.9483531	2293	0.5947183	2333	0.0869149
2174	0.1400708	2214	0.6981313	2254	0.9470667	2294	0.5791601	2334	0.0807566
2175	0.1491696	2215	0.7118706	2255	0.9454098	2295	0.5634745	2335	0.0749337
2176	0.1586548	2216	0.7252968	2256	0.9433803	2296	0.5476861	2336	0.0694366
2177	0.1685271	2217	0.7383947	2257	0.9409759	2297	0.5318199	2337	0.0642555
2178	0.1787857	2218	0.7511501	2258	0.9381936	2298	0.5159012	2338	0.0593799
2179	0.1894289	2219	0.7635504	2259	0.9350303	2299	0.4999557	2339	0.0547993
2180	0.2004532	2220	0.7755839	2260	0.9314826	2300	0.4840091	2340	0.0505026
2181	0.2118542	2221	0.7872403	2261	0.927547	2301	0.468087	2341	0.0464786
2182	0.2236259	2222	0.7985108	2262	0.9232197	2302	0.4522148	2342	0.0427162
2183	0.2357608	2223	0.8093875	2263	0.9184972	2303	0.436418	2343	0.0392039
2184	0.2482501	2224	0.8198637	2264	0.9133756	2304	0.4207213	2344	0.0359303
2185	0.2610837	2225	0.8299342	2265	0.9078514	2305	0.4051492	2345	0.0328841
2186	0.27425	2226	0.8395946	2266	0.9019213	2306	0.3897254	2346	0.030054
2187	0.287736	2227	0.8488417	2267	0.8955821	2307	0.374473	2347	0.0274289
2188	0.3015273	2228	0.8576732	2268	0.8888312	2308	0.3594142	2348	0.0249979
2189	0.3156083	2229	0.866088	2269	0.8816662	2309	0.3445705	2349	0.0227501

b. Trace la gráfica de la curva formada por la media real como abscisa contra la probabilidad de aceptar la hipótesis nula como ordenada, a la cual se le conoce como Curva Característica de Operación (CCO).

Figura 5.4



c. Estime la probabilidad del error tipo I, es decir, la probabilidad de rechazar que la media es de 2250 psi, cuando realmente es de 2250 psi.

$$\alpha = p\left(rechazar\_H_0 \mid H_0\_es\_cierta\right) = 1 - \left(aceptar\_H_0 \mid H_0\_es\_cierta\right)$$

$$\alpha = 1 - p\left(2200 \leq \mu_x \leq 2300 \mid \mu_x = 2250\right) = 0.05$$

d. Determine la probabilidad de aceptar que la media es de 2250 psi, cuando realmente es de 2200 psi.

$$\beta = \mathrm{p}\left(aceptar\_H_0 \mid H_0\_es\_falsa\right) = p\left(2200 \leq \mu_x \leq 2300 \mid \mu_x = 2200\right) = 0.484$$

e. ¿Para qué media real la probabilidad de aceptación es de 20%?

De acuerdo con la gráfica, existen dos valores de la media para las cuales la probabilidad de aceptación es del 20.04%: 2180 y 2320 psi.

5.2.2. Pruebas de hipótesis sobre la media de una población normal o tamaño de muestra grande con varianza desconocida

Suponga que x es normal $x \sim N(\mu_x, \sigma_x)$ o n es suficientemente grande para suponer que se cumple el teorema del límite central, es decir, $\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la media poblacional, con el estadístico t de Student; si dicho intervalo no contiene al valor μ_0 entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor μ_0 entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$t_0 = \frac{\bar{x} - \mu_0}{S_{n-1}/\sqrt{n}} \tag{5.13}$$

Para una prueba bilateral o de dos colas, donde H_1: $\mu_x \neq \mu_0$

$$Si \qquad |t_0| > t_{\frac{\alpha}{2}, n-1} \tag{5.14}$$

Es decir:

$$t_0 > t_{\frac{\alpha}{2}, n-1} \quad \text{o} \quad t_0 < -t_{\frac{\alpha}{2}, n-1}$$

Se rechaza la hipótesis nula

De lo contrario, si

$$Si \; -t_{\frac{\alpha}{2}, n-1} < t_0 < t_{\frac{\alpha}{2}, n-1} \tag{5.15}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x < \mu_0$

Si $t_0 > t_{\alpha, n-1}$ se rechaza la hipótesis es nula. (5.16)

De lo contrario, si $t_0 > t_{\alpha, n-1}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x > \mu_0$

Si $t_0 < t_{\alpha, n-1}$ se rechaza la hipótesis es nula. (5.17)

De lo contrario, si $t_0 < t_{\alpha, n-1}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(t > t_0)$ o $p = p(t < -t_0)$, utilizando la distribución t de Student

Para una prueba bilateral o de dos colas, donde H_1: $\mu_x \neq \mu_0$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x < \mu_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_x > \mu_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x \neq \mu_0$

Unilateral_Inferior
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x < \mu_0$

Unilateral_Superior (5.18)
$H_0 : \mu_x = \mu_0$
$H_1 : \mu_x > \mu_0$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico t de Student; la prueba puede ser de dos colas $t_{\alpha/2,\, n-1}$, unilateral inferior $t_{\alpha,\, n-1}$ o unilateral superior $-t_{\alpha,\, n-1}$

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.8

Del ejercicio 5.6. Un proveedor afirma que la resistencia del concreto hidráulico que él fabrica es mayor a 2475 psi. En el área de recibo de una empresa constructora un ingeniero civil realiza la prueba a 50 probetas, obteniendo los siguientes datos en unidades de psi:

2243	2310	2281	2277	2272
2246	2271	2235	2261	2251
2320	2208	2268	2263	2295
2215	2270	2264	2241	2205
2234	2285	2256	2305	2223
2271	2244	2306	2281	2242
2287	2254	2212	2252	2258
2267	2268	2279	2304	2240
2257	2230	2263	2297	2270
2263	2290	2219	2224	2262

Suponga que la resistencia a la compresión de dichas probetas es normal.

Realice una prueba de hipótesis estadística al 95 y 99% de nivel de confianza de la resistencia promedio del concreto a la compresión.

H_0: $\mu_x = 2475$
H_1: $\mu_x < 2475$

Nótese que la prueba es de una cola. Los valores críticos son:

$t_{0.05,49} = 1.67655$
$t_{0.01,49} = 2.40489$

Con el primer criterio, se calcula un intervalo inferior de confianza con el estadístico t de Student:

$\mu_x < 2267.76$ al 95% de nivel de confianza
$\mu_x < 2270.65$ al 99% de nivel de confianza

En ambos casos no contiene al valor que establece el proveedor, por lo que se rechaza contundentemente su afirmación.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$t_0 = (2260.78\text{-}2475)/(27.84/\text{raíz}(50)) = -54.41$

Nótese que, en ambos casos, $t_0 < -t_{\alpha,\, n-1}$, por lo que se rechaza la hipótesis nula del proveedor.

Con el tercer criterio, se calcula la probabilidad de que $t < -t_0$

$p = p(t < -t_0) = 0 < 0.01 < 0.05$, por lo tanto, se rechaza la hipótesis nula del proveedor.

5.2.3. Pruebas de hipótesis sobre la varianza de una población normal o para una muestra de tamaño grande

Suponga que x es normal $x \sim N\left(\mu_x, \sigma_x\right)$ o n es suficientemente grande.

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la varianza poblacional, con el estadístico ji cuadrada; si dicho intervalo no contiene al valor σ_0^2 entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor σ_0^2 entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$\chi_0^2 = \frac{(n-1)}{\sigma_0^2} S_{n-1}^2 \tag{5.19}$$

Para una prueba bilateral o de dos colas, donde H_1: $\sigma_x^2 \neq \sigma_0^2$

$$Si \quad \chi_0^2 > \chi_{\frac{\alpha}{2}, n-1}^2 \quad \text{o} \quad \chi_0{}^2 < \chi_{1-\frac{\alpha}{2}, n-1}^2 \tag{5.20}$$

Se rechaza la hipótesis nula

De lo contrario, si

$$Si \quad \chi_{1-\frac{\alpha}{2}, n-1}^2 < \chi_0^2 < \chi_{\frac{\alpha}{2}, n-1}^2 \tag{5.21}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\sigma_x^2 > \sigma_0^2$

$$Si \quad \chi_0^2 > \chi_{\alpha, n-1}^2 \text{ se rechaza la hipótesis nula} \tag{5.22}$$

De lo contrario, si $\chi_0^2 < \chi_{\alpha,\, n-1}^2$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\sigma_x^2 < \sigma_0^2$

$$Si \quad \chi_0^2 < \chi_{1-\alpha,\, n-1}^2 \text{ se rechaza la hipótesis nula} \tag{5.23}$$

De lo contrario, si $\chi_0^2 > \chi_{1-\alpha,\, n-1}^2$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(\chi^2 > \chi_0^2\right)$, utilizando la distribución ji cuadrada.

Para una prueba bilateral o de dos colas, donde H_1: $\sigma_x^2 \neq \sigma_0^2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\sigma_x^2 > \sigma_0^2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\sigma_x^2 < \sigma_0^2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.
Bilateral
$H_0 : \sigma_x^2 = \sigma_0^2$
$H_1 : \sigma_x^2 \neq \sigma_0^2$

Unilateral_Inferior
$H_0: \sigma_x^2 = \sigma_0^2$
$H_1: \sigma_x^2 < \sigma_0^2$

Unilateral_Superior (5.24)
$H_0: \sigma_x^2 = \sigma_0^2$
$H_1: \sigma_x^2 > \sigma_0^2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico ji cuadrada; la prueba puede ser de dos colas $\chi^2_{\frac{\alpha}{2},\, n-1}$ o $\chi^2_{1-\frac{\alpha}{2},\, n-1}$, unilateral inferior $\chi^2_{1-\alpha,\, n-1}$ o unilateral superior $\chi^2_{\alpha,\, n-1}$

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.9

Suponga que el mismo proveedor del ejercicio 5.6 afirma que la resistencia del concreto hidráulico que él fabrica presenta una desviación estándar menor de 20 psi. Con la misma muestra obtenida en dicho ejercicio realice una prueba de hipótesis estadística al 95 y 99% de nivel de confianza, de la varianza de la resistencia del concreto a la compresión, para probar la afirmación del proveedor:

2243	2310	2281	2277	2272
2246	2271	2235	2261	2251
2320	2208	2268	2263	2295
2215	2270	2264	2241	2205
2234	2285	2256	2305	2223
2271	2244	2306	2281	2242

2287	2254	2212	2252	2258
2267	2268	2279	2304	2240
2257	2230	2263	2297	2270
2263	2290	2219	2224	2262

Nótese que la prueba es de una cola. Los valores críticos son:

$\chi^2_{0.05,49} = 66.3386$

$\chi^2_{0.01,49} = 74.9195$

Con el primer criterio, se calcula un intervalo inferior de confianza con el estadístico t de Student:

$$\frac{(n-1)S^2_{n-1}}{\chi^2_{\alpha,\, n-1}} \leq \sigma^2_x \Rightarrow \sqrt{\frac{(n-1)S^2_{n-1}}{\chi^2_{\alpha,\, n-1}}} \leq \sigma_x$$

$23.9269 \leq \sigma_x \quad al\_95\%$

$22.515 \leq \sigma_x \quad al\_99\%$

Como se puede apreciar, en ambos casos el intervalo no contiene al valor que establece el proveedor, por lo que se rechaza contundentemente su afirmación.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$$\chi^2_0 = \frac{(n-1)}{\sigma^2_0} S^2_{n-1} = \frac{49^*775.0731}{20^2} = 94.9465$$

Nótese que, en ambos casos, $\chi^2_0 > \chi^2_{\alpha,\, n-1}$, por lo que se rechaza la hipótesis nula del proveedor.

Con el tercer criterio, se calcula la probabilidad $p = p\left(\chi^2 > \chi^2_0\right)$, utilizando la distribución ji cuadrada

p = DISTR.CHICUAD.CD(94.94645,49) = 9.12735E-5 = 0.000091

Como se puede apreciar $p = 0.000091 < 0.01 < 0.05$, por lo tanto, se rechaza la hipótesis nula del proveedor.

5.2.4. Pruebas de hipótesis para el número de elementos exitosos en una muestra de tamaño n o para la fracción o proporción de elementos exitosos *p* en una población

Suponga que x es binomial con $np > 5$ para $p < 0.5$, o sea,

$x \sim N\left(\mu_x = np,\ \sigma_x = \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}\right)$ o n es suficientemente grande para suponer el cumplimiento del teorema del límite central.

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al $100(1-\alpha)\%$ de nivel de confianza, para la fracción p, con el estadístico z; si dicho intervalo no contiene al valor p_0 entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor p_0 entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{p_0 - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}} \qquad N\_finita$$

$$z_0 = \frac{p_0 - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \qquad N\_infinita \tag{5.25}$$

Para una prueba bilateral o de dos colas, donde

$$Si \qquad z_0 > z_{\frac{\alpha}{2}} \ \text{o} \quad z_0 < -z_{\frac{\alpha}{2}} \tag{5.26}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -z_{\frac{\alpha}{2}} < z_0 < z_{\frac{\alpha}{2}} \tag{5.27}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p > p_0$

Si $z_0 > z_\alpha$ se rechaza la hipótesis nula

De lo contrario, si $z_0 > z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p > p_0$

Si $z_0 < -z_\alpha$ se rechaza la hipótesis nula

De lo contrario, si $z_0 < -z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(z > z_0)$ o $p = p(z < -z_0)$, utilizando la distribución normal estándar.

Para una prueba bilateral o de dos colas, donde $H_1 : p \neq p_0$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p > p_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p < p_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : p = p_0$
$H_1 : p \neq p_0$

Unilateral_Inferior (5.28)
$H_0 : p = p_0$
$H_1 : p < p_0$

Unilateral_Superior
$H_0 : p = p_0$
$H_1 : p > p_0$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico z; la prueba puede ser de dos colas $z_{\alpha/2}$, unilateral inferior o unilateral superior z_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.10 Un proveedor de remaches entrega un lote de $N = 5000$ piezas. El proveedor afirma que dicho lote tiene menos de 2.5% de defectuosos. Para demostrarlo, el área de recibo de la empresa cliente toma una muestra de $n = 300$ artículos, y los prueba funcionalmente (la prueba es destructiva). Suponga que después de hacer la inspección se obtuvieron 15 remaches defectuosos. Realice una prueba de hipótesis estadística al 95 y 99% de nivel de confianza de la fracción de artículos defectuosos, para probar la afirmación del proveedor.

Nótese que la prueba es de una cola. Los valores críticos son:

$$z_{0.05} = 1.64485$$

$$z_{0.01} = 2.32635$$

Con el primer criterio, se calcula un intervalo inferior de confianza con el estadístico z:

$$\hat{p} = \frac{15}{300} = 0.05$$

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}\left(\frac{N-n}{N-1}\right)} \leq p$$

$$0.0299 \leq p \qquad al\_95\%$$
$$0.0216 \leq p \qquad al\_99\%$$

Como se puede apreciar, al 95% del nivel de confianza el intervalo no contiene al valor que establece el proveedor, pero al 99% del nivel de confianza si lo contiene, por lo que cae en zona de duda, lo recomendable sería tomar una muestra más grande para tomar la decisión.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$$z_0 = \frac{p_0 - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}} = -2.049$$

Nótese que $-z_{0.01} < z_0 < -z_{0.05}$, por lo que se rechazaría la hipótesis nula al 95% pero se aceptaría al 99% de nivel de confianza, por lo que cae en zona de duda, lo recomendable sería tomar una muestra más grande para tomar la decisión.

Con el tercer criterio, se calcula la probabilidad $p = p(z > z_0)$, utilizando la distribución normal estándar.

p = DISTR.NORMAL.N(-2.049,0,1) = 0.02023

Como se puede apreciar $0.01 < 0.02023 < 0.05$, por lo que cae en zona de duda, lo recomendable sería tomar una muestra más grande para tomar la decisión.

Ejercicio 5.11

Una empresa cliente recibe lotes de $N = 2000$ balatas y para aceptar cada uno de ellos toma una muestra de $n = 200$; el criterio de decisión para aceptar cada lote es si se obtienen cinco o menos balatas defectuosas se acepta el lote, de lo contrario, se rechaza.

La prueba de hipótesis que se aplica es la siguiente:

$$H_0 : p = \frac{5}{200} = 0.025$$

$$H_1 : p > 0.025$$

O también

$$H_0 : x = np = 5$$

$$H_1 : x > 5$$

a. Formule la expresión para calcular la probabilidad de aceptar un lote

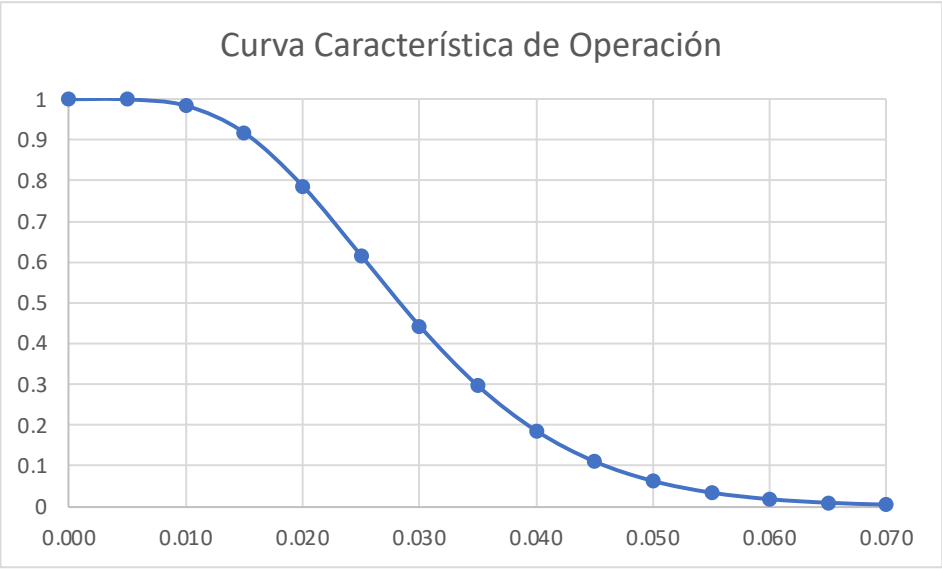
La distribución de probabilidad del número de defectuosos x es hipergeométrica, por lo que la probabilidad de aceptar el lote está dada por la expresión:

$$\beta = p(x \le 5 \mid p) = \sum_{x=0}^{x=5} \frac{\binom{D}{x}\binom{2000-D}{200-x}}{\binom{2000}{200}}$$

$$\beta = \sum_{x=0}^{x=5} \frac{\binom{2000\,p}{x}\binom{2000-2000\,p}{200-x}}{\binom{2000}{200}}$$

b. Trace la Curva Característica de Operación

Figura 5.4



5.2.5. Pruebas de hipótesis para el número de defectos, ocurrencias, éxitos o llegadas en n unidades, así como la fracción de defectos, ocurrencias, éxitos o llegadas por unidad

Suponga que x tiene distribución tipo Poisson con $c > 5$ para $p < 0.5$, o sea,

$x \sim N\left(\mu_x = c, \sigma_x = \sqrt{c\left(\frac{N-n}{N-1}\right)}\right)$ o n es suficientemente grande para suponer el cumplimiento del teorema del límite central.

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al $100(1-\alpha)\%$ de nivel de confianza, para la fracción u, con el estadístico z, si dicho intervalo no contiene al valor u_0 entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor u_0 entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{u_0 - \hat{u}}{\sqrt{\frac{\hat{u}}{n}\left(\frac{N-n}{N-1}\right)}} \qquad N\_finita$$

$$z_0 = \frac{u_0 - \hat{u}}{\sqrt{\frac{\hat{u}}{n}}} \qquad N\_infinita \tag{5.29}$$

Para una prueba bilateral o de dos colas, donde $H_1 : u \neq u_0$

$$Si \qquad z_0 > z_{\frac{\alpha}{2}} \quad \text{o} \quad z_0 < -z_{\frac{\alpha}{2}} \tag{5.30}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -z_{\frac{\alpha}{2}} < z_0 < -z_{\frac{\alpha}{2}} \tag{5.31}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : u > u_0$

$$Si \qquad z_0 > z_\alpha \text{ se rechaza la hipótesis nula} \tag{5.32}$$

De lo contrario, si $z_0 > z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : u < u_0$

$$Si \qquad z_0 < -z_\alpha \text{ se rechaza la hipótesis nula} \tag{5.33}$$

De lo contrario, si $z_0 < -z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(z > z_0)$ o $p = p(z < -z_0)$, utilizando la distribución normal estándar.

Para una prueba bilateral o de dos colas, donde $H_1 : u \neq u_0$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : u > u_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : u < u_0$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : u = u_0$
$H_1 : u \neq u_0$

Unilateral_Inferior (5.34)
$H_0 : u = u_0$
$H_1 : u < u_0$

Unilateral_Superior
$H_0 : u = u_0$
$H_1 : u > u_0$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico z; la prueba puede ser de dos colas $z_{\alpha/2}$, unilateral inferior o unilateral superior z_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.12

Un proveedor de remaches entrega un lote de N = 5000 piezas. El proveedor afirma que dicho lote tiene menos de 2.5% de defectuosos. Para demostrarlo, el área de recibo de la empresa cliente toma una muestra de n = 300 artículos, y los prueba funcionalmente (la prueba es destructiva). Suponga que después de hacer la inspección se obtuvieron 15 remaches defectuosos. Realice una prueba de hipótesis estadística al 95 y 99% de nivel de confianza de la fracción de artículos defectuosos para probar la afirmación del proveedor.

Nótese que la prueba es de una cola. Los valores críticos son:

$$z_{0.05} = 1.64485$$
$$z_{0.01} = 2.32635$$

Con el primer criterio, se calcula un intervalo inferior de confianza con el estadístico z:

$$\hat{p} = \frac{15}{300} = 0.05$$

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}\left(\frac{N-n}{N-1}\right)} \leq p$$

$$0.0299 \leq p\ al\_95\%$$
$$0.0216 \leq p\ al\_99\%$$

Como se puede apreciar, al 95% del nivel de confianza el intervalo no contiene al valor que establece el proveedor, pero al 99% del nivel de confianza sí lo contiene, por lo que cae en zona de duda, lo recomendable sería tomar una muestra más grande para tomar la decisión.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$$z_0 = \frac{p_0 - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}} = -2.049$$

Nótese que $-z_{0.01} < z_0 < -z_{0.05}$, por lo que se rechazaría la hipótesis nula al 95% pero se aceptaría al 99% de nivel de confianza, por lo que cae en zona de duda, lo recomendable sería tomar una muestra más grande para tomar la decisión.

Con el tercer criterio, se calcula la probabilidad , utilizando la distribución normal estándar.

p = DISTR.NORMAL.N(-2.049,0,1) = 0.02023

Como se puede apreciar $0.01 < 0.02023 < 0.05$, por lo que cae en zona de duda, lo recomendable sería una muestra más grande para tomar la decisión.

Ejercicio 5.13

Del ejercicio 4.20. Una empresa arrendadora de computadoras le renta a un laboratorio de pruebas $n = 200$ computadoras. Suponga que el número de computadoras que posee el corporativo es infinito. Por datos históricos el laboratorio de pruebas ha logrado contabilizar el número de defectos de las computadoras que se envían a mantenimiento correctivo en los últimos diez meses, como se muestra a continuación:

mes	No. Computadoras	No. Defectos
1	15	87
2	12	93
3	10	112
4	13	115
5	12	120
6	11	93
7	20	130
8	12	105
9	13	100
10	13	99
Suma =	131	1054

La empresa arrendadora sostiene en su contrato que el número de defectos por unidad de sus computadoras no llega a cinco, por lo que el laboratorio de pruebas realiza una prueba de hipótesis al 95 y al 99% de nivel de confianza, para probar la afirmación del proveedor.

Nótese que la prueba es de una cola. Los valores críticos son:

$$z_{0.05} = 1.64485$$
$$z_{0.01} = 2.32635$$

Con el primer criterio, se calcula un intervalo inferior de confianza con el estadístico z:

$$\hat{u} = \frac{1054}{131} = 8.0458$$

$$\hat{u} - z_{\alpha}\sqrt{\frac{\hat{u}}{n}} \leq u$$

$$7.7159 \leq u \qquad al\_95\%$$
$$7.5792 \leq u \qquad al\_99\%$$

Como se puede apreciar, en ambos niveles de confianza el intervalo no contiene al valor que establece el proveedor, por lo que se rechaza contundentemente la afirmación del proveedor.

Con el segundo criterio, se calcula el valor del estadístico de prueba

$$z_0 = \frac{u_0 - \hat{u}}{\sqrt{\frac{\hat{u}}{n}}} = -15.186$$

Nótese que $z_0 < -z_{0.01} < -z_{0.05}$, por lo que se rechaza contundentemente la afirmación del proveedor.

Con el tercer criterio, se calcula la probabilidad $p = p(z > z_0)$, utilizando la distribución normal estándar.

p = DISTR.NORMAL.N(-15.186,0,1) = 0

Como se puede apreciar $0 < 0.01 < 0.05$, por lo que se rechaza contundentemente la afirmación del proveedor.

Figura 5.5. Pruebas de Hipótesis para un parámetro de una población normal

Hipótesis nula	Condiciones iniciales	Valor del estadístico de prueba	Hipótesis alternativa	Criterio rechazo H_o	Criterio rechazo H_o	Parámetro k de la CCO
$H_0: \mu_x - \mu_0$	$x \sim N(\mu_x, \sigma_x)$ o n_grande N Finita σ_x^2 Conocida	$z_0 = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma_x}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}}$	$H_1: \mu_x \neq \mu_0$	$z_0 > z_{\alpha/2}$ $z_0 < -z_{\alpha/2}$	$p = p(z > z_0) < \alpha/2$ $p = p(z < -z_0) < \alpha/2$	$k = \lvert\mu - \mu_0\rvert / \sigma$
			$H_1: \mu_x > \mu_0$	$z_0 > z_\alpha$	$p < \alpha$	$k = (\mu - \mu_0)/\sigma$
			$H_1: \mu_x < \mu_0$	$z_0 < -z_\alpha$	$p < \alpha$	$k = (\mu_0 - \mu)/\sigma$
$H_0: \mu_x = \mu_0$	$x \sim N(\mu_x, \sigma_x)$ N Finita σ_x^2 Desconocida	$t_0 = \dfrac{\bar{x} - \mu_0}{\dfrac{S_{n-1}}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}}$	$H_1: \mu_x \neq \mu_0$	$t_0 > t_{\alpha/2,n-1}$ $t_0 < -t_{\alpha/2,n-1}$	$p = p(t > t_0) < \alpha/2$ $p = p(t < -t_0) < \alpha/2$	$k = \lvert\mu - \mu_0\rvert / \sigma$
			$H_1: \mu_x > \mu_0$	$t_0 > t_{\alpha,n-1}$	$p < \alpha$	$k = (\mu - \mu_0)/\sigma$
			$H_1: \mu_x < \mu_0$	$t_0 > t_{\alpha,n-1}$	$p < \alpha$	$k = (\mu_0 - \mu)/\sigma$
$H_0: \sigma_x^2 = \sigma_0^2$	$x \sim N(\mu_x, \sigma_x)$ N Finita	$\chi_0^2 = \dfrac{(n-1)S_{n-1}^2}{\sigma_0^2}\left(\dfrac{N-n}{N-1}\right)$	$H_1: \sigma_x^2 \neq \sigma_0^2$	$\chi_0^2 > \chi^2_{\frac{\alpha}{2},n-1}$ $\chi_0^2 < \chi^2_{1-\frac{\alpha}{2},n-1}$	$p = p(\chi^2_{\frac{\alpha}{2},n-1} > \chi_0^2) < \alpha/2$ $p = p(\chi^2_{1-\frac{\alpha}{2},n-1} < \chi_0^2) < \alpha/2$	$\lambda = \sigma / \sigma_0$
			$H_1: \sigma_x^2 > \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha,n-1}$	$p < \alpha$	$\lambda = \sigma / \sigma_0$
			$H_1: \sigma_x^2 < \sigma_0^2$	$\chi_0^2 < \chi^2_{1-\alpha,n-1}$	$p < \alpha$	$\lambda = \sigma / \sigma_0$
$H_0: \mathrm{p} = \mathrm{p}_0$	$x \sim N(\mu_x, \sigma_x)$ N Finita n Grande	$z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}\left(\dfrac{N-n}{N-1}\right)}}$	$H_1: p \neq p_0$	$z_0 > z_{\alpha/2}$ $z_0 < -z_{\alpha/2}$	$p = p(z > z_0) < \alpha/2$ $p = p(z < -z_0) < \alpha/2$	$k = \lvert p - p_0\rvert / \sigma$
			$H_1: p > p_0$	$z_0 > z_\alpha$	$p < \alpha$	$k = (p - p_0)/\sigma$
			$H_1: p < p_0$	$z_0 < -z_\alpha$	$p < \alpha$	$k = (p_0 - p)/\sigma$
$H_0: \mathrm{u} = u_0$	$x \sim N(\mu_x, \sigma_x)$ N Finita n Grande	$z_0 = \dfrac{\hat{u} - u_0}{\sqrt{\dfrac{u_0}{n}\left(\dfrac{N-n}{N-1}\right)}}$	$H_1: \mathrm{u} \neq u_0$	$z_0 > z_{\alpha/2}$ $z_0 < -z_{\alpha/2}$	$p = p(z > z_0) < \alpha/2$ $p = p(z < -z_0) < \alpha/2$	$k = \lvert u - u_0\rvert / \sigma$
			$H_1: \mathrm{u} > u_0$	$z_0 > z_\alpha$	$p < \alpha$	$k = (u - u_0)/\sigma$
			$H_1: \mathrm{u} < u_0$	$z_0 < -z_\alpha$	$p < \alpha$	$k = (u_0 - u)/\sigma$

5.3. Pruebas de hipótesis de un mismo parámetro para dos poblaciones

5.3.1. Prueba de hipótesis para demostrar la igualdad de medias de dos poblaciones normales o tamaños de muestras grandes con desviaciones estándar conocidas

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia de medias, con el estadístico z; si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}} \qquad N\_finita$$

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad N\_infinita \tag{5.35}$$

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

$$Si \qquad z_0 > z_{\frac{\alpha}{2}} \quad \text{o} \quad z_0 < -z_{\frac{\alpha}{2}} \tag{5.36}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -z_{\frac{\alpha}{2}} < z_0 - z_{\frac{\alpha}{2}} \tag{5.37}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 > \mu_2$

Si $z_0 > z_\alpha$ se rechaza la hipótesis nula (5.38)

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 < \mu_2$

Si $z_0 < -z_\alpha$ se rechaza la hipótesis nula (5.39)

De lo contrario, si $z_0 < -z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(z > z_0)$ o $p = p(z < -z_0)$, utilizando la distribución normal estándar.

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 > \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 < \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

Unilateral_Superior (5.40)
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$

Unilateral_Inferior
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico z; la prueba puede ser de dos colas $z_{\alpha/2}$, unilateral superior o unilateral inferior z_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.14 Del ejercicio 4.24. Una empresa del sector eléctrico usa dos tipos de material aislante (A y B) en las piezas que fabrica. Suponga que se fabrican 3000 piezas con el material A y 2000 con el material B diariamente. El nuevo gerente de planta desea

abatir costos de producción y descubre que el costo del material B es mayor que el costo del material A, por lo cual solicita realizar una prueba de hipótesis para determinar si ambos materiales tienen la misma resistencia. Para determinar la resistencia media del material A se tomaron n = 8 lecturas, y para el material B, nueve lecturas. Los datos obtenidos se muestran a continuación:

Material	1	2	3	4	5	6	7	8	9
A	1.25	1.16	1.33	1.15	1.23	1.2	1.32	1.28	
B	1.01	0.89	0.97	0.95	0.94	1.02	0.99	1.06	0.98

Suponga que la varianza en la resistencia del material A es 0.05 y la varianza en la resistencia del material B es 0.0025.

$\bar{x}_1 = 1.24$
$\bar{x}_2 = 0.97889$
$z_{0.05} = 1.95996$
$z_{0.01} = 2.57583$

i. Se obtiene un intervalo bilateral de confianza para la diferencia entre medias al 95 y al 99% de nivel de confianza

$0.10276 \leq \mu_1 - \mu_2 \leq 0.419466$ *al_95%*
$0.053 \leq \mu_1 - \mu_2 \leq 0.469224$ *al_99%*

Como se puede apreciar en los intervalos de confianza anteriores, ambos no contienen al cero, por lo que se concluye que las medias de resistencia dieléctrica de ambos materiales A y B no son iguales.

ii. Se calcula el estadístico z_0

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}} = \frac{\left(1.24 - 0.97889\right)}{\sqrt{\frac{0.05}{8}\left(\frac{3000 - 8}{3000 - 1}\right) + \frac{0.0025}{9}\left(\frac{2000 - 9}{2000 - 1}\right)}} = 3.23568$$

Como se puede apreciar $z_0 > 2.57583 > 1.95996$, por lo cual se rechaza la hipótesis de igualdad entre medias.

iii. $p = p(z > z0) = p(z > 3.23568) = 1\text{-}p(z < 3.23568) = 0.000607 < 0.01 < 0.05$, por lo cual se rechaza la hipótesis de igualdad entre medias.

5.3.2. Prueba de hipótesis para demostrar la igualdad de varianzas de dos poblaciones normales o con tamaños de muestras grandes

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para el cociente entre varianzas, con el estadístico F de Fisher-Snedecor, si dicho intervalo no contiene al valor uno entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor uno entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$F_0 = \frac{S^2_{n_1-1}}{S^2_{n_2-1}} \frac{\left(\frac{N_1 - n_1}{N_1 - 1}\right)}{\left(\frac{N_2 - n_2}{N_2 - 1}\right)} \tag{5.41}$$

Para una prueba bilateral o de dos colas, donde $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$Si \qquad F_0 > F_{\frac{\alpha}{2}, n_1-1, n_2-1} \quad \text{o} \quad F_0 < F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} \tag{5.42}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad F_{\frac{\alpha}{2}, n_1-1, n_2-1} \leq F_0 \leq F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} \tag{5.43}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \sigma_1^2 > \sigma_2^2$

$$Si \qquad F_0 > F_{\alpha, n_1-1, n_2-1} \qquad \text{se rechaza la hipótesis nula} \tag{5.44}$$

De lo contrario, si $F_0 < F_{\alpha,\, n_1-1,\, n_2-1}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(F > F_0\right)$, utilizando la distribución F de Fisher-Snedecor.

Para una prueba bilateral o de dos colas, donde $H_1 : \sigma_1^2 \neq \sigma_2^2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \sigma_1^2 > \sigma_2^2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad (5.45)$$

Unilateral_Superior

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico F; la prueba puede ser de dos colas $F_{\alpha/2}$, o unilateral F_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.15 Va a efectuarse una comparación de las precisiones de dos máquinas para extraer jugo de naranja usando los siguientes datos:

Máquina	n	N	S^2_{n-1}
A	25	1500	3.1
B	24	1200	1.4

i. Se obtiene un intervalo bilateral de confianza para el cociente entre varianzas al 95 y al 99% de nivel de confianza

$$F_{0.975,23,24} = 0.43499 \qquad F_{0.995,23,24} = 0.33103$$
$$F_{0.025,23,24} = 2.28207 \qquad F_{0.005,23,24} = 2.98779$$

$$\frac{\left(\frac{N_A - n_A}{N_A - 1}\right)}{\left(\frac{N_B - n_B}{N_B - 1}\right)} = \frac{\left(\frac{1500 - 25}{1500 - 1}\right)}{\left(\frac{1200 - 24}{1200 - 1}\right)} = \frac{0.98399}{0.98082} = 1.00323$$

$$\frac{S^2_{n_A-1}}{S^2_{n_B-1}} \frac{\left(\frac{N_A - n_A}{N_A - 1}\right)}{\left(\frac{N_B - n_B}{N_B - 1}\right)} F_{1-\frac{\alpha}{2}, n_B-1, n_A-1} \leq \frac{\sigma^2_A}{\sigma^2_B} \leq \frac{S^2_{n_A-1}}{S^2_{n_B-1}} \frac{\left(\frac{N_A - n_A}{N_A - 1}\right)}{\left(\frac{N_B - n_B}{N_B - 1}\right)} F_{\frac{\alpha}{2}, n_B-1, n_A-1}$$

$$0.96631 \leq \frac{\sigma^2_A}{\sigma^2_B} \leq 5.0695 \qquad \text{al\_95\%}$$
$$0.73537 \leq \frac{\sigma^2_A}{\sigma^2_B} \leq 6.63721 \qquad \text{al\_99\%}$$

Como se puede apreciar en los intervalos de confianza anteriores ambos contienen al valor uno, por lo que se concluye que las varianzas de las precisiones de ambas máquinas A y B son iguales.

ii. Se calcula el estadístico F_0

$$F_0 = \frac{S^2_{n_1-1}}{S^2_{n_2-1}} \frac{\left(\frac{N_1 - n_1}{N_1 - 1}\right)}{\left(\frac{N_2 - n_2}{N_2 - 1}\right)} = 2.22145$$

Como se puede apreciar F_0 cae en ambos intervalos de F al 95 y al 99% de nivel de confianza, por lo cual se acepta que ambas varianzas son iguales.

$F_{0.975,23,24} = 0.4382 \quad F_{0.975,23,24} = 0.3347$
$F_{0.025,23,24} = 2.29891 F_{0.005,23,24} = 3.02085$

iii. $p = p(F > F_0) = p(F > 2.22145) = 0.02996 > 0.025 > 0.005$, por lo cual se acepta que ambas varianzas son iguales.

5.3.3. Prueba de hipótesis para demostrar la igualdad de medias de dos poblaciones normales o con tamaños de muestras grandes, con varianzas poblacionales desconocidas pero iguales

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas pero iguales; si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5.46}$$

Donde

$$S_p = \sqrt{\frac{(n_1 - 1)\, S_1^2\left(\frac{N_1 - n_1}{N_1 - 1}\right) + (n_2 - 1)\, S_2^2\left(\frac{N_2 - n_2}{N_2 - 1}\right)}{n_1 + n_2 - 2}} \tag{5.47}$$

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

$$Si \qquad t_0 > t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \qquad \text{o} \qquad t_0 < -t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \tag{5.48}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \leq t_0 \leq t_{\frac{\alpha}{2}, n_1 + n_2 - 2} \tag{5.49}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

$$Si \qquad t_0 > t_{\alpha,\, n_1+n_2-2} \qquad \text{se rechaza la hipótesis nula} \tag{5.50}$$

De lo contrario, si $t_0 < t_{\alpha,\, n_1+n_2-2}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

$$Si \qquad t_0 < t_{\alpha,\, n_1+n_2-2} \qquad \text{se rechaza la hipótesis nula} \tag{5.51}$$

De lo contrario, si $t_0 > t_{\alpha,\, n_1+n_2-2}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(t > t_0\right)$ o $p = p\left(t < -t_0\right)$, utilizando la distribución t de Student.

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Unilateral_Superior (5.52)
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

Unilateral_Inferior
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico *t*; la prueba puede ser de dos colas $t_{\alpha/2}$, o unilateral t_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.16

Se están investigando los diámetros de las barras de acero fabricadas por dos diferentes máquinas de extruido. Se seleccionan dos muestras aleatorias de tamaño $n_1 = 12$ y $n_2 = 18$, y las medias y varianzas muestrales son $\bar{x}_1 = 8.75$, $S_1^2 = 0.29$, $\bar{x}_2 = 8.63$ y $S_2^2 = 0.34$, respectivamente.

a. Realice una prueba de hipótesis para demostrar la igualdad entre varianzas.

$$F_{0.025,11,17} = 0.30473 \qquad F_{0.975,11,17} = 2.86964$$
$$F_{0.005,11,17} = 0.20167 \qquad F_{0.975,11,17} = 4.04956$$

$$F_0 = \frac{S^2_{n_1-1}}{S^2_{n_2-1}} = \frac{0.29}{0.34} = 0.8529$$

Como se puede apreciar, F_0, cae dentro del intervalo de aceptación de la igualdad entre varianzas para ambos niveles de confianza al 95% y al 99%.

b. Aplique una prueba de hipótesis para determinar si las medias en los diámetros de las barras de ambas máquinas de extruido son iguales.

i. Obteniendo un intervalo de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas pero iguales, de acuerdo con lo demostrado en el inciso a.

$$t_{0.025,28} = 2.04841$$
$$t_{0.005,28} = 2.76326$$

$$S_p = \sqrt{\frac{(n_1-1)\,S_1^2\left(\frac{N_1-n}{N_1-1}\right) + (n_2-1)\,S_2^2\left(\frac{N_2-n_2}{N_2-1}\right)}{n_1+n_2-2}} = \sqrt{\frac{11(0.29)+17(0.34)}{28}} = 0.566$$

$$(\bar{x}_1-\bar{x}_2) - t_{\frac{\alpha}{2},\,n_1+n_2-2}\;S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le (\bar{x}_1-\bar{x}_2) + t_{\frac{\alpha}{2},\,n_1+n_2-2}\;S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$-0.3121 \le \mu_1-\mu_2 \le 0.55208 \quad al\_95\%$$
$$-0.4629 \le \mu_1-\mu_2 \le 0.70287 \quad al\_99\%$$

Como se puede apreciar, ambos intervalos de confianza contienen al cero, por lo cual, se acepta que la media de los diámetros de las barras de ambas máquinas es igual.

ii. Calculando el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{8.75-8.63}{\sqrt{\frac{1}{12}+\frac{1}{18}}} = 0.56889$$

Para una prueba bilateral o de dos colas, nótese que t_0 cae dentro del intervalo para ambos niveles de confianza, por lo cual, se acepta que las medias de los diámetros de las barras de ambas máquinas son iguales.

iii. Se calcula la probabilidad $p = p(t > t_0)$ o $p = p(t < -t_0)$, utilizando la distribución t de Student.

$$p(t > t_0) = 0.28698 > 0.025 > 0.005$$

Por lo cual, se acepta que las medias de los diámetros de las barras de ambas máquinas son iguales.

5.3.4. Prueba de hipótesis para demostrar la igualdad de medias de dos poblaciones normales o con tamaños de muestras grandes, con varianzas poblacionales desconocidas y diferentes

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas y diferentes, si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{5.53}$$

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

$$Si \qquad t_0 > t_{\frac{\alpha}{2},\upsilon} \text{ o } t_0 < -t_{\frac{\alpha}{2},\upsilon} \tag{5.54}$$

Donde

$$\upsilon = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2 \tag{5.55}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -t_{\frac{\alpha}{2},\upsilon} \leq t_0 \leq t_{\frac{\alpha}{2},\upsilon} \tag{5.56}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

$$Si \qquad t_0 > t_{\alpha,\upsilon} \qquad \text{se rechaza la hipótesis nula} \tag{5.57}$$

De lo contrario, si $t_0 < t_{\alpha,\upsilon}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

$$Si \qquad t_0 < t_{\frac{\alpha}{2},\upsilon} \qquad \text{se rechaza la hipótesis nula} \tag{5.58}$$

De lo contrario, si $t_0 > t_{\alpha,\upsilon}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(t > t_0\right)$ o $p = p\left(t < -t_0\right)$, utilizando la distribución t de Student.

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Unilateral_Superior (5.59)
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

Unilateral_Inferior
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico t; la prueba puede ser de dos colas $t_{\alpha/2}$, o unilateral t_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.17

El proceso de unión entre un soporte de acero y un contacto de plata se puede llevar a cabo por uno de dos procesos diferentes: por soldadura de aposte o por disparo termomagnético. Al utilizar el proceso de soldadura por disparo termomagnético, queda un residuo que da la impresión de que la pastilla de plata se

encuentra flameada, lo cual provoca que se deba hacer un proceso de limpieza que encarece al proceso. Se realiza un experimento para determinar cuál de los dos procesos arroja una mayor resistencia mecánica en la unión del soporte con la pastilla, para lo cual se toma una muestra de $n_1 = 16$ en el proceso de soldadura de aposte y $n_2 = 20$ en el proceso de unión por disparo termomagnético. Suponga que las poblaciones son infinitas. Los resultados obtenidos se muestran a continuación:

Proceso	n	media	varianza
Soldadura de Aposte	16	425	800
Disparo Termomagnético	20	445	3200

a. Realice una prueba de hipótesis para demostrar que las varianzas de ambos procesos son diferentes.

$$F_{0.025,15,19} = 0.36062 F_{0.975,15,19} = 2.61712$$
$$F_{0.005,15,19} = 0.25558 F_{0.975,15,19} = 3.58657$$

$$F_0 = \frac{S^2_{n_1-1}}{S^2_{n_2-1}} = \frac{800}{3300} = 0.24242$$

Como se puede apreciar, $F_0 = 0.24242 < 0.25558 < 0.36062$, cae fuera del intervalo de aceptación de la igualdad entre varianzas para ambos niveles de confianza al 95% y al 99%, por lo cual se comprueba que las varianzas son diferentes.

b. Aplique una prueba de hipótesis para determinar si las medias en los diámetros de las barras de ambas máquinas de extruido son iguales.

i. Obteniendo un intervalo de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas pero iguales, de acuerdo con lo demostrado en el inciso a.

$$\upsilon = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2 = \frac{\left(\frac{800}{16} + \frac{3300}{20}\right)^2}{\frac{\left(\frac{800}{16}\right)^2}{17} + \frac{\left(\frac{3300}{20}\right)^2}{21}} - 2 = 32.0231$$

$t_{0.025,34} = 2.03693$

$t_{0.005,34} = 2.73848$

$$\left(\bar{x}_1 - \bar{x}_2\right) - t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(x_1 - x_2\right) + t_{\frac{\alpha}{2},\upsilon}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$-49.867 \leq \mu_1 - \mu_2 \leq 9.86731$ *al*_95%

$-60.154 \leq \mu_1 - \mu_2 \leq 20.154$ *al*_99%

Como se puede apreciar, ambos intervalos de confianza contienen al cero, por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

ii. Calculando el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{425 - 445}{\sqrt{\frac{800}{16} + \frac{3300}{20}}} = -1.364$$

Para una prueba bilateral o de dos colas, nótese que t_0 cae dentro del intervalo para ambos niveles de confianza, por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

iii. Se calcula la probabilidad $p = p\left(t > t_0\right)$ o $p = p\left(t < -t_0\right)$, utilizando la distribución t de Student.

$p\left(t > t_0\right) = 0.09104 > 0.025 > 0.005$

Por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

5.3.5. Prueba de hipótesis para demostrar la igualdad de medias de dos poblaciones normales, con varianzas poblacionales desconocidas y diferentes para tamaños de muestra no grandes ($n < 30$)

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas y diferentes si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{5.60}$$

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

$$Si \qquad t_0 > t_{\frac{\alpha}{2},\upsilon} \quad \text{o} \quad t_0 < -t_{\frac{\alpha}{2},\upsilon} \tag{5.61}$$

Donde

$$\upsilon = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -t_{\frac{\alpha}{2},\upsilon} \leq t_0 \leq t_{\frac{\alpha}{2},\upsilon} \tag{5.62}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

$$Si \qquad t_0 > t_{\alpha,\upsilon} \text{ se rechaza la hipótesis nula} \tag{5.63}$$

De lo contrario, si $t_0 < t_{\alpha,\upsilon}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

$$Si \qquad t_0 < t_{\frac{\alpha}{2},\upsilon} \text{ se rechaza la hipótesis nula} \tag{5.64}$$

De lo contrario, si $t_0 > t_{\alpha,\upsilon}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p(t > t_0)$ o $p = p(t < -t_0)$, utilizando la distribución t de Student.

Para una prueba bilateral o de dos colas, donde H_1: $\mu_1 \neq \mu_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 > \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $\mu_1 < \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Unilateral_Superior (5.65)
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

Unilateral_Inferior
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico *t*; la prueba puede ser de dos colas $t_{\alpha/2}$, o unilateral ta.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.17

El proceso de unión entre un soporte de acero y un contacto de plata se puede llevar a cabo por uno de dos procesos diferentes: por soldadura de aposte o por disparo termomagnético. Al utilizar el proceso de soldadura por disparo termomagnético, queda un residuo que da la impresión de que la pastilla de plata se encuentra flameada, lo cual provoca que se deba hacer un proceso de limpieza que encarece al proceso. Se realiza un experimento para determinar cuál de los dos procesos arroja una mayor resistencia mecánica en la unión del soporte con la pastilla, para lo cual se toma una muestra de $n_1 = 16$ en el proceso de soldadura de aposte y $n_2 = 20$ en el proceso de unión por disparo termomagnético. Suponga que las poblaciones son infinitas. Los resultados obtenidos se muestran a continuación:

Proceso	n	media	Varianza
Soldadura de Aposte	16	425	800
Disparo Termomagnético	20	445	3200

a. Realice una prueba de hipótesis para demostrar que las varianzas de ambos procesos son diferentes.

$$F_{0.025,15,19} = 0.36062 \quad F_{0.975,15,19} = 2.61712$$

$$F_{0.005,15,19} = 0.25558 \quad F_{0.975,15,19} = 3.58657$$

$$F_0 = \frac{S^2_{n_1-1}}{S^2_{n_2-1}} = \frac{800}{3300} = 0.24242$$

Como se puede apreciar, $F_0 = 0.24242 < 0.25558 < 0.36062$ cae fuera del intervalo de aceptación de la igualdad entre varianzas para ambos niveles de confianza al 95% y al 99%, por lo cual se comprueba que las varianzas son diferentes.

b. Aplique una prueba de hipótesis para determinar si las medias en los diámetros de las barras de ambas máquinas de extruido son iguales.

   i. Obteniendo un intervalo de confianza, para la diferencia entre medias, con el estadístico t de Student, suponiendo varianzas poblacionales desconocidas pero iguales, de acuerdo con lo demostrado en el inciso *a*.

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2 = \frac{\left(\frac{800}{16} + \frac{3300}{20}\right)^2}{\frac{\left(\frac{800}{16}\right)^2}{17} + \frac{\left(\frac{3300}{20}\right)^2}{21}} - 2 = 32.0231$$

$t_{0.025,34} = 2.03693$

$t_{0.005,34} = 2.73848$

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$-49.867 \leq \mu_1 - \mu_2 \leq 9.86731$ *al*_95%

$-60.154 \leq \mu_1 - \mu_2 \leq 20.154$ *al*_99%

Como se puede apreciar, ambos intervalos de confianza contienen al cero, por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

ii. Al Calcular el estadístico de prueba

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{425 - 445}{\sqrt{\frac{800}{16} + \frac{3300}{20}}} = -1.364$$

Para una prueba bilateral o de dos colas, nótese que t_0 cae dentro del intervalo para ambos niveles de confianza, por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

iii. Se calcula la probabilidad $p = p(t > t_0)$ o $p = p(t < -t_0)$, utilizando la distribución t de Student.

$p(t < -t_0) = 0.09104 > 0.025 > 0.005$

Por lo cual, se acepta que la media de resistencia de la unión es la misma en ambos procesos.

5.3.6. Prueba de hipótesis para demostrar la igualdad entre medias para dos poblaciones normales con observaciones pareadas

Suponga que se obtiene la diferencia entre cada una de las observaciones pareadas, es decir, $D_1 = x_{11} - x_{21}$, $D_2 = x_{12} - x_{22}$, $\cdots$, $D_n = x_{1n} - x_{2n}$.

La media de la población de diferencias, dado que se supone que las variables x_1 y x_2 son normales, también será normal.

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al $100(1-\alpha)\%$ de nivel de confianza, para la diferencia entre medias, para dos poblaciones normales con observaciones pareadas, con el estadístico t de Student si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$t_0 = \frac{\overline{D}}{\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)} \qquad N\_finita \tag{5.66}$$

$$t_0 = \frac{\overline{D}}{\frac{S_D}{\sqrt{n}}} \qquad N\_infinita$$

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

$$Si \qquad t_0 > t_{\frac{\alpha}{2},n-1} \quad \text{o} \quad t_0 < t_{\frac{\alpha}{2},n-1} \tag{5.67}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -t_{\frac{\alpha}{2},n-1} \leq t_0 \leq t_{\frac{\alpha}{2},n-1} \tag{5.68}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

$$Si \qquad t_0 > t_{\alpha,n-1} \text{ se rechaza la hipótesis nula} \tag{5.69}$$

De lo contrario, si $t_0 < t_{\alpha,n-1}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

$$Si \qquad t_0 < t_{\frac{\alpha}{2},n-1} \quad \text{se rechaza la hipótesis nula} \tag{5.70}$$

De lo contrario, si $t_0 > t_{\alpha,n-1}$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(t > t_0\right)$ o $p = p\left(t < -t_0\right)$, utilizando la distribución t de Student.

Para una prueba bilateral o de dos colas, donde $H_1 : \mu_1 \neq \mu_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 > \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : \mu_1 < \mu_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : D = \mu_1 - \mu_2 = 0$
$H_1 : D \neq 0$

Unilateral_Superior (5.71)
$H_0 : D = \mu_1 - \mu_2 = 0$
$H_1 : D > 0$

Unilateral_Inferior
$H_0 : D = \mu_1 - \mu_2 = 0$
$H_1 : D < 0$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico t; la prueba puede ser de dos colas $t_{\alpha/2,n-1}$, o unilateral $t_{\alpha,n-1}$.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.18

Formule una prueba de hipótesis para verificar que la gasolina premium da mayor rendimiento en kilometraje por litro consumido que la magna. Para ello, se seleccionan ocho automóviles de diferentes marcas y se hacen las pruebas siguiendo la autopista de la gasolinera "Qué Chula es Puebla" a la gasolinera de la entrada a Puebla en la av. Aquiles Serdán; en cada automóvil usado se hace el mismo recorrido, bajo las mismas condiciones con gasolina premium y con gasolina magna, obteniendo los siguientes resultados:

	Rendimiento km/l	
Automóvil	Magna	Premium
Fiat 500	18.08	20.00
Civic Honda	13.09	17.23
Mazda 3	22.98	23.55
Mercedes C200	16.07	25.55
Prius C	26.17	26.59
Audi A3	22.77	24.65
BMW320iA	25.00	25.11
Ford Focus	19.33	21.56

Cabe señalar que los datos fueron generados artificialmente con distribución normal, tomando como referencia parámetros reales de la PROFECO.

Automóvil	Rendimiento km/l		
	Magna	Premium	D
Fiat 500	18.08	20.00	1.92
Civic Honda	13.09	17.23	4.14
Mazda 3	22.98	23.55	0.57
Mercedes C200	16.07	25.55	9.48
Prius C	26.17	26.59	0.42
Audi A3	22.77	24.65	1.88
BMW320iA	25.00	25.11	0.11
Ford Focus	19.33	21.56	2.23
Media =	20.44	23.03	2.59
DesvEst =	4.56	3.19	3.07

La prueba de hipótesis a realizar es la siguiente:

$H_0 : D = \mu_1 - \mu_2 = 0$
$H_1 : D > 0$

i. Obteniendo un intervalo de confianza, para la diferencia entre medias, con el estadístico z

$t_{0.05,7} = 1.89458$ *al_95%*
$t_{0.01,7} = 2.99795$ *al_99%*

Con los datos dados, se obtiene el siguiente intervalo de confianza:

$$2.59 - 1.89458 * \frac{3.07}{\sqrt{8}} = 0.53361 \leq D \qquad al\_95\%$$

$$-0.664 \leq D \qquad al\_99\%$$

Nótese que al 95% el intervalo no contiene al cero, por lo que se rechaza la hipótesis nula; sin embargo, al 99% el intervalo sí contiene al cero, por lo cual se cae en zona de duda, se sugiere aumentar el tamaño de muestra para hacer una afirmación más contundente.

ii. Calculando el estadístico de prueba

$$t_0 = \frac{\overline{D}}{\frac{S_D}{\sqrt{n}}} = 2.3862$$

Nótese que t_0 no cae dentro del intervalo para el 95% de nivel de confianza, pero si para el 99% de nivel de confianza, por lo cual, se sugiere aumentar el tamaño de muestra para hacer una afirmación más contundente.

iii. Se calcula la probabilidad $p = p\left(t > t_0\right)$, utilizando la distribución t de Student.

$p\left(t > 2.3862\right) = 0.02422$

Pero $0.025 > 0.02422 > 0.005$

Por lo cual, se sugiere aumentar el tamaño de muestra para hacer una afirmación más contundente.

5.3.7. Prueba de hipótesis para demostrar la igualdad entre proporciones o fracciones para dos poblaciones normales o con tamaños de muestras muy grandes

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia entre proporciones, para dos poblaciones normales o con tamaños de muestras muy grandes, con el estadístico z, si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}} \qquad N\_finita \tag{5.72}$$

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\left(1 - \hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1 - \hat{p}_2\right)}{n_2}}} \qquad N\_infinita$$

Para una prueba bilateral o de dos colas, donde H_1: $p_1 \neq p_2$

$$Si \qquad z_0 > z_{\frac{\alpha}{2}} \quad \text{o} \quad z_0 < -z_{\frac{\alpha}{2}} \tag{5.73}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -z_{\frac{\alpha}{2}} \leq z_0 \leq -z_{\frac{\alpha}{2}} \tag{5.74}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $p_1 > p_2$

$$Si \qquad z_0 > z_\alpha \text{ se rechaza la hipótesis nula} \tag{5.75}$$

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $p_1 < p_2$

$$Si \qquad z_0 < z_{\frac{\alpha}{2}} \text{ se rechaza la hipótesis nula} \tag{5.76}$$

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(z > z_0\right)$ o $p = p\left(z < -z_0\right)$, utilizando la distribución z de la normal.

Para una prueba bilateral o de dos colas, donde $H_1 : p_1 \neq p_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p_1 > p_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde $H_1 : p_1 < p_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$

Unilateral_Superior (5.77)
$H_0 : p_1 = p_2$
$H_1 : p_1 > p_2$

Unilateral_Inferior
$H_0 : p_1 = p_2$
$H_1 : p_1 < p_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico *z*; la prueba puede ser de dos colas $z_{\alpha/2}$, o unilateral t_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.19

En la Facultad de Ingeniería de la UNAM, se ha implantado una nueva modalidad en la aplicación de exámenes extraordinarios, denominada exámenes extraordinarios en tres etapas, la cual pretende elevar el nivel de acreditación en dichos exámenes. A continuación, se presentan los resultados publicados en el Informe de Actividades 2015 (https://www.planeacion.unam.mx/informes/PDF/FI-2015-2016.pdf). Suponga que existen 13 asignaturas de la División de Ciencias Básicas.

Asignatura	Semestre 2016-1			Semestre 2015-2		
	Inscritos	Presentados	% Aprobados	Inscritos	Presentados	% Aprobados
Álgebra	94	89	43.62	164	143	36.08
Álgebra Lineal	141	139	24.11	277	254	18.68
Cáculo Diferencial	155	153	40.65	326	307	32.03
Cálculo Integral	151	146	25.83	267	240	20.03
Ecuaciones Diferenciales	192	183	27.6	246	225	18.91
Geometría Analítica	198	192	31.31	212	187	25.57
Total	931	902	30.63	1492	1356	23.59

El porcentaje de aprobados en cada semestre fue obtenido como una media armónica en vez de un promedio, en virtud de tratarse de porcentajes, por lo que difiere de la publicación citada.

Realice una prueba de hipótesis para demostrar que la modalidad de exámenes extraordinarios en tres etapas sí ha aumentado el porcentaje de aprobados en dichos exámenes

Semestre	N	n	p
2016-1	13	6	0.3063
2015-2	13	6	0.2521

La prueba de hipótesis a realizar es la siguiente:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

i. Obteniendo un intervalo de confianza, para la diferencia entre proporciones, con el estadístico z

$z_{0.05} = 1.644854 \qquad al\_95\%$
$z_{0.01} = 2.326335 \qquad al\_99\%$

Con los datos dados, se obtienen los siguientes intervalos de confianza:

$$p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

$p_1 - p_2 \leq 0.324785 \qquad al\_95\%$
$p_1 - p_2 \leq 0.729935 \qquad al\_99\%$

Ambos intervalos contienen al valor cero, por lo cual no se tiene evidencia estadística que demuestre que la proporción de aprobados se ha elevado.

ii. Calculando el estadístico de prueba

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}} = 0.274492$$

Nótese que z_0 cae dentro de ambos intervalos, por lo cual no existe evidencia estadística para afirmar que se ha elevado la proporción de aprobados debido al examen extraordinario en tres etapas.

iii. Se calcula la probabilidad $p = p(z > z_0)$, utilizando la distribución normal.

$p(p > 0.2744992) = 0.39185$

$p = 0.39185 > 0.05 > 0.01$

Por lo cual, no existe evidencia estadística para afirmar que se ha elevado la proporción de aprobados debido al examen extraordinario en tres etapas.

5.3.8. Prueba de hipótesis para demostrar la igualdad entre fracciones de defectos o éxitos por unidad para dos poblaciones normales o con tamaños de muestras grandes

La prueba puede llevarse a cabo de tres formas diferentes:

i. Obtener un intervalo de confianza al 100(1–α)% de nivel de confianza, para la diferencia entre fracciones de defectos o éxitos, para dos poblaciones normales o con tamaños de muestras muy grandes, con el estadístico z; si dicho intervalo no contiene al valor cero entonces se tiene evidencia estadística para rechazar la hipótesis nula. Si el intervalo contiene al valor cero entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello. El intervalo de confianza puede ser bilateral, unilateral inferior o unilateral superior, dependiendo de lo que se pretenda demostrar.

ii. Se calcula el estadístico de prueba

$$z_0 = \frac{\hat{u}_1 - \hat{u}_2}{\sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\hat{u}_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}} \qquad N\_finita \tag{5.78}$$

$$z_0 = \frac{\hat{u}_1 - \hat{u}_2}{\sqrt{\frac{\hat{u}_1}{n_1} + \frac{\hat{u}_2}{n_2}}} \qquad N\_infinita$$

Para una prueba bilateral o de dos colas, donde H_1: $u_1 \neq u_2$

$$Si \qquad z_0 > z_{\alpha/2} \quad \text{o} \quad z_0 < -z_{\alpha/2} \tag{5.79}$$

Se rechaza la hipótesis nula

De lo contrario,

$$Si \qquad -z_{\alpha/2} \leq z_0 < z_{\alpha/2} \tag{5.80}$$

Entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $u_1 > u_2$

Si $z_0 > z_\alpha$ se rechaza la hipótesis nula (5.81)

De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $u_1 < u_2$ (5.82)

Si $z_0 < z_{\alpha/2}$ se rechaza la hipótesis nula
De lo contrario, si $z_0 < z_\alpha$ entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

iii. Se calcula la probabilidad $p = p\left(z > z_0\right)$ o $p = p\left(z < -z_0\right)$, utilizando la distribución z de la normal.

Para una prueba bilateral o de dos colas, donde H_1: $u_1 \neq u_2$

Si $p < \alpha/2$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha/2$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $u_1 > u_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Para una prueba unilateral, donde H_1: $u_1 < u_2$

Si $p < \alpha$ se rechaza la hipótesis nula. De lo contrario, si $p > \alpha$, entonces no se tiene suficiente evidencia estadística para poder rechazar la hipótesis nula y se termina por aceptarla aunque no se esté convencido de ello.

Se aplican los pasos establecidos para llevar a cabo una prueba de hipótesis:

a. Se plantean las hipótesis nula y alternativa.

Bilateral
$H_0 : u_1 = u_2$
$H_1 : u_1 \neq u_2$

Unilateral_Superior (5.83)
$H_0 : u_1 = u_2$
$H_1 : u_1 > u_2$

Unilateral_Inferior
$H_0 : u_1 = u_2$
$H_1 : u_1 < u_2$

b. Se selecciona el nivel de significancia α, el cual puede ser 0.10, 0.05, 0.01, 0.0027 o el que establezca una norma o especificación. En la práctica, a veces es conveniente hacer la prueba para dos niveles de significancia al 0.05 y al 0.01; si se rechaza o no para ambos niveles de significancia, la prueba es contundente; si se rechaza para 0.05 pero no para 0.01, se cae en zona de duda y se sugiere tomar una muestra mayor.

c. Se identifica el estadístico de prueba. Para este caso particular, se trata del estadístico z; la prueba puede ser de dos colas $z_{\alpha/2}$, o unilateral z_{α}.

d. Se formula la regla de decisión de acuerdo con uno de los métodos establecidos anteriormente.

e. Se toma una muestra de cada población, se calcula el valor del estadístico de prueba, se compara con el valor crítico y se decide si se rechaza H_0 o se concluye que no existe evidencia estadística suficiente para rechazar la hipótesis nula.

Ejercicio 5.20

El jefe de un corporativo necesita contratar a una secretaria. Suponga que existen dos empleadas que trabajan en dos áreas diferentes y decide hacer una prueba de hipótesis para elegir a la más adecuada. Por datos históricos las áreas donde trabajan actualmente tienen un expediente de cada una de ellas y realizan un cuadro comparativo sobre su desempeño. Los datos que le proporcionan de los últimos doce meses, son los siguientes:

mes	Empleada 1		Empleada 2	
	Actividades	No. Defectos	Actividades	No. Defectos
1	15	12	25	13
2	12	9	17	9
3	10	11	12	11
4	13	11	14	11
5	12	8	17	9
6	11	9	13	9
7	20	13	20	12
8	12	10	15	11
9	13	10	15	10
10	13	9	17	10
11	17	9	17	9
12	14	10	12	8
Total	162	121	194	122

Las actividades que realizan cada una de ellas son llamadas por teléfono, elaborar oficios, agendar reuniones, archivar documentos, actualizar directorio, etcétera. Los errores pueden ser no contestar una llamada telefónica o contestarla de mala gana, olvidar fechas de reuniones, no archivar adecuadamente un documento, no registrar los datos de un cliente, errores de tipografía o de ortografía en los documentos elaborados, etcétera. Suponga que el número de actividades que van a realizar en promedio es de 20 por mes.

Realice una prueba de hipótesis bilateral, al 95% y al 99% de nivel de confianza, para determinar si el número de errores que cometen por cada actividad en promedio es la misma entre ellas o hay diferencias.

La prueba de hipótesis a realizar es la siguiente:

$$H_0 : u_1 = u_2$$
$$H_1 : u_1 \neq u_2$$

i. Obteniendo un intervalo de confianza, para la diferencia entre la fracción de defectos por actividad de cada una de ellas, con el estadístico z

$z_{0.05} = 1.95996$ *al_95%*
$z_{0.01} = 2.57583$ *al_99%*

Con los datos dados, se obtienen los siguientes intervalos de confianza:

$$\left(\hat{u}_1 - \hat{u}_2\right) - z_{\alpha/2}\sqrt{\frac{\hat{u}_1}{n_1} + \frac{\hat{u}_2}{n_2}} \leq \left(u_1 - u_2\right) \leq \left(\hat{u}_1 - \hat{u}_2\right) + z_{\alpha/2}\sqrt{\frac{\hat{u}_1}{n_1} + \frac{\hat{u}_2}{n_2}}$$

$-0.396 \leq u_1 - u_2 \leq 0.632$ *al_95%*
$-0.5575 \leq u_1 - u_2 \leq 0.79363$ *al_99%*

Ambos intervalos contienen al valor cero, por lo cual no se tiene evidencia estadística que demuestre que una de las empleadas comete menos errores por actividad que la otra.

ii. Calculando el estadístico de prueba

$$z_0 = \frac{\hat{u}_1 - \hat{u}_2}{\sqrt{\frac{\hat{u}_1}{n_1} + \frac{\hat{u}_2}{n_2}}} = \frac{0.74691 - 0.62887}{\sqrt{\frac{0.74691}{20} + \frac{0.62887}{20}}} = 0.45009$$

Nótese que z_0 cae dentro de ambos intervalos al 95% y al 99%, por lo cual no existe evidencia estadística para afirmar que una de las empleadas comete menos errores por actividad que la otra.

iii. Se calcula la probabilidad $p = p\left(z > z_0\right)$, utilizando la distribución normal.

$p\left(z > 0.45009\right) = 0.32632$

$p = 0.32632 > 0.05 > 0.01$

Por lo cual, no existe evidencia estadística para afirmar que una de las empleadas comete menos errores por actividad que la otra.

Figura 5.6. Pruebas de Hipótesis para un mismo parámetro de dos poblaciones normales

Hipótesis nula	Condiciones iniciales	Valor del estadístico de prueba	Hipótesis alternativa	Criterio rechazo H_0	Criterio rechazo H_0	Parámetro k de la CCO
$H_0:\sigma_1^2=\sigma_2^2$	$x_1\sim N(\mu_1,\sigma_1)$ $x_2\sim N(\mu_2,\sigma_2)$ N_1 Finita N_2 Finita	$F_0=\dfrac{S_1^2\left(\frac{N_1-n_1}{N_1-1}\right)}{S_2^2\left(\frac{N_2-n_2}{N_2-1}\right)}$	$H_1:\sigma_1^2\neq\sigma_2^2$	$F_0>F_{\frac{\alpha}{2},n_1-1,n_2-1}$ $F_0<F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$	$p=p(F>F_0)<\alpha/2$	$\lambda=\sigma_1/\sigma_2$
			$H_1:\sigma_1^2>\sigma_2^2$	$F_0>F_{\alpha,n_1-1,n_2-1}$	$p=p(F>F_0)<\alpha/2$	$\lambda=\sigma_1/\sigma_2$
$H_0:\mu_1=\mu_2$	$x_1\sim N(\mu_1,\sigma_1)$ o n_1 grande $x_2\sim N(\mu_2,\sigma_2)$ o n_2 grande N_1,N_2 Finitas σ_1^2,σ_2^2 Conocidas	$z_0=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\sigma_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}}$	$H_1:\mu_1\neq\mu_2$	$z_0>z_{\alpha/2}$ $z_0<-z_{\alpha/2}$	$p=p(z>z_0)<\alpha/2$ $p=p(z<-z_0)<\alpha/2$	$k=\lvert\mu_1-\mu_2\rvert/\sqrt{\sigma_1^2+\sigma_2^2}$
			$H_1:\mu_1>\mu_2$	$z_0>z_\alpha$	$p<\alpha$	$k=(\mu_1-\mu_2)/\sqrt{\sigma_1^2+\sigma_2^2}$
			$H_1:\mu_1<\mu_2$	$z_0<-z_\alpha$	$p<\alpha$	$k=(\mu_2-\mu_1)/\sqrt{\sigma_1^2+\sigma_2^2}$
$H_0:\mu_1=\mu_2$	$x_1\sim N(\mu_1,\sigma_1)$ o n_1 grande $x_2\sim N(\mu_2,\sigma_2)$ o n_2 grande N_1,N_2 Finitas σ_1^2,σ_2^2 Desconocidas pero $\sigma_1^2=\sigma_2^2$	$t_0=\dfrac{\bar{x}_1-\bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ Donde $Sp=\sqrt{\dfrac{(n_1-1)S_1^2\left(\frac{N_1-n_1}{N_1-1}\right)+(n_2-1)S_2^2\left(\frac{N_2-n_2}{N_2-1}\right)}{n_1+n_2-2}}$	$H_1:\mu_1\neq\mu_2$	$t_0>t_{\frac{\alpha}{2},n_1+n_2-2}$ $t_0<-t_{\frac{\alpha}{2},n_1+n_2-2}$	$p=p(t>t_0)<\alpha/2$ $p=p(t<-t_0)<\alpha/2$	$k=\lvert\mu_1-\mu_2\rvert/2\sigma$
			$H_1:\mu_1>\mu_2$	$t_0>t_{\alpha,n_1+n_2-2}$	$p=p(t>t_0)<\alpha$	$k=(\mu_1-\mu_2)/2\sigma$
			$H_1:\mu_1<\mu_2$	$t_0<-t_{\alpha,n_1+n_2-2}$	$p=p(t<-t_0)<\alpha$	$k=(\mu_2-\mu_1)/2\sigma$
$H_0:\mu_1=\mu_2$	$x_1\sim N(\mu_1,\sigma_1)$ o n_1 grande $x_2\sim N(\mu_2,\sigma_2)$ o n_2 grande N_1,N_2 Finitas $\sigma_1^2\neq\sigma_2^2$ Desconocidas	$z_0=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}}$ $\upsilon=\dfrac{\left(\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)\right)^2}{n_1+1}+\frac{\left(\frac{S_2^2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)\right)^2}{n_2+1}}-2$	$H_1:\mu_1\neq\mu_2$	$t_0>t_{\frac{\alpha}{2},\upsilon}$ $t_0<-t_{\frac{\alpha}{2},\upsilon}$	$p=p(t>t_0)<\alpha/2$ $p=p(t<-t_0)<\alpha/2$	$k=\lvert\mu_1-\mu_2\rvert/2\sigma$
			$H_1:\mu_1>\mu_2$	$t_0>t_{\alpha,\upsilon}$	$p=p(t>t_0)<\alpha$	$k=(\mu_1-\mu_2)/2\sigma$
			$H_1:\mu_1<\mu_2$	$t_0<-t_{\alpha,\upsilon}$	$p=p(t<-t_0)<\alpha$	$k=(\mu_2-\mu_1)/2\sigma$
$H_0:D=\mu_1-\mu_2=0$	$x_1\sim N(\mu_1,\sigma_1)$ o n_1 grande $x_2\sim N(\mu_2,\sigma_2)$ o n_2 grande $N_1=N_2=N$ Finita $n_1=n_2=n$ Observaciones pareadas	$t_0=\dfrac{\bar{D}}{\frac{S_D}{\sqrt{n}}\left(\frac{N-n}{N-1}\right)}$	$H_1:D\neq 0$	$t_0>t_{\frac{\alpha}{2},n-1}$ $t_0<-t_{\frac{\alpha}{2},n-1}$	$p=p(t>t_0)<\alpha/2$ $p=p(t<-t_0)<\alpha/2$	$k=\lvert D\rvert/\sigma$
			$H_1:D>0$	$t_0>t_{\alpha,n-1}$	$p=p(t>t_0)<\alpha$	$k=D/\sigma$
			$H_1:D<0$	$t_0<-t_{\alpha,n-1}$	$p=p(t<-t_0)<\alpha$	$k=-D/\sigma$
$H_0:p_1=p_2$	$x_1\sim N\left(n_1\hat{p}_1,\sqrt{n_1\hat{p}_1(1-\hat{p}_1)\left(\frac{N_1-n_1}{N_1-1}\right)}\right)$ $x_2\sim N\left(n_2\hat{p}_2,\sqrt{n_2\hat{p}_2(1-\hat{p}_2)\left(\frac{N_2-n_2}{N_2-1}\right)}\right)$ N_1,N_2 Finitas	$z_0=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}}$	$H_1:p_1\neq p_2$	$z_0>z_{\alpha/2}$ $z_0<-z_{\alpha/2}$	$p=p(z>z_0)<\alpha/2$ $p=p(z<-z_0)<\alpha/2$	
			$H_1:p_1>p_2$	$z_0>z_\alpha$	$p<\alpha$	
			$H_1:p_1<p_2$	$z_0<-z_\alpha$	$p<\alpha$	
$H_0:u_1=u_2$	$x_1\sim N\left(\hat{u}_1,\sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)}\right)$ $x_2\sim N\left(\hat{u}_2,\sqrt{\frac{\hat{u}_2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}\right)$ N_1,N_2 Finitas	$z_0=\dfrac{\hat{u}_1-\hat{u}_2}{\sqrt{\frac{\hat{u}_1}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right)+\frac{\hat{u}_2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}}$	$H_1:u_1\neq u_2$	$z_0>z_{\alpha/2}$ $z_0<-z_{\alpha/2}$	$p=p(z>z_0)<\alpha/2$ $p=p(z<-z_0)<\alpha/2$	
			$H_1:u_1>u_2$	$z_0>z_\alpha$	$p<\alpha$	
			$H_1:u_1<u_2$	$z_0<-z_\alpha$	$p<\alpha$	

5.4. Pruebas de bondad de ajuste entre una muestra obtenida empíricamente y la distribución de probabilidad de un modelo teórico dado

En lo correspondiente a la Inferencia Estadística está integrada por el capítulo cuatro, donde se analizaron los estimadores puntuales y por intervalos, y por el capítulo cinco, donde se han desarrollado las pruebas de hipótesis; ambos capítulos parten de suposiciones iniciales como lo son:

a. Las muestras obtenidas deben ser representativas, lo que también indica que deben ser aleatorias.
b. Debe existir independencia estadística entre un estadístico y otro.
c. La distribución de probabilidad de un estadístico debe obedecer a uno de los modelos probabilísticos (en la mayoría de dichas condiciones que se formularon se supuso que la distribución de la población es normal).

Para poder hacer uso de todas las expresiones deducidas en los capítulos cuatro y cinco, debe comprobarse el cumplimiento de estas condiciones iniciales, de otra forma, no hay la seguridad de que las decisiones que se tomen sean realmente verdaderas, de allí la necesidad de aplicar criterios o pruebas para verificar el cumplimiento de dichas condiciones. Cuando se cumplen las condiciones iniciales se pueden emplear todas las expresiones deducidas anteriormente, a lo cual se le denomina Estadística Paramétrica.

Las pruebas paramétricas, por ejemplo, de las distribuciones t o F, se basan en una variedad de fuertes suposiciones a las que está sujeto su uso. Cuando los datos de una investigación pueden ser analizados adecuadamente por una prueba paramétrica, se tiene el medio más poderoso para rechazar una hipótesis nula H_0 falsa. Sin embargo, las condiciones iniciales que deben cumplirse pueden ser estrictas y no realizarse en muchos casos. Por ejemplo, las condiciones en las que la prueba de la distribución t es la más poderosa, y sin las cuales no se puede tener confianza en cualquier aseveración de probabilidad obtenida, son por lo menos las siguientes:

i. Las observaciones deben ser independientes entre sí.
ii. Las observaciones deben hacerse en poblaciones distribuidas normalmente.

En el caso del análisis de varianza por medio de la distribución F se agrega otra condición:

iii. Las medias de las poblaciones normales y homocedásticas deberán ser combinaciones lineales de efectos aditivos.

En estadística el concepto de homocedasticidad se aplica en modelos de regresión lineal múltiple y en modelos predictivos (que se verán en otros volúmenes de esta serie), los cuales presentan homocedasticidad cuando la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones. Se habla de homocedasticidad si el error cometido por el modelo tiene siempre la misma varianza. Cuando no se cumple esta situación, se dice que existe heterocedasticidad, que es cuando la varianza de cada término de perturbación no es un número constante.

La Estadística No Paramétrica es una rama de la Inferencia Estadística que estudia las pruebas y modelos estadísticos cuya distribución de probabilidad subyacente no se ajusta a los llamados criterios paramétricos. Su distribución no puede ser definida a priori, pues son los datos observados los que la determinan. La utilización de la estadística no paramétrica se hace recomendable cuando no se puede asumir que los datos se ajusten a una distribución conocida, como lo es el caso más común de la distribución normal.

En general, la potencia de las pruebas no paramétricas es menor que la potencia de las pruebas paramétricas equivalentes. Aun así, el uso adecuado de los tamaños muestrales disminuye la posibilidad de cometer errores tipo II, puesto que aumenta al mismo tiempo la eficacia de la prueba.

En este subtema se analizarán algunas pruebas paramétricas y no paramétricas para demostrar el cumplimiento de algunas de las condiciones iniciales, como lo son la verificación de que un parámetro de una población presenta una distribución de probabilidad conocida como lo es la normal.

Para iniciar considere los siguientes ejemplos:

Ejercicio 5.21 Un jugador profesional de dados pretende jugar una apuesta con un alumno de Estadística. Para iniciar el juego, el alumno impone como condición lanzar $n = 600$ veces el dado y registrar el número de veces que cae cada cara del dado.

Los resultados obtenidos fueron los siguientes:

Cara Superior	1	2	3	4	5	6
Frecuencia	72	125	108	106	118	71

Compruebe estadísticamente que el dado no se encuentra cargado.

Ejercicio 5.22

Los datos que se muestran a continuación representan el tiempo en horas de 40 dispositivos electrónicos que funcionaron desde el inicio de su operación hasta antes de su falla.

5403	3703	9810	34694
13780	25858	5083	543
1482	14263	10506	8157
947	6852	3229	4860
27090	817	28225	12785
2925	10306	573	14122
3572	1530	2407	3749
644	30807	6813	31642
27059	6372	896	5927
15675	12770	5687	195

Se sospecha que, por tratarse de tiempos de vida, el modelo probabilístico de comportamiento es exponencial negativo, compruebe esta hipótesis.

Ejercicio 5.23

Se escoge a 64 obreros para que cada uno de ellos extraiga una muestra de 50 artículos de lotes diferentes de tamaño $N = 600$, donde se sabe que existen $D = 100$ defectuosos. Cada obrero revisa su muestra y cuenta los artículos defectuosos que obtuvo. Suponga que los resultados fueron los siguientes:

5	7	8	5	7	4	12	8
10	9	3	14	7	8	7	6
6	6	12	7	10	7	6	6
5	8	5	4	10	6	11	6
10	7	9	4	8	10	11	6
9	9	6	7	11	8	9	7
6	12	11	5	9	5	13	5
11	8	9	6	6	9	5	10

Pruebe que estos datos presentan distribución normal.

Pruebas de hipótesis sobre normalidad de un conjunto de datos

Para determinar si una característica de calidad presenta una distribución de probabilidad conocida, se emplean alguno de los siguientes métodos para probarlo:

a. Comparación del histograma de frecuencias observado contra el histograma de probabilidad esperado.
b. Comparación de los valores observados contra los esperados con cierta distribución de probabilidad, y utilizando un gráfico de papel probabilístico.
c. Prueba de bondad de ajuste Ji-cuadrada o prueba de Pearson.
d. Prueba no paramétrica D de Kolmogorov- Smirnov o prueba de Lilliefors.
e. Otro tipo de pruebas exclusivamente sobre normalidad de una población, como Anderson-Darling, Cramer-Von Mises, Shapiro-Wilk, Shapiro-Francia, entre otros.

Cabe señalar que existen una gran cantidad de métodos para hacer el análisis de la bondad de ajuste o en particular de la normalidad de los datos y puede usarse una variedad copiosa de software para ello. En este volumen se emplean tres de ellos, Excel, Minitab y R, pero es indispensable conocer las bases teóricas, las condiciones iniciales, los criterios para aceptar o rechazar la hipótesis nula de cada uno de los métodos citados, para así poder aplicarlos.

5.4.1. Por comparación del histograma de frecuencias observado contra el histograma de probabilidad esperado

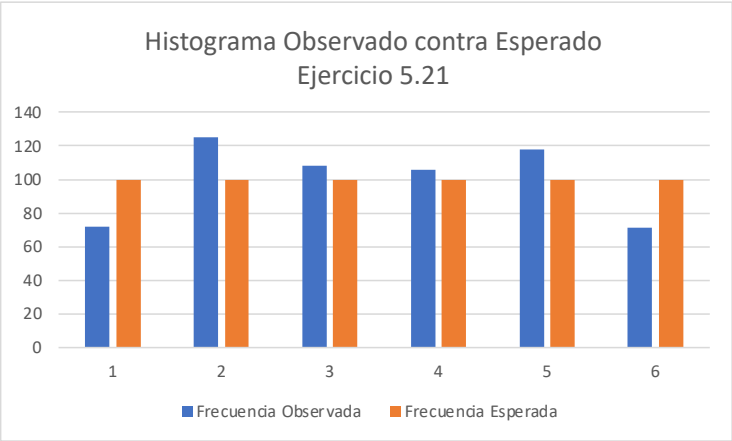
De todos los métodos existentes, este es el más gráfico y por lo mismo, más simple; sin embargo, es también el más débil en potencia de la prueba porque no aporta algún indicador estadístico que valide la decisión que se tome. Lo primero que se realiza es obtener una tabla de frecuencias observadas a partir de los datos de la muestra y agregar un renglón con las frecuencias esperadas del modelo teórico que se piensa que cumple, con base en los intervalos de clase que se definan. Posteriormente se trazan los histogramas observado y esperado, y se comparan a "ojo de pájaro". Para ilustrar la aplicación del método, se resolverán los ejercicios 5.21, 5.22 y 5.23, explicando cómo se construye la tabla de frecuencias y el trazado de su histograma, usando Excel, Minitab y R.

Para el ejercicio 5.21, la tabla de frecuencias se muestra a continuación:

Cara Superior	1	2	3	4	5	6
Frecuencia Observada	72	125	108	106	118	71
Frecuencia Esperada	100	100	100	100	100	100

En este caso no es necesario calcular las frecuencias esperadas, porque se trata de un dado con seis caras homogéneas y se aplica el enfoque de Laplace de que todas las caras tienen la misma posibilidad de caer, por lo que en 600 tiros se espera que cada cara caiga en promedio 100 veces; en consecuencia, la distribución de probabilidad de la cara que cae hacia arriba al lanzar un dado se supone que es uniforme discreta con espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$.

Su histograma se muestra a continuación:



La comparación visual entre ambos histogramas permite percibir claramente qué tanto se aproximan, pero no hay un indicador que mida que tanto se ajustan.

Para el ejercicio 5.22 la tabla de frecuencias es:

No.	Lim Inf Int	Lim Sup Int	Marca Clase	Frec Obs	Frec Esp
1	0	2500	1250	10	8.81387
2	2500	5000	3750	6	6.87176
3	5000	7500	6250	7	5.35759
4	7500	10000	8750	2	4.17706
5	10000	12500	11250	2	3.25666
6	12500	15000	13750	5	2.53907
7	15000	17500	16250	1	1.97959
8	17500	20000	18750	0	1.54339
9	20000	22500	21250	0	1.20331
10	22500	25000	23750	0	0.93817
11	25000	27500	26250	3	0.73144
12	27500	30000	28750	1	0.57027
13	30000	32500	31250	2	0.44462
14	32500	35000	33750	1	0.34665

Las frecuencias esperadas en la tabla anterior se calculan usando un modelo esperado exponencial negativo, donde su función de probabilidad acumulada está dada por la expresión

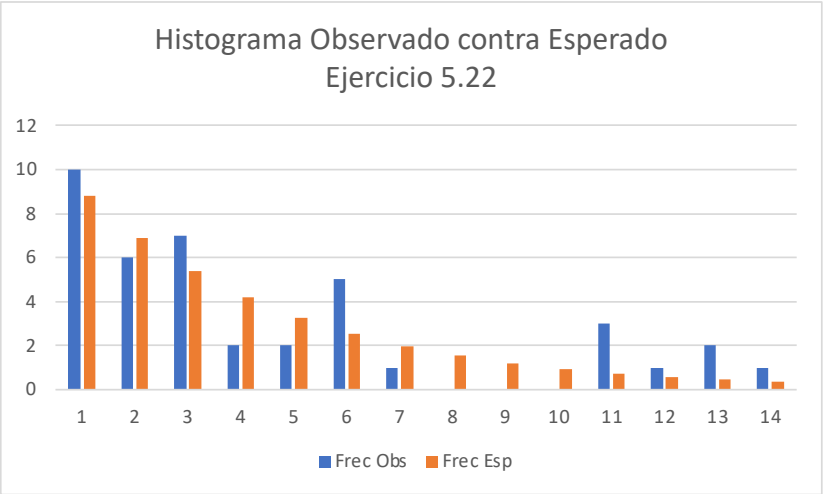
$$F(x) = 1 - e^{-\lambda x}$$

En donde el valor de $\hat{\lambda} = 1/\bar{x}$ se estima puntualmente como el recíproco de la media muestral de los datos dados. Las frecuencias esperadas de cada intervalo de clase se obtienen usando la siguiente expresión

$$f_1 = n * p\left(LimInf_i < x < LimSup_i\right) = n * \left[\left(1 - \exp\left(-\frac{LimSup_i}{\bar{x}}\right)\right) - \left(1 - \exp\left(-\frac{LimInf_i}{\bar{x}}\right)\right)\right]$$

$$f_1 = 50 * \left[\left(1 - \exp\left(-\frac{LimSup_i}{\bar{x}}\right)\right) - \left(1 - \exp\left(-\frac{LimInf_i}{\bar{x}}\right)\right)\right]$$

El histograma de frecuencias observadas contra esperadas se muestra a continuación:



La comparación visual entre ambos histogramas permite percibir claramente qué tanto se aproximan, pero no hay un indicador que mida qué tanto se ajustan.

En la figura anterior se observa que existe una inconsistencia en los datos dados, no existen frecuencias observadas en los intervalos 8, 9 y 10 correspondientes al intervalo entre 17500 y 23750 horas. Si las lecturas fueron bien tomadas, pues así deben considerarse; sin embargo, esto puede estar reflejando que se revolvieron datos de una población con datos de otra.

Para el ejercicio 5.23 la tabla de frecuencias es:

Marca Clase	Frec Acum Obs	Frec Obs	Frec Acum Esp	Frec Esp
0	0	0	0.000847313	0.054228009
1	0	0	0.003127909	0.145958141
2	0	0	0.009929104	0.435276509
3	1	1	0.027179179	1.104004769
4	4	3	0.064390864	2.381547853
5	12	8	0.132665913	4.369603144
6	24	12	0.239214675	6.819120751
7	33	9	0.380645954	9.051601893
8	40	7	0.54032859	10.21968871
9	48	8	0.693679729	9.814472903
10	54	6	0.818945601	8.017015784

Marca Clase	Frec Acum Obs	Frec Obs	Frec Acum Esp	Frec Esp
11	59	5	0.905980165	5.570212091
12	62	3	0.957415106	3.291836241
13	63	1	0.983268886	1.654641903
14	64	1	0.994321846	0.707389444
15	64	0	0.998340764	0.257210752
16	64	0	0.999583563	0.079539124
17	64	0	0.999910404	0.020917842
18	64	0	0.999983501	0.004678201
19	64	0	0.999997403	0.000889707
20	64	0	0.999999651	0.00014388

Las frecuencias esperadas en la tabla anterior se calculan usando un modelo esperado normal

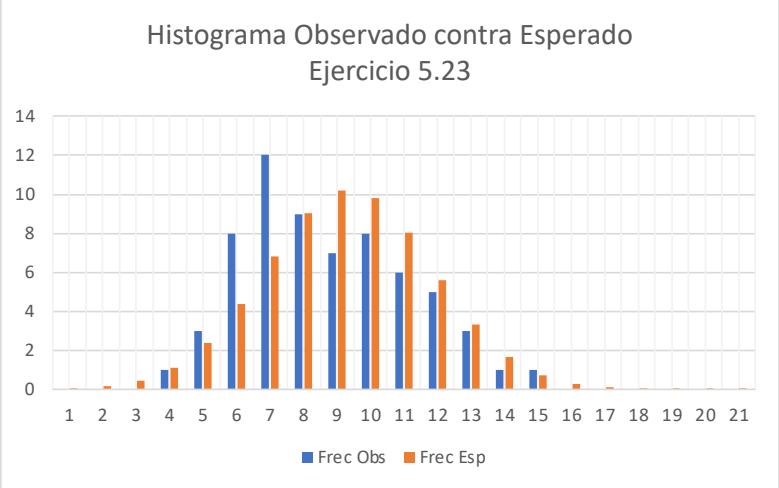
$$x - N\left(\hat{\mu}_x = \overline{x}, \hat{\sigma}_x = S_{n-1}\right)$$

Las frecuencias esperadas de cada intervalo de clase se obtienen usando la siguiente expresión

$$f_1 = np\left(LimInf_i < x < LimSup_i\right)$$

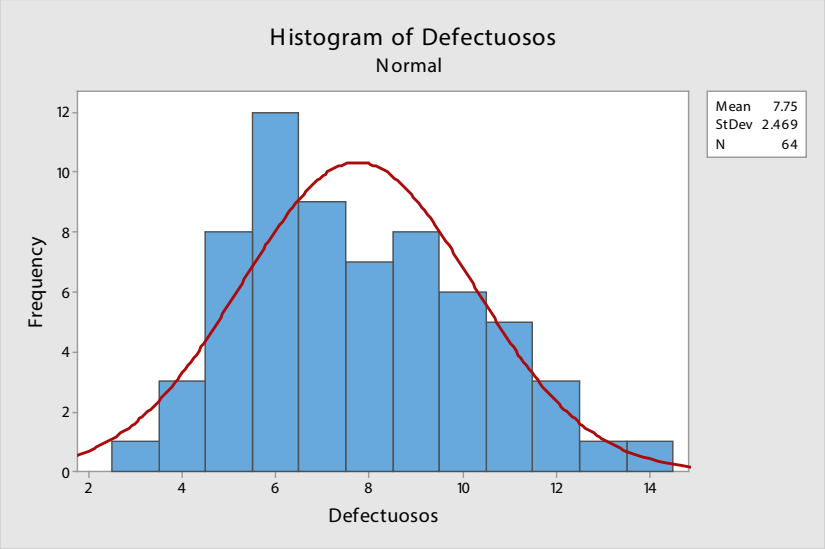
$$f_1 = n * \left[Distric.Norm.N\left(LimSup, \overline{x}, S_{n-1}\right) - Distric.Norm.N\left(LimInf, \overline{x}, S_{n-1}\right)\right]$$

El histograma de frecuencias observadas contra esperadas se muestra a continuación:



La comparación visual entre ambos histogramas permite percibir claramente qué tanto se aproximan, pero no hay un indicador que mida qué tanto se ajustan.

Para este ejemplo también se puede usar Minitab:



5.4.2. Por comparación de los valores observados contra los esperados con cierta distribución de probabilidad, y utilizando un gráfico de papel probabilístico

Este método es bastante antiguo y se aplicaba cuando no existían todavía las computadoras personales. Consistía en vaciar los datos dados, ordenados de menor a mayor, en un papel con una escala especial que representaba a la distribución de probabilidad que se suponía cumplían los datos. Así, uno podía ir a algunas papelerías especializadas y comprar el papel probabilístico normal, o papel probabilístico exponencial, o de la distribución de probabilidad que se requería. Actualmente es prácticamente imposible conseguir este tipo de papel, por lo que se utilizan las computadoras para graficarlo.

El principio básico del método del papel probabilístico es que al vaciar los datos ordenados de menor a mayor en dicho papel, si los puntos caen sobre una línea recta entonces se comportan como la función de probabilidad del papel adquirido. Para aplicar este método se requiere completar una tabla en la que en la primera columna se coloca un número consecutivo, empezando desde uno hasta n. En la segunda columna se coloca $(j\text{-}0.5)/n$, en donde j representa al número consecutivo situado a la izquierda. Estos valores obtenidos corresponden a la frecuencia observada acumulada (probabilidad esperada desde el punto de vista frecuentista o de Von Mises).

Al vaciar los datos de frecuencia observada contra frecuencia esperada, en el papel probabilístico ya diseñado, los puntos deben caer sobre una línea recta aproximadamente. Como ya no se reproducen estos papeles probabilísticos lo que se hace es calcular el valor de x esperado, para una frecuencia acumulada esperada dada por la expresión $(j\text{-}0.5)/n$; estos valores de x se obtienen como la inversa de la función de probabilidad esperada y se colocan como una tercer columna. Para ilustrar la aplicación del método del papel probabilístico se resolverán los problemas 5.22 y 5.23 anteriormente definidos.

Para el problema 5.22 se conforma la tabla siguiente:

j	(j-0.5)/n	Observado	Esperado
1	0.0125	195	126.341
2	0.0375	543	383.892
3	0.0625	573	648.222
4	0.0875	644	919.696

j	(j-0.5)/n	Observado	Esperado
5	0.1125	817	1198.71
6	0.1375	896	1485.7
7	0.1625	947	1781.13
8	0.1875	1482	2085.52
9	0.2125	1530	2399.42
10	0.2375	2407	2723.44
11	0.2625	2925	3058.27
12	0.2875	3229	3404.65
13	0.3125	3572	3763.4
14	0.3375	3703	4135.44
15	0.3625	3749	4521.8
16	0.3875	4860	4923.61
17	0.4125	5083	5342.17
18	0.4375	5403	5778.93
19	0.4625	5687	6235.55
20	0.4875	5927	6713.92
21	0.5125	6372	7216.23
22	0.5375	6813	7744.98
23	0.5625	6852	8303.12
24	0.5875	8157	8894.11
25	0.6125	9810	9522.06
26	0.6375	10306	10191.9
27	0.6625	10506	10909.6
28	0.6875	12770	11682.6
29	0.7125	12785	12520.1
30	0.7375	13780	13433.8
31	0.7625	14122	14439.1
32	0.7875	14263	15556.2
33	0.8125	15675	16813.3
34	0.8375	25858	18250.6
35	0.8625	27059	19928.5
36	0.8875	27090	21944
37	0.9125	28225	24468.2
38	0.9375	30807	27847.7
39	0.9625	31642	32978.4
40	0.9875	34694	44012.9

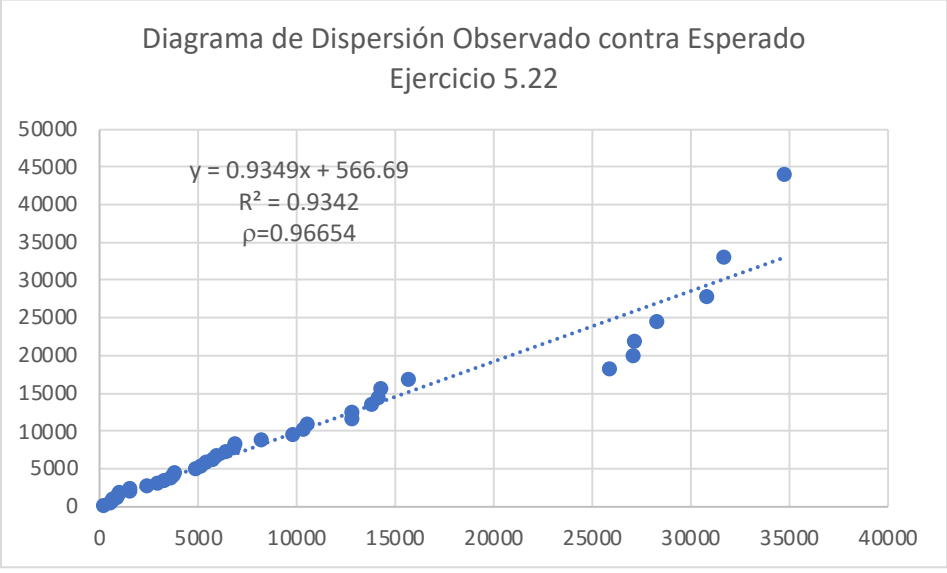
La tercera columna son los valores observados ordenados de menor a mayor. Para obtener la cuarta columna se calcula la inversa de la función de probabilidad exponencial (función de probabilidad supuesta), tomando como estimador del parámetro $\hat{\lambda} = 1/\bar{x}$, el recíproco de la media aritmética de los datos dados, es decir

$$F(x) = 1 - e^{-\lambda x}$$

$$Se\_despeja\_x$$

$$x = \frac{1}{\lambda} Ln\left(\frac{1}{1 - F(x)}\right)$$

$$Valor\_Esperado = \bar{x} * Ln\left(\frac{1}{1 - F(j - 0.5 * n)}\right)$$



Nuevamente, en la figura anterior se observa que existe una inconsistencia en los datos dados, no existen frecuencias observadas en los intervalos 8, 9 y 10 correspondientes al intervalo entre 17500 y 23750 horas. Si las lecturas fueron bien tomadas, pues así deben considerarse; sin embargo, esto puede estar reflejando que se revolvieron datos de una población con datos de otra.

Con respecto al ejercicio 5.23 se conforma la siguiente tabla:

j	$(j\text{-}0.5)/n$	Observado	Esperado
1	0.007813	3	1.78
2	0.023438	4	2.84
3	0.039063	4	3.40
4	0.054688	4	3.80
5	0.070313	5	4.11
6	0.085938	5	4.38
7	0.101563	5	4.61
8	0.117188	5	4.81
9	0.132813	5	5.00
10	0.148438	5	5.17
11	0.164063	5	5.34
12	0.179688	5	5.49
13	0.195313	6	5.63
14	0.210938	6	5.77
15	0.226563	6	5.90
16	0.242188	6	6.02
17	0.257813	6	6.14
18	0.273438	6	6.26
19	0.289063	6	6.38
20	0.304688	6	6.49
21	0.320313	6	6.60
22	0.335938	6	6.70
23	0.351563	6	6.81
24	0.367188	6	6.91
25	0.382813	7	7.01
26	0.398438	7	7.11
27	0.414063	7	7.21
28	0.429688	7	7.31
29	0.445313	7	7.41
30	0.460938	7	7.51
31	0.476563	7	7.60
32	0.492188	7	7.70
33	0.507813	7	7.80

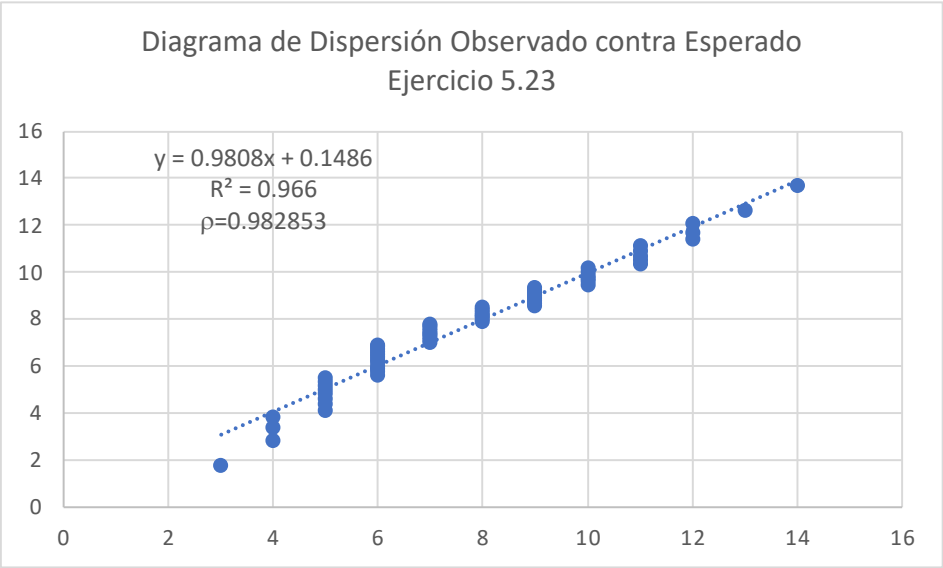
j	$(j\text{-}0.5)/n$	Observado	Esperado
34	0.523438	8	7.90
35	0.539063	8	7.99
36	0.554688	8	8.09
37	0.570313	8	8.19
38	0.585938	8	8.29
39	0.601563	8	8.39
40	0.617188	8	8.49
41	0.632813	9	8.59
42	0.648438	9	8.69
43	0.664063	9	8.80
44	0.679688	9	8.90
45	0.695313	9	9.01
46	0.710938	9	9.12
47	0.726563	9	9.24
48	0.742188	9	9.36
49	0.757813	10	9.48
50	0.773438	10	9.60
51	0.789063	10	9.73
52	0.804688	10	9.87
53	0.820313	10	10.01
54	0.835938	10	10.16
55	0.851563	11	10.33
56	0.867188	11	10.50
57	0.882813	11	10.69
58	0.898438	11	10.89
59	0.914063	11	11.12
60	0.929688	12	11.39
61	0.945313	12	11.70
62	0.960938	12	12.10
63	0.976563	13	12.66
64	0.992188	14	13.72

La tercera columna son los valores observados ordenados de menor a mayor. Para obtener la cuarta columna se calcula la inversa de la función de probabilidad normal (función de probabilidad supuesta), tomando como estimadores

de la media y de la desviación estándar a la media aritmética y a la desviación estándar de los datos dados, es decir

$$Valor\_Esperado = INV.\,NORM\left(j - 0.5 / n, \overline{x}, S_{n-1}\right)$$

Su diagrama de dispersión es



Como se puede apreciar en el diagrama anterior, el ajuste de los datos a una normal es bastante adecuado, su coeficiente de correlación de 0.98 es cercano a uno, por lo cual sí se puede afirmar que los datos tienen distribución normal.

5.4.3. Prueba de bondad de ajuste χ^2 o prueba de Pearson

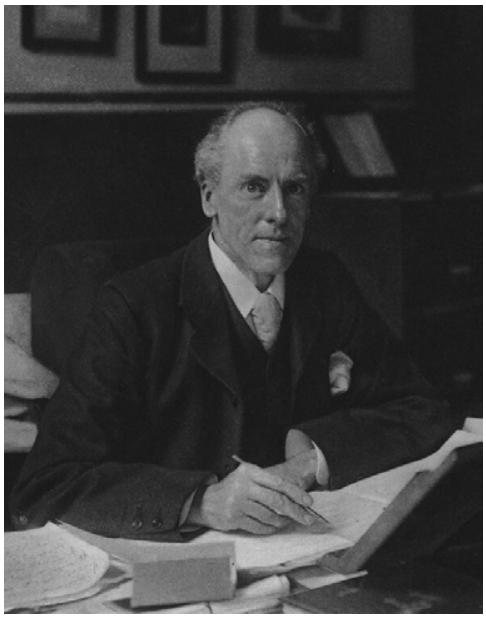


Figura 5.5. Karl Pearson (1857-1936)

Karl Pearson recuperado de: https://es.wikipedia.org/wiki/Karl_Pearson

La prueba χ^2 de Pearson es no paramétrica y mide la discrepancia entre una frecuencia empírica u observada O_i, y otra frecuencia teórica E_i (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis. También se utiliza para probar la independencia de dos variables entre sí, mediante la presentación de los datos en tablas de contingencia.

La hipótesis estadística que se formula es que una frecuencia empírica (observada) O_i se comporta como una frecuencia teórica (esperada) E_i, con cierta distribución de probabilidad conocida.

Pearson propuso el estadístico

$$\chi_0^2 = \sum_{i=1}^{i=n} \left(\frac{E_i - 0_i}{E_i} \right)^2 \tag{5.84}$$

Para el cual, suponiendo que la hipótesis nula H_0 es cierta debe tomar valores pequeños. Si al tomar una muestra su valor es grande, eso pone en evidencia

que la hipótesis inicial es probablemente falsa. Para probar lo anterior, enunció y demostró un teorema al cual denominó Ley asintótica para χ_0^2, el cual demostró que se comporta probabilísticamente como una función de probabilidad ji cuadrada con m-p-h grados de libertad, es decir

$$\chi_0^2 \sim \chi^2_{m-p-h} \tag{5.85}$$

Donde los grados de libertad $k = m - p - h$ dependen de los siguientes factores:

i. El número de intervalos de clase m de la tabla de frecuencias.

ii. El número p de parámetros requeridos para calcular las frecuencias E_i (para el caso de la exponencial negativa se requiere un solo parámetro, por lo que $p = 1$; para el caso de la normal se requieren dos parámetros; la media y la desviación estándar, por lo que $p = 2$; para el caso de la hipergeométrica se requieren tres parámetros D, n y N, por lo que $p = 3$).

iii. El número h de relaciones o condiciones impuestas a las E_i. Por ejemplo, si $\sum_{i=1}^{i=m} E_i = n$, entonces $h = 1$.

Cuanto mayor sea el valor de χ_0^2, menos verosímil es que la hipótesis nula (la cual supone igualdad entre ambas distribuciones) sea correcta. De la misma forma, cuanto más se aproxima a cero el valor de χ_0^2, más ajustadas están ambas distribuciones.

El criterio de decisión es el siguiente, si

$$\chi_0^2 > \chi^2_{\alpha/2,\, m-p-h} \text{ se rechaza la hipótesis nula.} \tag{5.86}$$

Si $\chi_0^2 < \chi^2_{\alpha/2,\, m-p-h}$ entonces no se tiene suficiente evidencia estadística para rechazar H_0 y debe aceptarse aunque no se esté de acuerdo en ello.

Una condición necesaria para poder usar esta prueba de bondad de ajuste es sobre la magnitud de las frecuencias esperadas. No hay un acuerdo entre los expertos, los cuales sugieren que las frecuencias esperadas sean mayores a tres; otros, a cuatro; y otros, a cinco. Esta condición puede cumplirse si se suman algunos intervalos de clase de tal manera que cada frecuencia esperada sea mayor al valor que se fije.

Para ilustrar la aplicación de esta prueba se resolverán los ejercicios 5.21, 5.22 y 5.23.

Con respecto al ejercicio 5.21, se conforma la siguiente tabla

i	O_i	E_i	$(E_i-O_i)^2/E_i$
1	72	100	7.84
2	125	100	6.25
3	108	100	0.64
4	106	100	0.36
5	118	100	3.24
6	71	100	8.41

$$\chi_0^2 = 26.74$$

$$\chi^2_{0.05,5} = 11.07049769$$

$$\chi^2_{0.01,5} = 15.08627247$$

Nótese que todas las frecuencias esperadas son mayores de cinco, por lo cual es perfectamente factible aplicar la prueba de bondad de ajuste ji cuadrada. Se puede apreciar que $m = 6$ intervalos de clase; $p = 0$, por lo que no se requirió estimar ningún parámetro para obtener las frecuencias esperadas; y $h = 1$ porque la única condición es que la suma de las frecuencias esperadas fuera igual a $n = 600$; por lo que se toman $k = m\text{-}p\text{-}h = 5$ grados de libertad y la prueba se realiza a dos niveles de confianza, al 95% y al 99%. En ambos casos el estadístico resultó ser mayor, por lo que se rechaza contundentemente la hipótesis nula de que el lanzamiento del dado presenta distribución uniforme, lo que implica que el dado no es homogéneo.

Con respecto al ejercicio 5.22

No.	Lim Inf Int	Lim Sup Int	Marca Clase	Frec Obs	Frec Esp
1	0	2500	1250	10	8.81387
2	2500	5000	3750	6	6.87176
3	5000	7500	6250	7	5.35759
4	7500	10000	8750	2	4.17706

No.	Lim Inf Int	Lim Sup Int	Marca Clase	Frec Obs	Frec Esp
5	10000	12500	11250	2	3.25666
6	12500	15000	13750	5	2.53907
7	15000	17500	16250	1	1.97959
8	17500	20000	18750	0	1.54339
9	20000	22500	21250	0	1.20331
10	22500	25000	23750	0	0.93817
11	25000	27500	26250	3	0.73144
12	27500	30000	28750	1	0.57027
13	30000	32500	31250	2	0.44462
14	32500	35000	33750	1	0.34665

Nótese que en la tabla de frecuencias anterior existen frecuencias esperadas que no cumplen la condición inicial, por lo que se agrupan los intervalos de clase de la siguiente forma:

i. Se agrupan los intervalos cinco y seis en un solo intervalo, sumando sus frecuencias observadas por un lado $O_5 + O_6 = 2 + 5 = 7$ y sus frecuencias esperadas por otro, $E_5 + E_6 = 3.25666 + 2.53907 = 5.79573$;

ii. Se agrupan los intervalos de clase del 7 al 14 en un solo intervalo, sumando sus frecuencias observadas por un lado $O_7 + O_8 + O_9 + O_{10} + O_{11} + O_{12} + O_{13} + O_{14} = 8$ y sus frecuencias esperadas por otro, $E_7 + E_8 + E_9 + E_{10} + E_{11} + E_{12} + E_{13} + E_{14} = 7.75744$, por lo que queda la siguiente tabla ya resumida:

No.	Lim Inf Int	Lim Sup Int	O_i	E_i	$(E_i - O_i)^2 / E_i$
1	0	2500	10	8.81387	0.159623908
2	2500	5000	6	6.87176	0.110592555
3	5000	7500	7	5.35759	0.503493289
4	7500	10000	2	4.17706	1.134671334
5	10000	15000	7	5.79573	0.250230123
6	15000	35000	8	7.75744	0.007584378

$$\chi_0^2 = 2.166195587$$

$$\chi^2_{0.05,5} = 9.487729037$$

$$\chi^2_{0.01,5} = 13.27670414$$

Nótese que todas las frecuencias esperadas son mayores de cuatro, por lo cual es perfectamente factible aplicar la prueba de bondad de ajuste ji cuadrada. Se puede apreciar que $m = 6$ intervalos de clase, $p = 1$, ya que se requirió estimar el parámetro l para obtener las frecuencias esperadas, y $h = 1$, porque la única condición es que la suma de las frecuencias esperadas fuera igual a $n = 40$, por lo que se toman $k = m–p–h = 4$ grados de libertad y la prueba se realiza a dos niveles de confianza al 95% y al 99%. En ambos casos el estadístico resultó ser menor, por lo que se puede suponer que la distribución de los datos obedece a un modelo exponencial negativo.

Para el ejemplo 5.23, la tabla de frecuencias es

Marca Clase	Frec Acum Obs	Observada	Frec Acum Esp	Esperada
0	0	0	0.000847	0.05
1	0	0	0.003128	0.15
2	0	0	0.009929	0.44
3	1	1	0.027179	1.10
4	4	3	0.064391	2.38
5	12	8	0.132666	4.37
6	24	12	0.239215	6.82
7	33	9	0.380646	9.05
8	40	7	0.540329	10.22
9	48	8	0.693680	9.81
10	54	6	0.818946	8.02
11	59	5	0.905980	5.57
12	62	3	0.957415	3.29
13	63	1	0.983269	1.65
14	64	1	0.994322	0.71
15	64	0	0.998341	0.26
16	64	0	0.999584	0.08
17	64	0	0.999910	0.02
18	64	0	0.999984	0.00
19	64	0	0.999997	0.00
20	64	0	1.000000	0.00
Suma =		64		64.00

Nótese que en la tabla de frecuencias anterior, existen algunas esperadas que no cumplen la condición inicial, por lo que se agrupan los intervalos de clase, sumando sus frecuencias observadas por un lado y las esperadas por otro, por lo que queda la siguiente tabla ya resumida:

No.	Intervalo	O_i	E_i	$(E_i-O_i)^2/E_i$
1	0-4	4	4.12	0.0035537
2	5	8	4.37	3.0162422
3	6	12	6.82	3.9362127
4	7	9	9.05	0.0002942
5	8	7	10.22	1.0143553
6	9	8	9.81	0.3354548
7	10	6	8.02	0.5074647
8	11	5	5.57	0.0583715
9	12-20	5	6.02	0.1719709

$$\chi_0^2 = 2.0879115$$

$$\chi^2_{0.05,6} = 12.591587$$

$$\chi^2_{0.01,5} = 16.811894$$

Nótese que todas las frecuencias esperadas son mayores de cuatro, por lo cual es perfectamente factible aplicar la prueba de bondad de ajuste ji cuadrada. Se puede apreciar que $m = 9$ intervalos de clase, $p = 2$, ya que se requirió estimar la media y la desviación estándar para obtener las frecuencias esperadas, y $h = 1$, porque la única condición es que la suma de las frecuencias esperadas fuera igual a $n = 64$, por lo que se toman $k = m–p–h = 6$ grados de libertad y la prueba se realiza a dos niveles de confianza al 95% y al 99%. En ambos casos el estadístico resultó ser menor, por lo que se puede suponer que la distribución de los datos obedece a una función de probabilidad normal.

5.4.4. Prueba D de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (también conocida como prueba K-S) es una prueba no paramétrica que determina la bondad de ajuste de una distribución de frecuencias empírica (observada) a una distribución de probabilidad teórica (esperada). Lleva ese nombre debido a las contribuciones de dos excelentes matemáticos rusos Andréi Nikolaevich Kolmogórov(1903-1987) y Nikolai Vasilyevich Smirnov (1900-1966).

La prueba de Kolmogorov-Smirnov es más sensible a los valores cercanos a la mediana que a los extremos de la distribución.

Para una demostración de la prueba puede consultar el libro de Jean Dickinson Gibbons y Subhabrata Chakraborti, *Nonparametric Statistical Inference,* CRC Press, Fith Edition, section 4.3, pag 108-125.

La prueba de hipótesis a aplicar establece lo siguiente:

$$H_0 : F\left(x\right) = F_0\left(x\right) \tag{5.87}$$

$$H_1 : F\left(x\right) \neq F_0\left(x\right)$$

Donde $F_0(x)$ es la función de probabilidad acumulada esperada que se supone cumple una población y $S_n(x)$; una distribución de frecuencias observadas a partir de los datos de una muestra aleatoria $\{x_1, x_2, x_3, \cdots, x_n\}$, ordenada de menor a mayor, es decir, x_1 representa al valor mínimo de la muestra y x_n al valor máximo de la muestra. Sea $F_0\left(x\right) = p\left(X \leq x\right)$ la proporción de casos esperados que tienen puntajes menores o iguales que x. Sea $S_n\left(x\right) = k/n$ la función de frecuencia acumulativa observada de la muestra tomada aleatoriamente de n observaciones, donde k es el número de observaciones menor o igual a x. Conforme a la hipótesis nula H_0, se espera que la diferencia entre $S_n(x)$ y $F_0(x)$ sea lo más pequeña posible y que la diferencia entre ellas se deba exclusivamente al error aleatorio. Kolmogorov y Smirnov hicieron el análisis matemático de la distribución de probabilidad que presenta el estadístico D_n de máxima desviación entre $F_0(x)$ y $S_n(x)$, a través del siguiente algoritmo:

$$D_n = max\left\{D_n^+, D_n^-\right\}$$

Donde

$$D_n^+ = max\left[\frac{k}{n} - F_0(x_k)\right] \tag{5.88}$$

$$D_n^- = \max\left\{\max\left[F_0(x_k) - \frac{k-1}{0}\right], 0\right\}$$

La función de probabilidad de $D_{n,\alpha}$ conforme a H_0 fue deducida por Smirnov y se encuentra tabulada en la figura 5.6.

Los pasos para aplicar la prueba D de Kolmogórov-Smirnov son:

1. Se ordenan los datos de la muestra observada ordenados de menor a mayor.

2. Se calcula la frecuencia observada acumulada de cada dato de la muestra, ordenado de menor a mayor, $S_n(x) = k/n$, donde k es el número de observaciones menor o igual a x.

3. Se determina la probabilidad acumulada $F_0(x < x_k)$ usando el modelo teórico que se supone cumple la población a partir de la hipótesis nula.

4. Se calcula D_n usando el siguiente algoritmo

   $D_n = max\left\{D_n^+, D_n^-\right\}$

   Donde

   $$D_n^+ = max\left[\frac{k}{n} - F_0(x_k)\right]$$

   $$D_n^- = \max\left\{\max\left[F_0(x_k) - \frac{k-1}{0}\right], 0\right\}$$

5. Se obtienen $D_{n,\alpha}$ $\alpha = 0.05, 0.01$ usando la tabla mostrada en la figura 5.6.

6. Se comparan los valores anteriores con D_n y se toma la decisión sobre H_0.

$$
\begin{aligned}
&D_n > D_{n,0.05} > D_{n,0.01} && Se\_rechaza\_H_0 \\
&D_n > D_{n,0.05} > D_{n,0.01} && Se\_acepta\_H_0 \\
&D_{n,0.05} > D_n > D_{n,0.01} && Se\_toma\_muestra\_más\_grande
\end{aligned} \tag{5.89}
$$

O también

$$
\begin{aligned}
&p\left(D_n > D_{n,\alpha}\right) < \alpha \quad \alpha=0.05,\ 0.01 && Se\_rechaza\_H_0 \\
&p\left(D_n > D_{n,\alpha}\right) > \alpha \quad \alpha=0.05,\ 0.01 && Se\_acepta\_H_0 \\
&0.05 > p\left(D_n > D_{n,\alpha}\right) > 0.01 && Se\_toma\_muestra\_más\_grande
\end{aligned}
$$

Cabe señalar que cuando las muestras son pequeñas y por consiguiente, los intervalos de clase deben agruparse antes de que χ^2 pueda calcularse apropiadamente, la prueba χ^2 es definitivamente menos poderosa que la prueba de Kolmogórov-Smirnov. Además, para muestras muy pequeñas, la prueba χ^2 no es aplicable en modo alguno, en cambio, la prueba de Kolmogórov-Smirnov sí lo es.

Figura 5.7 Tabla de valores críticos de D en la prueba K-S para bondad de ajuste de una muestra

	Nivel de Significancia para D=máximo $\lvert F_0(x)-S_n(x)\rvert$				
n	0.2	0.15	0.1	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.210	0.220	0.240	0.270	0.320
30	0.190	0.200	0.220	0.240	0.290
35	0.180	0.190	0.210	0.230	0.270
> 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Fuente: Massey, F. J., Jr. 1951. J. Amer. Statistic. Ass., 46, 70.

Para el ejercicio 5.21, la tabla de frecuencias acumuladas se muestra a continuación:

Cara	Frecuencia Observada	Frec Obs Acum	Sn(x)	Frecuencia Esperada	Frec Esp Acum	$F_0(x)$	$Sn(x)-F_0(x)$	$Sn(x-e)-F_0(x)$	$\lvert Sn(x)-F_0(x)\rvert$	$\lvert Sn(x-e)-F_0(x)\rvert$
1	72	72	0.1200	100	100	0.1667	-0.0467	-0.0483333	0.046667	0.048333
2	125	197	0.3283	100	200	0.3333	-0.0050	-0.0066667	0.005000	0.006667
3	108	305	0.5083	100	300	0.5000	0.0083	0.00666667	0.008333	0.006667
4	106	411	0.6850	100	400	0.6667	0.0183	0.01666667	0.018333	0.016667
5	118	529	0.8817	100	500	0.8333	0.0483	0.04666667	0.048333	0.046667
6	71	600	1.0000	100	600	1.0000	0.0000	-0.0016667	0.000000	0.001667
	600			600				$D_n^+ =$	0.048333	
								$D_n^- =$		0.048333
						$D_{600,0.05}=$	0.05552177		Dn=	0.048333
						$D_{600,0.01}=$	0.06654447			

Nótese que en ambos casos $D < D_{n,\alpha}$ al 95% y al 99% de nivel de confianza, por lo que no existe evidencia estadística suficiente para rechazar la hipótesis nula y debe aceptarse que los datos del dado presentan distribución uniforme discreta entre uno y seis. En este ejercicio nótese que todas las frecuencias son mayores de cinco por lo que resulta más poderosa la prueba χ^2 que la *K-S*.

Para el ejercicio 5.22, la tabla de frecuencias acumuladas se muestra a continuación:

No.	tiempo rep	Sn(x)	$F_0(x)$	$Sn(x)-F_0(x)$	$Sn(x-e)-F_0(x)$	$\lvert Sn(x)-F_0(x)\rvert$	$\lvert Sn(x-e)-F_0(x)\rvert$
1	195	0.025	0.019227	0.005773	-0.019227	0.005773	0.019227
2	543	0.05	0.052627	-0.002627	-0.027627	0.002627	0.027627
3	573	0.075	0.055452	0.019548	-0.005452	0.019548	0.005452
4	644	0.1	0.062106	0.037894	0.012894	0.037894	0.012894
5	817	0.125	0.078122	0.046878	0.021878	0.046878	0.021878
6	896	0.15	0.085345	0.064655	0.039655	0.064655	0.039655
7	947	0.175	0.089977	0.085023	0.060023	0.085023	0.060023
8	1482	0.2	0.137182	0.062818	0.037818	0.062818	0.037818
9	1530	0.225	0.141296	0.083704	0.058704	0.083704	0.058704
10	2407	0.25	0.213094	0.036906	0.011906	0.036906	0.011906
11	2925	0.275	0.252649	0.022351	-0.002649	0.022351	0.002649
12	3229	0.3	0.274930	0.025070	0.000070	0.025070	0.000070
13	3572	0.325	0.299273	0.025727	0.000727	0.025727	0.000727
14	3703	0.35	0.308353	0.041647	0.016647	0.041647	0.016647
15	3749	0.375	0.311513	0.063487	0.038487	0.063487	0.038487
16	4860	0.4	0.383609	0.016391	-0.008609	0.016391	0.008609
17	5083	0.425	0.397143	0.027857	0.002857	0.027857	0.002857
18	5403	0.45	0.416048	0.033952	0.008952	0.033952	0.008952
19	5687	0.475	0.432328	0.042672	0.017672	0.042672	0.017672
20	5927	0.5	0.445732	0.054268	0.029268	0.054268	0.029268
21	6372	0.525	0.469753	0.055247	0.030247	0.055247	0.030247
22	6813	0.55	0.492531	0.057469	0.032469	0.057469	0.032469
23	6852	0.575	0.494497	0.080503	0.055503	0.080503	0.055503
24	8157	0.6	0.556089	0.043911	0.018911	0.043911	0.018911
25	9810	0.625	0.623451	0.001549	-0.023451	0.001549	0.023451
26	10306	0.65	0.641595	0.008405	-0.016595	0.008405	0.016595
27	10506	0.675	0.648661	0.026339	0.001339	0.026339	0.001339
28	12770	0.7	0.719565	-0.019565	-0.044565	0.019565	0.044565
29	12785	0.725	0.719983	0.005017	-0.019983	0.005017	0.019983
30	13780	0.75	0.746393	0.003607	-0.021393	0.003607	0.021393
31	14122	0.775	0.754883	0.020117	-0.004883	0.020117	0.004883
32	14263	0.8	0.758300	0.041700	0.016700	0.041700	0.016700
33	15675	0.825	0.789999	0.035001	0.010001	0.035001	0.010001
34	25858	0.85	0.923807	-0.073807	-0.098807	0.073807	0.098807
35	27059	0.875	0.932394	-0.057394	-0.082394	0.057394	0.082394
36	27090	0.9	0.932602	-0.032602	-0.057602	0.032602	0.057602
37	28225	0.925	0.939804	-0.014804	-0.039804	0.014804	0.039804
38	30807	0.95	0.953450	-0.003450	-0.028450	0.003450	0.028450
39	31642	0.975	0.957163	0.017837	-0.007163	0.017837	0.007163
40	34694	1	0.968388	0.031612	0.006612	0.031612	0.006612
					$D_n^+ =$	0.085023	
					$D_n^- =$		0.098807
			$D_{40,0.05}=$	0.215035		Dn=	0.098807
			$D_{40,0.01}=$	0.257726			

Nótese que en ambos casos $D < D_{n,\alpha}$ al 95% y al 99% de nivel de confianza, por lo que no existe evidencia estadística suficiente para rechazar la hipótesis nula y debe aceptarse que los datos del dado presentan distribución exponencial.

Para el ejercicio 5.23, la tabla de frecuencias acumuladas se muestra a continuación:

j	Def	$S_n(x)$	$F_0(x)$	$S_n(x)-F_0(x)$	$S_n(x-e)-F_0(x)$	$\lvert S_n(x)-F_0(x)\rvert$	$\lvert S_n(x-e)-F_0(x)\rvert$
1	3	0.02	0.027179	-0.011554	-0.027179	0.011554	0.027179
2	4	0.03	0.064391	-0.033141	-0.048766	0.033141	0.048766
3	4	0.05	0.064391	-0.017516	-0.033141	0.017516	0.033141
4	4	0.06	0.064391	-0.001891	-0.017516	0.001891	0.017516
5	5	0.08	0.132666	-0.054541	-0.070166	0.054541	0.070166
6	5	0.09	0.132666	-0.038916	-0.054541	0.038916	0.054541
7	5	0.11	0.132666	-0.023291	-0.038916	0.023291	0.038916
8	5	0.13	0.132666	-0.007666	-0.023291	0.007666	0.023291
9	5	0.14	0.132666	0.007959	-0.007666	0.007959	0.007666
10	5	0.16	0.132666	0.023584	0.007959	0.023584	0.007959
11	5	0.17	0.132666	0.039209	0.023584	0.039209	0.023584
12	5	0.19	0.132666	0.054834	0.039209	0.054834	0.039209
13	6	0.20	0.239215	-0.036090	-0.051715	0.036090	0.051715
14	6	0.22	0.239215	-0.020465	-0.036090	0.020465	0.036090
15	6	0.23	0.239215	-0.004840	-0.020465	0.004840	0.020465
16	6	0.25	0.239215	0.010785	-0.004840	0.010785	0.004840
17	6	0.27	0.239215	0.026410	0.010785	0.026410	0.010785
18	6	0.28	0.239215	0.042035	0.026410	0.042035	0.026410
19	6	0.30	0.239215	0.057660	0.042035	0.057660	0.042035
20	6	0.31	0.239215	0.073285	0.057660	0.073285	0.057660
21	6	0.33	0.239215	0.088910	0.073285	0.088910	0.073285
22	6	0.34	0.239215	0.104535	0.088910	0.104535	0.088910
23	6	0.36	0.239215	0.120160	0.104535	0.120160	0.104535
24	6	0.38	0.239215	0.135785	0.120160	0.135785	0.120160
25	7	0.39	0.380646	0.009979	-0.005646	0.009979	0.005646
26	7	0.41	0.380646	0.025604	0.009979	0.025604	0.009979
27	7	0.42	0.380646	0.041229	0.025604	0.041229	0.025604
28	7	0.44	0.380646	0.056854	0.041229	0.056854	0.041229
29	7	0.45	0.380646	0.072479	0.056854	0.072479	0.056854
30	7	0.47	0.380646	0.088104	0.072479	0.088104	0.072479
31	7	0.48	0.380646	0.103729	0.088104	0.103729	0.088104

j	Def	$S_n(x)$	$F_0(x)$	$S_n(x)-F_0(x)$	$S_n(x-e)-F_0(x)$	$\lvert S_n(x)-F_0(x)\rvert$	$\lvert S_n(x-e)-F_0(x)\rvert$
32	7	0.50	0.380646	0.119354	0.103729	0.119354	0.103729
33	7	0.52	0.380646	0.134979	0.119354	0.134979	0.119354
34	8	0.53	0.540329	-0.009079	-0.024704	0.009079	0.024704
35	8	0.55	0.540329	0.006546	-0.009079	0.006546	0.009079
36	8	0.56	0.540329	0.022171	0.006546	0.022171	0.006546
37	8	0.58	0.540329	0.037796	0.022171	0.037796	0.022171
38	8	0.59	0.540329	0.053421	0.037796	0.053421	0.037796
39	8	0.61	0.540329	0.069046	0.053421	0.069046	0.053421
40	8	0.63	0.540329	0.084671	0.069046	0.084671	0.069046
41	9	0.64	0.693680	-0.053055	-0.068680	0.053055	0.068680
42	9	0.66	0.693680	-0.037430	-0.053055	0.037430	0.053055
43	9	0.67	0.693680	-0.021805	-0.037430	0.021805	0.037430
44	9	0.69	0.693680	-0.006180	-0.021805	0.006180	0.021805
45	9	0.70	0.693680	0.009445	-0.006180	0.009445	0.006180
46	9	0.72	0.693680	0.025070	0.009445	0.025070	0.009445
47	9	0.73	0.693680	0.040695	0.025070	0.040695	0.025070
48	9	0.75	0.693680	0.056320	0.040695	0.056320	0.040695
49	10	0.77	0.818946	-0.053321	-0.068946	0.053321	0.068946
50	10	0.78	0.818946	-0.037696	-0.053321	0.037696	0.053321
51	10	0.80	0.818946	-0.022071	-0.037696	0.022071	0.037696
52	10	0.81	0.818946	-0.006446	-0.022071	0.006446	0.022071
53	10	0.83	0.818946	0.009179	-0.006446	0.009179	0.006446
54	10	0.84	0.818946	0.024804	0.009179	0.024804	0.009179
55	11	0.86	0.905980	-0.046605	-0.062230	0.046605	0.062230
56	11	0.88	0.905980	-0.030980	-0.046605	0.030980	0.046605
57	11	0.89	0.905980	-0.015355	-0.030980	0.015355	0.030980
58	11	0.91	0.905980	0.000270	-0.015355	0.000270	0.015355
59	11	0.92	0.905980	0.015895	0.000270	0.015895	0.000270
60	12	0.94	0.957415	-0.019915	-0.035540	0.019915	0.035540
61	12	0.95	0.957415	-0.004290	-0.019915	0.004290	0.019915
62	12	0.97	0.957415	0.011335	-0.004290	0.011335	0.004290
63	13	0.98	0.983269	0.001106	-0.014519	0.001106	0.014519
64	14	1.00	0.994322	0.005678	-0.009947	0.005678	0.009947

$D_n^+ = 0.135785$

$D_n^- = 0.120160$

$D_{40,0.05} = 0.170000$ $\quad D_n = 0.135785$

$D_{40,0.01} = 0.203750$

Nótese que en ambos casos $D < D_{n,\alpha}$ al 95% y al 99% de nivel de confianza, por lo que no existe evidencia estadística suficiente para rechazar la hipótesis nula y debe aceptarse que los datos del número de defectuosos en la muestra presentan distribución normal.

Lo valioso de los tres métodos anteriores que se han visto para probar la bondad de ajuste entre una distribución empírica de datos y un modelo probabilístico particular es que los tres, aunque antiguos, son aplicables a cualquier modelo probabilístico; por ejemplo, el ejercicio 5.21 trató el caso de una función de probabilidad uniforme discreta, el ejercicio 5.22 abordó el caso de una distribución exponencial negativa que es un modelo continuo, y el ejercicio 5.23 trató el caso de un modelo normal.

En el subtema que sigue se abordarán pruebas de bondad de ajuste exclusivamente para la distribución normal.

5.4.5. Pruebas de hipótesis de normalidad de un conjunto de datos o pruebas de bondad de ajuste a una curva normal

Una parte considerable de las aplicaciones que se han desarrollado de la probabilidad y la estadística, se basa en el supuesto de normalidad de los datos. Por ello, numerosos investigadores se han dedicado a analizar métodos específicos para probar la hipótesis de normalidad de un conjunto de datos.

Para probar la bondad de ajuste de un conjunto de datos a un modelo de probabilidad normal se citan una gran cantidad de pruebas:

1. Prueba de Anderson-Darling (A-D).
2. Prueba de Cramer-von Mises (CVM).
3. Prueba de Lilliefors (Kolmogorov-Smirnov).
4. Prueba $\chi 2$ de Pearson.
5. Prueba de Shapiro-Wilk (S-W) o su similar Ryan-Joyner (R-J).
6. Prueba de Shapiro-Francia (S-F).
7. Prueba de Jarque-Bera (J-B).
8. Prueba de Frosini.
9. Prueba de Geary.
10. Prueba de Hegazy-Green.
11. Prueba de Curtosis.
12. Prueba de Asimetría.
13. Prueba de Spiegelhalter.
14. Prueba de Weisberg-Bingham (W-B).
15. Prueba de Agostino-Shapiro (A-S).

Prueba de Normalidad de Anderson Darling (A-D)

Es una prueba no paramétrica sobre si los datos de una muestra provienen de una distribución normal. El estadístico de prueba que se usa tiene la siguiente expresión:

$$A = -n - \frac{1}{n}\sum_{i=1}^{i=n}\left[2i-1\right]\left[1n\left(p\left(i\right)\right) + 1n\left(\mathrm{i}-p(n-i+1)\right)\right] \qquad (5.90)$$

Donde

$$p(i) = \Phi\left(\left[\frac{x(i)-x}{S_{n-1}}\right]\right) \Phi(x) \sim N\left(\hat{\mu}_x = \overline{x}, \hat{\sigma}_x = S_{n-1}\right)$$

Donde el valor de p (p-value) se obtiene al calcular la probabilidad

$$p = p\left(z \leq A\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)\right) \tag{5.91}$$

A continuación, se usará Minitab y R para resolver el ejercicio 5.23, usando la prueba de Anderson-Darling.

Con Minitab

1. Se capturan los datos dados en la columna C1 de Minitab o se migran directamente de Excel.

2. Se da click en el menú Stat, luego en el submenú Basic Statistics, después en Normality Test, como se muestra en la figura 5.8.

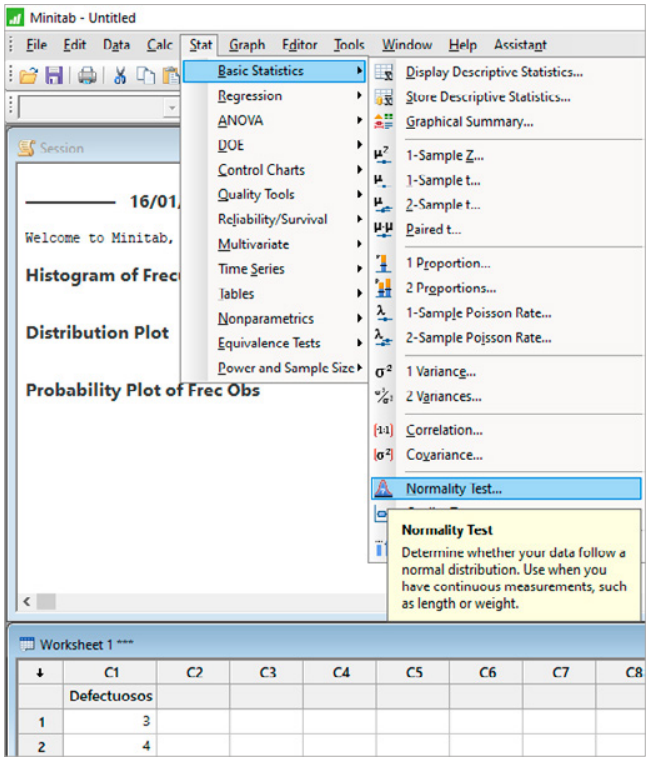
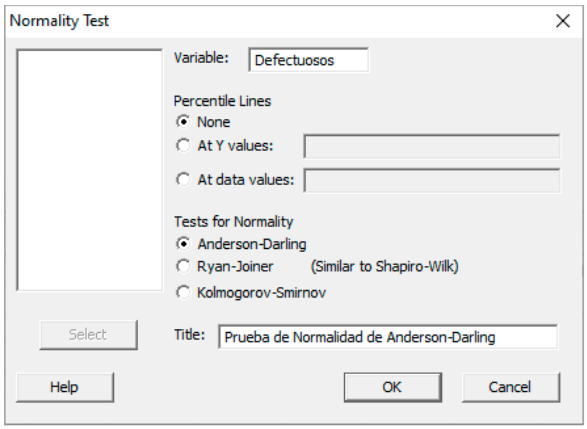


Figura 5.8.

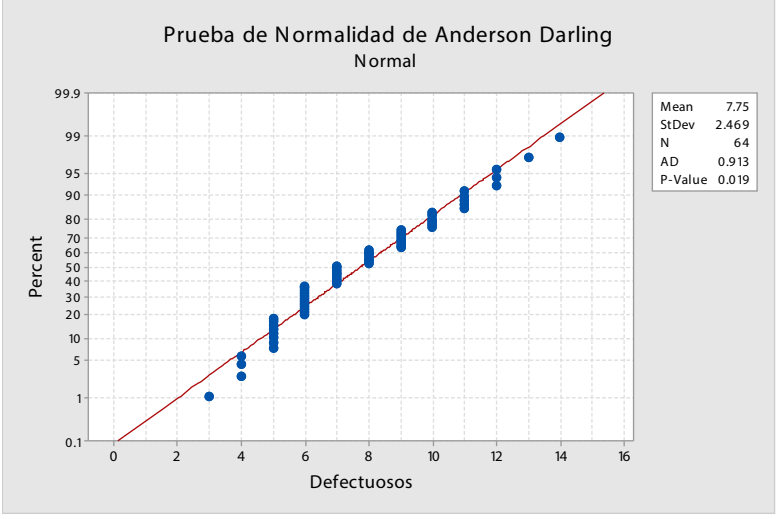
3. Aparece la siguiente pantalla, en la cual se indica la variable a analizar, se selecciona Anderson-Darling y se coloca el título de la prueba.

Figura 5.9.



4. Se obtiene la respuesta gráficamente, la cual se muestra en la figura 5.10.

Figura 5.10.



De acuerdo con esta prueba A-D, el valor de p (*p-value* en la figura 5.10) es menor a 0.05 pero no es menor para 0.01, por lo que para 0.05 se debe rechazar H_0, pero para 0.01 no se puede rechazar la hipótesis nula de que los datos dados presentan una distribución normal. Dado que se cae en zona de duda es conveniente tomar una muestra mayor para que el resultado sea más verosímil.

Con R

Se teclean las siguientes líneas:

```
> library(nortest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> ad.test(def);

        Anderson-Darling normality test

data:  def
A=0.91347, p-value=0.01889
```

De acuerdo con esta prueba A-D en R, el valor de p es menor a 0.05 pero no es menor para 0.01, por lo que para 0.05 se debe rechazar H_0, pero para 0.01 no se puede rechazar la hipótesis nula de que los datos dados presentan una distribución normal. Dado que se cae en zona de duda es conveniente tomar una muestra mayor para que el resultado sea más verosímil.

Prueba de Normalidad de Cramér-von Mises (CVM)

El criterio lleva los apellidos de Harald Cramér (1893-1985), matemático sueco que fue profesor y rector de la Universidad de Estocolmo, y Richard Edler von Mises (1883-1953), físico austrohúngaro y profesor de las universidades de Berlín y Harvard, ambos fueron los primeros en exponerlo entre 1928 y1930. Es una prueba muy útil para pequeñas muestras y usa los momentos como criterio.

El estadístico de prueba que se usa en CVM tiene la siguiente expresión:

$$w = \frac{1}{12n} + \sum_{i=1}^{i=n} \left[p(i) - \frac{2i-1}{2n} \right]^2$$

Donde (5.92)

$$p(i) = \Phi\left(\left[\frac{x(i) - \bar{x}}{S_{n-1}} \right] \right) \Phi(x) \sim N\left(\hat{\mu}_x = \bar{x}, \hat{\sigma}_x = S_{n-1} \right)$$

Donde el valor de p(p-*value*) se obtiene al calcular la probabilidad

$$p = p\left(z \leq w\left(1 + \frac{0.5}{n}\right)\right) \tag{5.93}$$

A continuación, se usará R para resolver el ejercicio 5.23, usando la prueba de Cramér-Von Mises (Minitab no la tiene habilitada).

```
> library(nortest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> cvm.test(def);

    Cramer-von Mises normality test

data:  def
W=0.1602, p-value=0.01686
```

De acuerdo con esta prueba CVM, el valor de p es menor a 0.05 pero no es menor para 0.01, por lo que para 0.05 se debe rechazar H_0, pero para 0.01 no se puede rechazar la hipótesis nula de que los datos dados presentan una distribución normal. Dado que se cae en zona de duda es conveniente tomar una muestra mayor para que el resultado sea más verosímil.

Prueba de Normalidad de Lilliefors

La prueba de Lilliefors debe su nombre a Hubert Lilliefors Whitman (1928-2008) un estadístico estadounidense, profesor en la Universidad George Washington de EUA. Esta prueba se basa en la prueba D de Kolmogórov-Smirnov aplicada para probar la normalidad de un conjunto de datos de una muestra.

La prueba de Lilliefors procede de la siguiente manera:

i. Estimar la media de la población y la varianza de la población con base en los datos empíricos dados.

ii. Obtener la Dn máxima entre la función de distribución empírica $S_n(x)$ y la función de distribución acumulativa $F_0(x)$ de la distribución normal con la media estimada y la varianza estimada. Al igual que en la prueba de Kolmogórov-Smirnov, esta será la estadística de prueba.

iii. Evaluar si la discrepancia máxima es lo suficientemente grande como para ser estadísticamente significativa, lo que requiere el rechazo de la hipótesis nula. El profesor Lilliefors fue el primero en calcular las tablas para esta distribución, usando una computadora a través de la generación de números aleatorios usando los métodos Monte Carlo.

A continuación, se usará Minitab y R para resolver el ejercicio 5.23, usando la prueba de Lilliefors.

Con Minitab, se sigue el mismo camino que se aplicó para la prueba de Anderson-Darling, utilizando la prueba señalada como K-S, obteniéndose la gráfica de la figura 5.11.

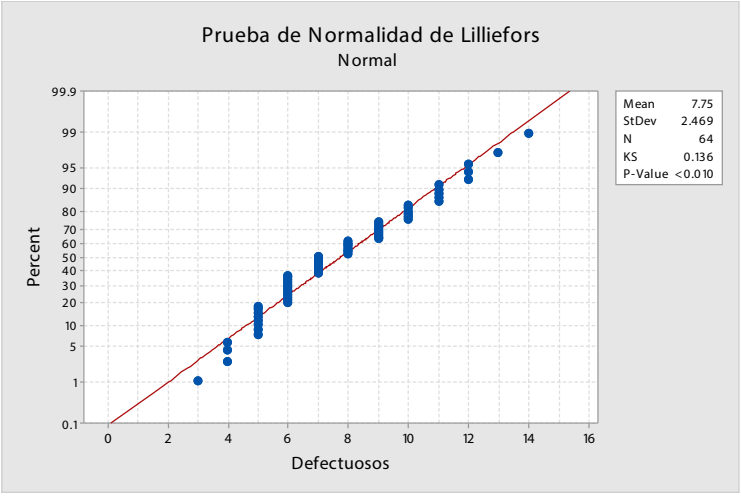


Figura 5.11.

De acuerdo con el valor de *p* (*p-value* en la figura 5.11) se debe rechazar la hipótesis nula de que los datos dados presentan una distribución normal.

Con R

Se teclean las siguientes líneas (de una vez se realizan todas las pruebas de normalidad que trae incluidas R:

```
> library(nortest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> lillie.test(def);

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  def
D = 0.13579, p-value = 0.005072
```

De acuerdo con esta prueba de Lilliefors, el valor de p es menor a 0.05 y menor a 0.01, por lo que se debe rechazar H_0, es decir los datos dados no presentan una distribución normal.

Prueba de Normalidad χ^2 de Pearson

Esta prueba ya fue analizada en el subtema 5.4.3, por lo cual sólo se resolverá el ejercicio 5.23 utilizando R. Está basada en una distribución ji cuadrada y corresponde con una prueba de bondad de ajuste.

```
> library(nortest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);

> pearson.test(def);

    Pearson chi-square normality test

data:  def
P = 32.938, p-value = 6.321e-05
```

De acuerdo con esta prueba χ^2, el valor de p es menor a 0.05 y menor a 0.01, por lo que se debe rechazar contundentemente H_0, es decir, los datos dados no presentan una distribución normal.

Prueba de Shapiro-Wilk (o su similar Ryan-Joyner)

Esta prueba fue publicada en 1965 por Samuel Shapiro (1930 -), profesor emérito de estadística en la Universidad Internacional de Florida y Martin Bradbury Wilk (1922 – 2013), estadístico canadiense. Es más poderosa al compararse con otras pruebas de normalidad cuando la muestra es pequeña.

La hipótesis nula es que un conjunto de datos de una muestra aleatoria proviene de una población normalmente distribuida con media y varianza desconocidas.

El estadístico de prueba w se define a través de la siguiente expresión:

$$w = \frac{\left(\sum_{i=1}^{i=n} \alpha(i)\, x(i)\right)^2}{\sum_{i=1}^{i=n} \left(x(i) - \bar{x}\right)^2} \tag{5.94}$$

Las variables ai son las componentes de un vector con la siguiente expresión vectorial

$$\vec{a} = \left(a_1, a_2, \cdots, a_n\right) = \frac{\vec{m}\,\vec{V}^{-1}}{\sqrt{\vec{m}^{\,T}\vec{V}^{-1}\vec{m}}} \tag{5.95}$$

Donde

$$\vec{m} = \left(m_1, m_2, \cdots, m_n\right)$$

$$\vec{V} = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,n}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 & \cdots & \sigma_{2,n}^2 \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n,1}^2 & \sigma_{n,2}^2 & \cdots & \sigma_{n,n}^2 \end{bmatrix} \tag{5.96}$$

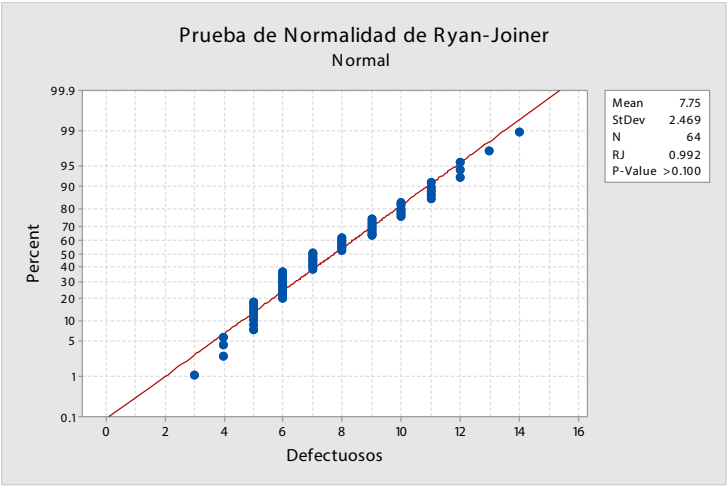
Cada mi del vector $\vec{m}$ representa el valor medio del estadístico ordenado, de variables aleatorias estadísticamente independientes e idénticamente distribuidas normales y $\vec{V}$ es la matriz de covarianzas de ese estadístico de orden.

Si el valor de p (p-*value*) es menor a a entonces H_0 es rechazada, es decir, se concluye que los datos no provienen de una distribución normal. Si el valor de p (p-*value*) es mayor a a, se concluye que no se puede rechazar H_0.

La prueba de Ryan-Joiner opera de forma similar a la de Shapiro-Wilk.

A continuación, se emplea el método de Ryan-Joiner cuyo algoritmo está habilitado por el software Minitab, obteniéndose la gráfica de la figura 5.12.

Figura 5.12



De acuerdo con esta prueba R-J, el valor de p(p-*value* en la figura 5.12) es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar la hipótesis nula de que los datos dados presentan una distribución normal.

Con R

```
> library(nortest);
> def < -c(5,7,8,5,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,11,6,
10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> shapiro.test(def);

    Shapiro-Wilk normality test

data:  def
W = 0.96398, p-value = 0.05867
```

De acuerdo con esta prueba S-W, el valor de p (*p-value*) es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar la hipótesis nula de que los datos dados presentan una distribución normal.

Prueba de Normalidad de Shapiro-Francia

Esta prueba fue creada por S.S. Shapiro y R. S. Francia en 1972 como una forma de simplificar la prueba de Shapiro-Wilk.

El estadístico de prueba que usa tiene la siguiente expresión:

$$w' = \frac{cov(x, m)}{\sigma_x \ \sigma_m} = \frac{\sum_{i=1}^{i=n} \left(x(i) - \bar{x}\right)\left(m_i - \overline{m}\right)}{\sqrt{\left(\sum_{i=1}^{i=n} \left(x(i) - \bar{x}\right)^2\right)\left(\sum_{i=1}^{i=n} \left(n_i - \overline{m}\right)^2\right)}} \qquad (5.97)$$

Bajo la hipótesis nula los datos presentan una distribución normal.

Esta prueba comparada con la de Shapiro-Wilk es más fácil de aplicar computacionalmente porque no requiere invertir la matriz de covarianzas.

A continuación, se usará R para resolver el ejercicio 5.23, usando la prueba de Shapiro-Francia (Minitab no la tiene integrada).

```
> library(nortest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> sf.test(def);

    Shapiro-Francia normality test

data:  def
W=0.96687, p-value=0.07727
```

De acuerdo con esta prueba S-F, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Normalidad de Jarque-Bera

Esta prueba de normalidad fue desarrollada por el economista mexicano Carlos M. Jarque Uribe (1954 -) y el economista estadounidense de origen hindú Anil K. Bera (1955 -), profesor de la Universidad de Illinois en Urbana-Champaign.

El estadístico de prueba definido por ellos se expresa a continuación:

$$JB = \frac{n}{6}\left(\hat{\gamma}_1^2 + \frac{1}{4}\hat{\gamma}_2^2\right) \tag{5.98}$$

Donde $\hat{\gamma}_1$ representa al estimador del coeficiente de asimetría y $\hat{\gamma}_2$ al estimador del coeficiente de curtosis, definidos por las siguientes expresiones

$$\hat{\gamma}_1 = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}$$

$$\hat{\gamma}_1 = \frac{\hat{\mu}_4}{\hat{\mu}_2^{2}} = \frac{\frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2\right)^{2}} \tag{5.99}$$

Jarque y Bera dedujeron que si los datos provienen de una distribución normal, el estadístico *JB* se comporta como una distribución χ^2 con dos grados de libertad.

La hipótesis nula considera conjuntamente las hipótesis de que el coeficiente de asimetría y el coeficiente de curtosis son cero, como sucede en el caso de la normal.

Para pequeñas muestras la aproximación del estadístico *JB* a la distribución χ^2 es sumamente sensible, por lo cual, el algoritmo de Jarque-Bera está considerado en Matlab para tamaños de muestra grandes ($n > 2000$). Para pequeñas

muestras se usan tablas obtenidas por simulación para interpolar el valor de *p* (*p-value*).

Con R, en este caso se tiene que cargar otro paquete llamado normtest, por lo cual las instrucciones a teclear son las siguientes:

```
> library(normtest);
> def < -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> jb.norm.test(def);

    Jarque-Bera test for normality

data:  def
JB = 2.4694, p-value = 0.1525
```

De acuerdo con esta prueba J-B, el valor de *p* es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Frosini

Con R

```
>library(normtest);
> def < -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> frosini.norm.test(def);

    Frosini test for normality

data:  def
B = 0.30944, p-value = 0.0205
```

De acuerdo con esta prueba, el valor de *p* es menor a 0.05 y mayor a 0.01, por lo que se recomendaría aumentar el tamaño de muestra antes de tomar la decisión de establecer si los datos dados presentan una distribución normal.

Prueba de Geary

Con R

```
>library(normtest);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> geary.norm.test(def);

    Geary test for normality

data:  def
d = 0.83882, p-value = 0.063
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Hegazy-Green

Con R

```
>library(normtest);
>def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,11,6,10,
7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> hegazy1.norm.test(def);

    Hegazy-Green test for normality

data:  def
T = 0.14947, p-value = 0.0335
```

De acuerdo con esta prueba, el valor de p es menor a 0.05 y mayor a 0.01, por lo que se recomendaría aumentar el tamaño de muestra antes de tomar la decisión de establecer si los datos dados presentan una distribución normal.

Prueba de Curtosis

Con R

```
library(normtest);
> def < -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,11,6,
10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> kurtosis.norm.test(def);

    Kurtosis test for normality

data:  def
T = 2.4613, p-value = 0.3065
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Asimetría

Con R

```
> library(normtest);
> def < -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,11,6,10,
7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> skewness.norm.test(def);

    Skewness test for normality

data:  def
T = 0.39868, p-value = 0.1545
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Spiegelhalter

Con R

```
library(normtest);
> def< -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,11,6,10,
7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> spiegelhalter.norm.test(def);

    Spiegelhalter test for normality

data:  def
T = 1.1923, p-value = 0.9635
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Weisberg-Bingham

Con R

```
library(normtest);
> def< -c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,
6,11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> wb.norm.test(def);

    Weisberg-Bingham test for normality

data:  def
WB = 0.96687, p-value = 0.074
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

Prueba de Normalidad D'Agostino-Pearson (D-P)

Considere una muestra de n observaciones, las cuales se ordenan de menor a mayor, asignándoles un consecutivo en función de este orden. Se calculan, la media y la desviación estándar de la media, a partir de las cuales se obtiene el estadístico T

$$T = \sum_{i=1}^{i=n} \left(i - \frac{n+1}{2} \right) x_i = \sum_{i=1}^{i=n} i x_i - \frac{n(n+1)}{2} \bar{x}$$

$$D = \frac{T}{n^2 S_{n-1}}$$

D'Agostino tabuló una tabla de valores de $D_{n,\alpha}$ para los cuales si $D < D_{n,\alpha}$ o $D > D_{n,\alpha}$, se rechaza la hipótesis nula de normalidad; por de lo contrario, si $D_{n,\alpha} < D < D_{n,\alpha}$, se asume la hipótesis de normalidad de los datos. Para realizar esta prueba se establece que $n > 10$.

Con R

```
> library(moments);
> def<-c(5,7,8,5,7,4,12,8,10,9,3,14,7,8,7,6,6,6,12,7,10,7,6,6,5,8,5,4,10,6,
11,6,10,7,9,4,8,10,11,6,9,9,6,7,11,8,9,7,6,12,11,5,9,5,13,5,11,8,9,6,6,9,5,10);
> agostino.test(def);

        D'Agostino skewness test

data:  def
skew = 0.39868, z = 1.38630, p-value = 0.1657
alternative hypothesis: data have a skewness
```

De acuerdo con esta prueba, el valor de p es mayor a 0.05 y mayor a 0.01, por lo que no se puede rechazar H_0, es decir, los datos dados presentan una distribución normal.

En todas las pruebas de normalidad citadas anteriormente, el valor de p (*p-value*) muestra la probabilidad de haber obtenido el resultado suponiendo que la hipótesis nula H_0 es cierta. Valores de p por arriba de α no permiten rechazar

la H_0, mientras que valores de p por debajo de α sí permiten rechazar la H_0. Cuando el valor de p es inferior al nivel de significación a, lo más probable es que la hipótesis de partida sea falsa, aunque también es posible que se esté en presencia de una observación atípica. En este caso, se estaría cometiendo el error estadístico tipo I, es decir, rechazar la hipótesis nula cuando esta es cierta.

Es importante recalcar que una prueba de hipótesis no permite aceptar una hipótesis; simplemente la rechaza o no la rechaza, es decir que la tacha de verosímil (lo que no significa obligatoriamente que sea cierta, simplemente que es más probable de serlo) o inverosímil.

Se han efectuado análisis comparativos tratando de demostrar cuál de todas estas pruebas de normalidad es la mejor y no existen resultados concluyentes de carácter general. Cuando la muestra es pequeña, las pruebas de normalidad estudiadas tienen un poder 1–β muy bajo. Todas ellas son sensibles ante una correlación fuerte de los datos. Las tres primeras pruebas tienden a ser las más adecuadas para identificar una distribución no normal cuando la distribución es asimétrica. Por lo general, entre las pruebas que se basan en la función de distribución empírica, la prueba de Anderson-Darling tiende a ser más efectiva para detectar desviaciones en las colas de la distribución.

Ejercicios del Capítulo 5

1. Se recibe un lote de $N = 9000$ resortes de tracción utilizados para interruptores de flotador, los cuales deben presentar una carga especificada de 6.5 ± 0.25 libras. Para proceder a aceptar el lote se realiza un plan de muestreo aleatorio simple, con un tamaño de muestra de $n = 90$ resortes, los cuales son probados con un medidor de cargas, arrojando los siguientes resultados.

6.87	6.82	6.76	6.79	6.52	6.90	6.70	7.07	6.63
6.82	6.51	6.45	6.59	6.71	6.65	6.88	6.81	6.52
6.74	6.78	6.70	6.82	6.85	6.80	6.79	6.82	6.58
7.12	6.74	6.80	6.60	6.82	7.18	6.89	6.60	6.77
6.82	6.67	7.06	6.69	6.42	6.51	6.72	6.96	6.96
6.60	7.16	6.99	6.78	6.86	7.00	6.64	6.78	6.57
6.53	6.84	6.58	6.63	6.72	6.47	6.75	6.69	6.48
6.99	6.48	6.90	6.67	6.77	6.54	6.82	6.75	6.63
6.68	6.70	6.47	6.98	6.69	6.94	6.71	6.49	6.94
6.73	6.71	6.87	6.37	7.05	6.79	6.97	6.72	6.88

   a. Pruebe la hipótesis de que los datos presentan distribución normal, utilizando para ello todos los métodos vistos en este tema.
   b. Pruebe la hipótesis de que la media de carga de los resortes es mayor a 7.0.
   c. Pruebe la hipótesis de que la varianza en la carga del lote de resortes es menor a 0.025.
   d. Pruebe la hipótesis de que la fracción de resortes en el lote que presenta una carga por arriba de siete libras es mayor a 10%.

2. Una oficina expendedora de pasaportes en la Ciudad de México establece un control de los errores que cometen al tramitar 500 pasaportes diariamente, los cuales operan por cita. Para analizar su desempeño decide recopilar una

muestra de $n = 20$ días. Cabe señalar que los errores que cometen son muy diversos, foto borrosa, foto equivocada, errores de tipografía, fecha de nacimiento equivocada, fecha de emisión equivocada, nombre mal escrito, tiempo de espera excedido, tiempo de emisión excedido, poca cortesía al recibir a los ciudadanos, trato descortés, personas que se cuelan a la fila, etcétera.

La oficina diariamente levanta una encuesta y le solicita a cada cliente le indique los errores cometidos, los cuales pueden ser críticos, mayores, menores, etcétera; nótese que el número de errores puede ser mayor que el número de pasaportes emitidos, ya que cada pasaporte puede presentar desde cero hasta una cantidad incontable de errores. Suponga que los errores o defectos durante esos $n = 20$ días fueron los siguientes:149, 150, 154, 139, 145, 152, 142, 142, 154, 138, 148, 130, 149, 143, 151, 131, 130, 115, 159, 149.

a. Pruebe la hipótesis de que los datos presentan distribución de Poisson utilizando la prueba de bondad de ajuste ji cuadrada y la prueba D de Kolmogorov-Smirnoff.
b. Pruebe la hipótesis de que los datos presentan distribución normal utilizando todas las pruebas vistas en este capítulo.
c. Pruebe la hipótesis estadística que el número de defectos por día presenta distribución normal utilizando todas las pruebas vistas en este capítulo.
d. Pruebe la hipótesis estadística que el número de defectos por día es en promedio de uno por cada cuatro pasaportes.

3. En un puerto marítimo el tiempo de descarga en horas de un barco carguero es crítico. Existe una cláusula en el contrato de la compañía naviera con el puerto, de no tardarse más de 20 horas en la descarga. Por cada hora de retraso, un barco le cobrará al puerto $50,000.00. El puerto decide realizar un análisis del tiempo que le lleva descargar cada barco y recolecta una muestra aleatoria de $n = 60$ barcos a lo largo de un mes. Los resultados en horas se muestran a continuación.

6.2	2.5	13.8	2.3	20.6	0
8	10.7	9.7	16.1	4.8	4.1
4.8	4.4	14	1.1	1	32.1
15.1	0.6	4	25	4.4	7.7

15.6	7.9	10.7	29.4	11.7	0.9
4.3	14.5	1.7	6.5	6.1	7.3
1.7	39	2.8	14.9	11.8	2.2
25.7	5.9	4.7	3.9	3.2	2.5
12.6	8.2	0.9	6.5	0.5	3.8
12.6	12.5	7.1	0.4	3.3	23.3

a. Pruebe la hipótesis de que la distribución de los datos es normal.
b. Pruebe la hipótesis de que la media del tiempo de descarga es de seis horas.
c. Pruebe la hipótesis de que la varianza de los datos es menor a 25.
d. Revise las hipótesis básicas para las pruebas de hipótesis de los incisos anteriores, ¿considera usted que se cumplen?

4. La durabilidad de una llanta se mide con base en el número promedio de kilómetros recorridos bajo condiciones controladas, hasta que aparezcan bandas de desgaste con una profundidad de rodadura de 1.6 mm de espesor o menos. Los compuestos de una llanta están diseñados para funcionar idealmente por un máximo de 5 años. El siguiente ejemplo es hipotético y se utilizan datos generados aleatoriamente que no son reales de las marcas que se mencionan, pero es ilustrativo del empleo que debe darse a los intervalos de confianza para dos poblaciones. Suponga que se desea adquirir un lote de N = 1600 llantas 245/40r20 Run Flat 99y. Para ello, existen dos marcas diferentes Pirelli y Michelin. Se realiza una prueba de durabilidad, usando ocho coches, Mercedes Benz C200, del mismo modelo y año, con las mismas condiciones de uso, con la misma carga, recién alineados y balanceados bajo las mismas especificaciones. Al primer equipo A de cuatro coches se les montan cuatro llantas Pirelli y al otro equipo B de cuatro coches, llantas Michelin. Se ponen a rodar bajo las mismas condiciones hasta que aparecen bandas de desgaste con una profundidad de rodadura de 1.6 mm de espesor o menos. Los resultados se muestran a continuación.

a. Pruebe la hipótesis de normalidad de los datos para cada una de las marcas utilizando todas las técnicas vistas en este capítulo.
b. Pruebe la hipótesis de igualdad entre varianzas para ambas marcas.
c. Pruebe la hipótesis de igualdad entre medias para ambas marcas.
d. Pruebe la hipótesis de igualdad entre fracciones de llantas que duran más de 52,000 Km.

Michelin	Pirelli	Michelin	Pirelli
50,133.93	60,014.73	52,109.23	53,801.49
49,954.33	54,336.33	48,483.41	57,228.75
49,164.83	54,940.62	46,079.35	54,037.71
51,300.54	52,895.42	47,043.97	55,067.46
48,249.71	57,123.37	44,855.49	55,990.88
48,494.13	57,535.76	52,853.32	54,487.21
51,038.79	51,359.57	49,679.16	52,606.81
53,306.49	58,214.95	47,965.69	53,523.96

5. Una 'caricia' en este texto, es una unidad de reconocimiento, es una forma o manera de medir el reconocimiento, aprecio o afecto que se le tiene a una persona. Hay caricias, sinestésicas o físicas, verbales y gestuales. Otra clasificación de caricias es que existen caricias positivas, negativas y con descuento, estas últimas son aquellas a las que después de decirles lo positivo de algo, van acompañadas a continuación del término 'pero'.

   Suponga que una joven tiene dos pretendientes; ella trata de decidir cuál es el pretendiente más adecuado y para ello mide el afecto que cada uno de ellos le tiene, con base en el número de caricias que le prodiga por día, las caricias positivas las cuenta con números positivos, las negativas con números negativos y a las de descuento les asigna el valor de cero, posteriormente las suma y obtiene un número que puede ser positivo, negativo o cero. Suponga que pone a prueba a sus pretendientes y contabiliza el número de caricias por día a lo largo de un mes, obteniendo los siguientes resultados.

   a. Pruebe la hipótesis de que los datos de ambas opciones tienen distribución de Poisson.
   b. Pruebe la hipótesis de que los datos de ambas opciones tienen distribución normal.
   c. Pruebe la hipótesis de que ambos pretendientes le proporcionan a la interesada el mismo número de caricias.

Pretendiente A		Pretendiente B	
Puntos de Contacto	No. Caricias	Puntos de Contacto	No. Caricias
4	5	1	3
7	5	4	4
6	5	3	2
8	9	3	2
10	6	7	3
7	6	5	3
7	3	2	5
1	5	5	1
11	2	8	0
4	4	4	5
8	1	3	1
5	7	7	1
10	5	8	3
7	4	6	0
3	3	9	3
14	2	2	3
5	7	5	4
3	4	3	4
3	7	3	3
6	4	3	4
4	6	3	1
5	5	3	4
6	3	2	1
6	3	6	1
9	3	8	4
9	5	7	4
6	5	5	2
10	2	4	5
8	3	6	2
5	5	5	3

6. Regresión y correlación lineal simple

6.1. Estadística multivariable y la distribución multinomial

La Estadística Multivariable o Multivariante es la rama de la Estadística que analiza conjuntos de datos multivariables o multivariantes en el sentido de que en un fenómeno o variable de interés intervienen diversas variables aleatorias para cada elemento estudiado.

Ejercicio 6.1

En el caso de una persona, cada individuo presenta diversas características como edad, sexo, peso, estatura y tipo de sangre, entre otras características fisiológicas. Por ejemplo, se considera el Índice de Masa Corporal de una persona como el cociente del peso entre el cuadrado de su estatura. La organización Mundial de Salud (OMS) establece una tabla de clasificación ideada por el estadístico belga Adolphe Quetelet (1796–1874), la cual se muestra en la figura 6.1. El autor de este texto sostiene que el IMC depende de más factores, es decir,

Clasificación de la OMS del estado nutricional de acuerdo con el IMC[6]

Clasificación	IMC (kg/m²)	
	Valores principales	Valores adicionales
Peso bajo	**<18,50**	**<18,50**
Delgadez severa	<16,00	<16,00
Delgadez moderada	16,00-16,99	16,00-16,99
Delgadez leve	17,00-18,49	17,00-18,49
Normal	18,5-24,99	18,5-22,99
		23,00-24,99
Sobrepeso	**≥25,00**	**≥25,00**
Preobesidad	25,00-29,99	25,00-27,49
		27,50-29,99
Obesidad	**≥30,00**	**≥30,00**
Obesidad leve	30,00-34,99	30,00-32,49
		32,50-34,99
Obesidad media	35,00-39,99	35,00-37,49
		37,50-39,99
Obesidad mórbida	≥40,00	≥40,00

Figura 6.1. Clasificación de la OMS sobre el peso

Clasificación de la OMS del estado nutricional de acuerdo con el IMC 2022 recuperada de: https://es.wikipedia.org/wiki/%C3%8Dndice_de_masa_corporal

$IMC = f\left(w = Peso,\ \text{h} = Estatura,\ \text{g} = Género,\ \text{a} = Edad,\ \text{b} = Masa\_Corporal\right)$
$IMC = f\left(\vec{x}\right)$

$\vec{x} = \left(w, h, g, a, b\right)$

Ejercicio 6.2

En la mayoría de las gasolineras de México se venden dos tipos de gasolina: magna y premium. Muchos consumidores de gasolina se preguntan qué gasolina es mejor, la magna o la premium, algunos autores sostienen que es la premium basándose solamente en el número de octanos que tiene la gasolina. Según Pemex, la Premium contiene 92 octanos, en cambio, la magna contiene 87 octanos. El factor que consideran como respuesta para medir qué gasolina es mejor es el rendimiento en kilometraje por litro consumido. El problema se ha vuelto más complejo desde el momento que se privatizó la venta de gasolina y que ahora existen muchas empresas que venden supuestamente los mismos dos tipos de gasolina, ahora se tienen entre otras Pemex, Shell, Exxon Mobil, G500, Gasmart, Redco, Total BP, Chevron Texaco, Gulf, Hidrosina, Grupo ECO, Lodemo, OXXO, Rendichicas, etcétera, es decir, alrededor de 35 en total.

En realidad, la respuesta a todos estos tipos de gasolina presenta varios factores de variación:

$\vec{R} = f\left(\vec{x}\right)$

$\vec{R} = \left(Rendimiento\_\ Km\ /l,\ Precio,\ desgaste\_vehiculo,\ grado\_contaminación\right)$

$$\vec{x} = \begin{pmatrix} tipo\_gasolina,\ empresa,\ cilindraje\_vehiculo,\ marca\_vehículo, \\ condiciones\_automóvil,\ peso\_conductor,\ condiciones\_manejo, \\ condiciones\_ambientales\_ruta \end{pmatrix}$$

Nótese que el vector $\vec{x}$ relaciona a un vector con nueve variables aleatorias y el vector $\vec{R}$ tiene cinco variables aleatorias. El objetivo del estudio sería determinar qué gasolina y de qué empresa es la más adecuada.

La distribución multinomial y la distribución normal multivariada

Este modelo se puede ver como una generalización del Binomial en el que, en lugar de tener dos posibles resultados, se tienen *r* resultados posibles.

Supóngase que el resultado de una determinada experiencia puede ser *r* valores distintos: A_1, A_2, ⋯, A_r cada uno de ellos con probabilidad p_1, p_2, ⋯, p_r, respectivamente.

$$p(A_1) = p_1\,,\ p(A_2) = p_2 \cdots p(A_r) = p_r$$

Si se repite la experiencia *n* veces en condiciones independientes, se puede preguntar la probabilidad de que el suceso A_1 aparezca x_1 veces, el suceso A_2, x_2 veces y así sucesivamente:

$$p\left[(A_1 = \mathrm{x}_1) \cap (A_2 = x_2) \cap \cdots \cap (A_r = x_r)\right]$$

Al modelo estadístico que da dicha probabilidad se le denomina Función de Probabilidad Multinomial, y su función de probabilidad viene dada por:

$$f(x_1, x_2, \cdots, x_r) = p\left[(A_1 = x_1) \cap (A_2 = x_2) \cap \cdots \cap (A_r = x_r)\right]$$

$$= \frac{n!}{x1!\ x2! \cdots x_r!}\ {p_1}^{x1}\ {p_2}^{x2} \cdots {p_r}^{x_r}$$

Donde

$$\sum_{i=1}^{i=r} x_i = n \quad \text{y} \quad \sum_{i=1}^{r} p(A_i) = 1 \tag{6.1}$$

como se ve, el modelo multinomial queda definido por los parámetros (*n*, p_1, p_2, ⋯, p_r). La fórmula anterior puede deducirse de forma análoga al caso Binomial. En realidad, si se toma $r = 2$ se llega exactamente al modelo Binomial.

Las medias y las varianzas de cada x_i en la distribución multinomial están dadas por las expresiones:

$$\mu_1 = E\{x_i\} = np_i \tag{6.2}$$

$$\sigma_i^2 = var\{x_i\} = np_i\left(1 - p_i\right) \tag{6.3}$$

Se debe destacar que este modelo es un ejemplo de distribución multivariante, es decir, de distribución conjunta de varias (r) variables aleatorias. En efecto, si se define la variable aleatoria k_1 como número de veces que se produce el suceso A_1 de un total de n experiencias, y así sucesivamente, se tiene un conjunto de r variables aleatorias discretas cuya función de densidad conjunta (valorada a la vez) viene definida por la anterior fórmula. Nótese que si se considera cada una de estas variables x_i ($i = 1, 2, \cdots, r$) por separado, su distribución es la Binomial de parámetros n y p_i.

Ejercicio 6.3

Tres empresas A_1, A_2 y A_3 ofrecen la posibilidad de hacer una estancia de prácticas profesionales en su planta; tres alumnos deciden concursar teniendo la posibilidad de entrar de 0.4, 0.3 y 0.3 respectivamente. Las estancias se van a asignar en forma independiente. ¿Cuál es la probabilidad de que un solo alumno reciba las tres ofertas?

$$p\left(x_1=3, x_2=0, x_3=0\right) + p\left(x_1=0, x_2=3, x_3=0\right) + \mathrm{p}\left(x_1=0, x_2=0, x_3=3\right) =$$

$$\frac{3!}{3!\,0!\,0!}p_1^3 p_2^0 p_3^0 + \frac{3!}{0!\,3!\,0!}p_1^0 p_2^3 p_3^0 + \frac{3!}{0!\,0!\,3!}p_1^0 p_2^0 p_3^3 + \frac{3!}{3!\,0!\,0!}\left(p_1^3 + p_2^3 + p_3^3\right) =$$

$$\left(0.4^3 + 0.3^3 + 0.3^3\right) = 0.118$$

La distribución más importante de la estadística multivariada es la distribución normal multivariable; la gran importancia de esta distribución radica en el hecho de que a menudo la suma estandarizada de vectores siguiendo cualquier distribución multivariable, en grandes muestras tiende a comportarse como una distribución normal multivariada. Esta conclusión se deduce de una generalización directa del teorema del límite central univariado.

La función de densidad de probabilidades de la distribución normal multivariable está definida por la siguiente expresión:

$$f(x) = \frac{e^{\left[-\frac{1}{2}(\bar{x}-\bar{\theta})'\Sigma^{-1}(\bar{x}-\bar{\theta})\right]}}{(2\pi)^{p/2}|\Sigma|^{1/2}} \tag{6.4}$$

Donde $|\Sigma|$ es el determinante de Σ. El vector μ en estas circunstancias es $E\{x\}$ y la matriz $\Sigma = AA^T$ es la matriz de covarianza de las componentes x_i. Es importante comprender que la matriz de covarianza puede ser singular.

Este caso aparece con frecuencia en estadística; por ejemplo, en la distribución del vector de residuos en problemas ordinarios de regresión lineal. Nótese también que los x_i son en general no independientes; pueden verse como el resultado de aplicar la transformación lineal A a una colección de variables normales Z.

Un caso particular de la distribución normal multivariada es la distribución normal bivariada, cuya expresión matemática se obtiene al hacer $p = 2$, la cual está definida por la expresión:

$$f(\vec{x}) = f(x_1, x_2) =$$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho_2)}\left[\left(\frac{x_1-\theta_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\theta_1}{\sigma_1}\right)\left(\frac{x_2-\theta_2}{\sigma_2}\right) + \left(\frac{x_2-\theta_2}{\sigma_2}\right)\right]\right\}$$

Donde

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \text{ y } \Sigma^{-1} = \begin{bmatrix} \dfrac{1}{\sigma_1^2(1-\rho^2)} & -\dfrac{\rho}{\sigma_1\sigma_2(1-p^2)} \\ -\dfrac{\rho}{\sigma_1\sigma_2(1-\rho^2)} & \dfrac{1}{\sigma_2^2(1-\rho^2)} \end{bmatrix} \qquad (6.5)$$

La gráfica de esta distribución de probabilidad normal bivariada se muestra en la figura 6.2.

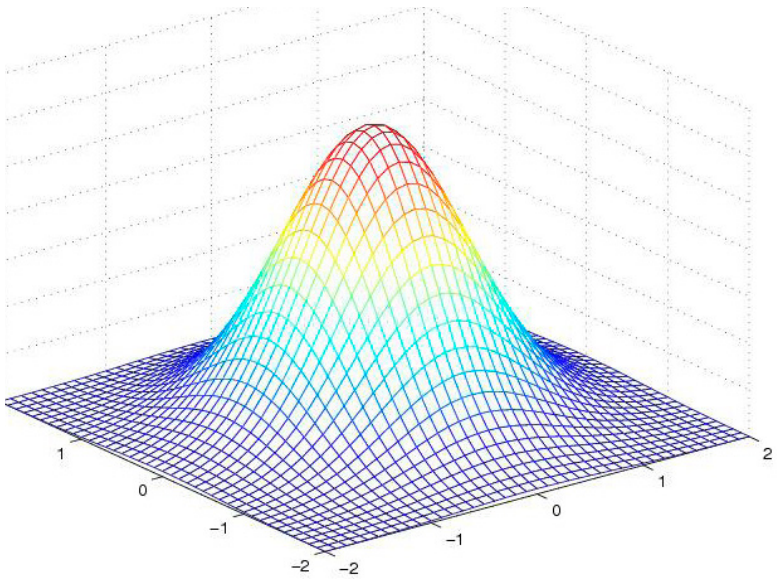


Figura 6.2. Distribución de probabilidad normal bivariada

Recuperada de: https://docplayer.es/docs-images/76/73127434/images/4-0.jpg

Ejercicio 6.4

(Problema 7-41, página 205, Hines, Montgomery, Goldsman y Borror, editorial Patria, Cuarta Edición, 2013, México). La vida útil de un bulbo x_1 y el diámetro del filamento x_2, se distribuyen en forma conjunta como una variable normal bivariada con los siguientes parámetros $\mu_1 = 2000$ horas, $\mu_2 = 0.10$ pulgadas, $\sigma_1^2 = 2500$ horas2, $\sigma_2^2 = 0.01$ pulgadas^2 y $\rho = 0.87$. El gerente de control de calidad desea determinar la vida útil de cada bulbo, midiendo el diámetro del filamento. Si el diámetro de un filamento es menor a 0.098, ¿cuál es la probabilidad de que el bulbo dure al menos 1950 horas?

$$p\left(x_1 > 1950, x_2 \leq 0.98\right) = \frac{1}{2\pi\sqrt{2500}\sqrt{0.01}\sqrt{1-0.87^2}} *$$

$$\iint exp\left\{-\frac{1}{2\left(1-0.87^2\right)}\left[\left(\frac{x_1-2000}{\sqrt{2500}}\right)^2 - 2\left(0.87\right)\left(\frac{x_1-2000}{\sqrt{2500}}\right)\left(\frac{x_2-0.10}{\sqrt{0.01}}\right) + \left(\frac{x_2-0.10}{\sqrt{0.01}}\right)^2\right]\right\} dx_2\, dx_1$$

$1950 \leq x_1 < \infty$
$0 \leq x_2 \leq 0.98$

6.2. Ajuste de la recta de regresión mediante el método de mínimos cuadrados

Una regresión es un ajuste de un conjunto de puntos dados a un modelo matemático en particular de dos o más variables del tipo $y = f(x_1, x_2, \cdots, x_n)$. Cuando el modelo matemático corresponde con una función del tipo $y = f(x)$, se denomina regresión simple. Si el modelo de una sola variable independiente es una recta, se denomina regresión lineal simple. Existen muchos tipos de regresión simple: lineal, exponencial, polinomial, trigonométrica, etcétera. Si el modelo presenta más de una variable independiente se denomina regresión múltiple.

Una correlación corresponde con la definición de indicadores que permitan determinar qué tan bueno es el ajuste entre el conjunto de puntos dados y el modelo matemático usado. La correlación indica el grado de intensidad y la dirección de una relación lineal. Se considera que dos variables cuantitativas están correlacionadas, cuando los valores de una de ellas varían sistemáticamente con respecto a los valores correspondientes de la otra. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad.

Método para llevar a efecto una regresión simple y medir su grado de correlación:

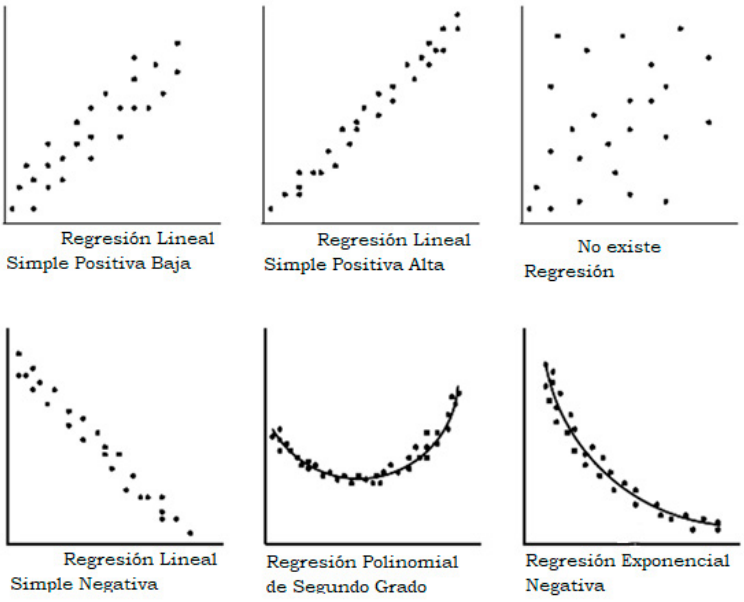
1. Determinar por muestreo un conjunto de n parejas ordenadas (x, y), de las cuales se sospecha que existe algún tipo de relación funcional. Definir exactamente las variables que se analizarán, diseñar una hoja de verificación para su recopilación y captura, y, llevar a cabo dicho proceso. Se requiere recolectar entre 30 y 100 parejas de datos.

2. Graficar en un sistema coordenado xy, al conjunto de puntos (x, y) y observar a ′ojo de pájaro′ si existe algún tipo de relación funcional. A este gráfico se le conoce como Diagrama de Dispersión.

   Para trazar un diagrama de dispersión es necesario dibujar un sistema cartesiano, etiquetando los ejes con los nombres de las variables a graficar y vaciar las parejas de datos obtenidas en el paso previo. Si se encuentra que algunos valores se repiten, se deben rodear los puntos marcados que se repiten con tantas circunferencias como veces en que se repitan dichos datos. También se requiere rotular el diagrama, colocando la fecha de elaboración y los nombres de los que lo elaboraron.

En la figura 6.3 se muestran algunos ejemplos de diagramas de dispersión, que muestran la posible relación entre la *x* (abscisa) y la y (ordenada).

Figura 6.3. Ejemplos de diagramas de dispersión



3. Dependiendo de la gráfica obtenida en el punto (2), establecer a qué modelo se ajustará el conjunto de puntos: lineal, exponencial, logarítmico, polinomial, trigonométrico, etcétera.

4. Por el método de los mínimos cuadrados, esto es, minimizando el Error Cuadrático Medio, determinar los coeficientes o parámetros del modelo matemático, de tal manera que se minimice dicho error.

5. Se define el modelo matemático que mejor se ajusta, especificando su expresión matemática.

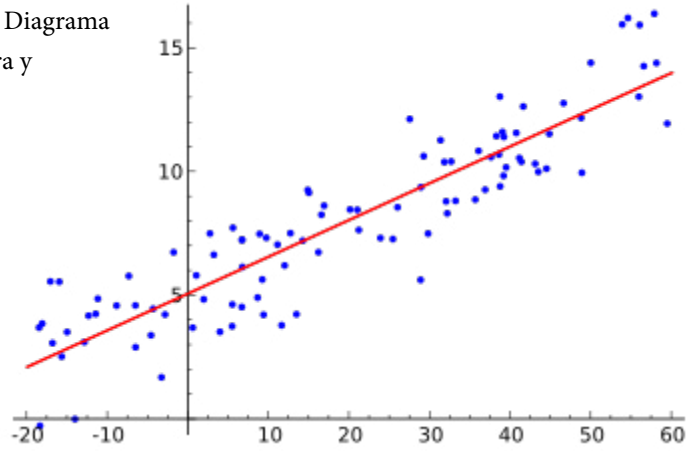
6. Se obtienen estimadores puntuales y por intervalos de los coeficientes o parámetros del modelo, así como de posibles valores de la variable dependiente.

7. Se calculan los coeficientes de determinación y de correlación que permitan estimar el grado de ajuste entre el modelo matemático y los datos obtenidos.

Se expresará a continuación la teoría necesaria para la regresión lineal.

Suponga que se toma aleatoriamente una muestra de *n* parejas ordenadas (x, y) de ciertos valores de dos variables que se sospecha se encuentran relacionadas linealmente, como se muestra en la figura 6.4:

Figura 6.4. Ejemplo de Diagrama de Dispersión de x contra y



Recuperada de: https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/350px-Linear_regression.svg.png

La muestra de parejas ordenadas se lista a continuación:

x	y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮
x_n	y_n

Sea $\hat{y} = \beta_1 x + \beta_0$ la expresión matemática de la recta que se ajustará al conjunto de puntos dados, donde β_1 es el parámetro que representa a la pendiente de la recta y β_0 su ordenada al origen, ambos coeficientes a estimar por el método de los mínimos cuadrados.

Hipótesis básicas o condiciones iniciales para aplicar un modelo de regresión lineal

Se define una desviación o error entre el valor real yi obtenido empíricamente y el valor teórico $\hat{y} = \beta_1 x + \beta_0$ como

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \beta_1 x_i - \beta_0 \tag{6.6}$$

Para poder realizar regresión lineal entre un conjunto de n parejas ordenadas (x_i, y_i), $i = 1, 2, \cdots, n$, y un modelo teórico lineal $\hat{y}_i = \beta_1 x_i + \beta_0$ se deben cumplir ciertas hipótesis o condiciones iniciales, las cuales se enuncian a continuación:

i. La esperanza matemática del error debe ser cero, es decir, $\mu_\varepsilon = E\{\varepsilon\} = 0$

ii. Homocedasticidad. La varianza del error ε_i debe ser constante para cada valor de i, o sea, $\sigma_i^2 = var\{\varepsilon_i\} = \sigma^2$, la cual es desconocida.

iii. Incorrelación o independencia entre ε_i y ε_j para todo $i \neq j$, es decir, $\sigma_{ij}^2 = cov\{\varepsilon_i, \varepsilon_j\} = E\{(\varepsilon_i - \mu_{\varepsilon i})(\varepsilon_j - \mu_{\varepsilon j})\} = 0$. Las covarianzas entre las distintas desviaciones son nulas, lo que quiere decir que no están correlacionadas. Esto implica que el valor de la desviación para cualquier observación muestral no viene influenciada por los valores de las desviaciones correspondientes a otras observaciones muestrales.

iv. Independencia lineal. No existen relaciones lineales exactas entre los regresores.

v. No existen errores de especificación en el modelo, ni errores de medida en las variables explicativas.

vi. Normalidad de las desviaciones, es decir, $\varepsilon \sim N(\mu_\varepsilon = 0, \sigma_\varepsilon^2 = \sigma^2)$.

El error cuadrático medio del ajuste está dado por la siguiente expresión:

$$ECM\{y\} \frac{1}{n} \sum_{i=1}^{i=n} (\gamma_i - \hat{\gamma}_i)^2 = \frac{1}{n} \sum_{i=1}^{i=n} (\gamma_i - \beta_1 x_i - \beta_0)^2 \tag{6.7}$$

Con el objeto de minimizar este error cuadrático medio, se aplica el método de la primera derivada, de manera que derivando e igualando a cero se obtienen las siguientes expresiones:

$$\frac{\partial ECM}{\partial \beta_1} = -\frac{2}{n}\sum_{i=1}^{i=n}\left(y_i - \beta_1 x_i - \beta_0\right)\left(-x_1\right) = 0$$

$$\frac{\partial ECM}{\partial \beta_0} = -\frac{2}{n}\sum_{i=1}^{i=n}\left(y_i - \beta_1 x_i - \beta_0\right)\left(-1\right) = 0$$

Se obtiene el siguiente sistema de dos ecuaciones con dos incógnitas:

$$\left(\sum_{i=1}^{i=n} x_i^2\right)\beta_1 + \left(\sum_{i=1}^{i=n} x_i\right)\beta_0 = \sum_{i=1}^{i=n} x_i\, y_i \qquad (6.8)$$

$$\left(\sum_{i=1}^{i=n} x_i\right)\beta_1 + n\beta_0 = \sum_{i=1}^{i=n} y_i$$

Para resolver el sistema de dos ecuaciones con dos incógnitas, se utilizará el método de la Regla de Kramer, visto en Álgebra Lineal:

$$\beta_1 = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i y_1 & \left(\sum_{i=1}^{n} x_i\right) \\ \sum_{i=1}^{n} y_1 & n \end{vmatrix}}{\begin{vmatrix} \left(\sum_{i=1}^{n} x_i^2\right) & \left(\sum_{i=1}^{n} x_i\right) \\ \left(\sum_{i=1}^{n} x_i\right) & n \end{vmatrix}} = \frac{n\sum_{i=1}^{n} x_i y_1 - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum x_i y_1 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\beta_0 = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i y_1 & \left(\sum_{i=1}^{n} x_i\right) \\ \sum_{i=1}^{n} y_1 & n \end{vmatrix}}{\begin{vmatrix} \left(\sum_{i=1}^{n} x_i^2\right) & \left(\sum_{i=1}^{n} x_i\right) \\ \left(\sum_{i=1}^{n} x_i\right) & n \end{vmatrix}} = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - \beta_1 \bar{x}$$

Si se definen S_{xx}, S_{yy} y S_{xy} de la siguiente forma:

$$S_{xx} = \sum_{i=1}^{i=n} (x_i - \bar{x})^2 = \sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n} \tag{6.9}$$

$$S_{yy} = \sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} y_i^2 - \frac{\left(\sum_{i=1}^{i=n} y_i\right)^2}{n} \tag{6.10}$$

$$S_{xy} = \sum_{i=1}^{i=n} y_i (x_i - \bar{x}) = \sum_{i=1}^{i=n} x_i y_i - \frac{\left(\sum_{i=1}^{i=n} x_i\right)\left(\sum_{i=1}^{i=n} y_i\right)}{n} \tag{6.11}$$

Entonces:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{6.12}$$

$$\hat{\beta}_0 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{i=n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{i=n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6.13}$$

6.3. Los coeficientes de correlación lineal y de determinación

La suma de los cuadrados de los residuos, es decir, la suma de los cuadrados de los errores o desviaciones se puede escribir de la siguiente forma:

$$(y_i - \hat{y}_i) = (y_i - \bar{y}_i) - (\hat{y}_i - \bar{y})$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

Donde

$$\sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = S_{yy}$$

$$\sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = SS_E$$

$$\sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 = \beta_1 S_{xy}$$

Es decir

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy} \qquad (6.14)$$

Por lo que, un estimador insesgado de la varianza poblacional sería la media de cuadrados del error, o sea:

$$\hat{\sigma}^2 = MSE = \frac{SS_E}{n-2} = \frac{S_{yy} - \beta_1 S_{xy}}{n-2} \qquad (6.15)$$

Cabe señalar que durante la deducción de la última expresión (6.15), se dividió una suma de cuadrados en dos sumandos que también son sumas de cuadrados, al dividir una suma de cuadrados entre sus grados de libertad se generan variables aleatorias χ^2; de acuerdo al teorema de aditividad de la distribución χ^2, al sumar dos variables aleatorias χ^2 con n_1 y n_2 grados de libertad respectivamente, el resultado da a su vez una variable aleatoria con $n_1 + n_2$ grados de libertad; lo anterior implica que la suma de cuadrados $\sum_{i=1}^{i=n}(y_i - \bar{y})^2$ presenta $(n-1)$ grados de libertad; el sumando $\sum_{i=1}^{i=n}(\hat{y}_i - \bar{y})^2$ presenta un grado de libertad, por lo que el sumando $\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2$ debe presentar $n-2$ grados de libertad, es decir, $(n-1) = 1 + (n-2)$.

Los parámetros que permitirán medir qué tan bueno es el ajuste a la recta de regresión lineal se definen a continuación:

Coeficiente de determinación:

$$R^2 = \frac{\sum_{i=1}^{i=n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{i=n}(y_i - \bar{y})^2} = 1 - \frac{SS_E}{S_{xy}} \tag{6.16}$$

Este coeficiente se emplea a menudo para juzgar la suficiencia de un modelo de regresión. En el caso específico en el que tanto x como y son variables aleatorias distribuidas en forma conjunta, entonces este coeficiente R^2 representa el cuadrado del coeficiente de correlación entre x y y. Es claro que $0 < R^2 < 1$.

Coeficiente de correlación ρ

$$\rho = \frac{\sum_{i=1}^{i=n} y_i\,(x_i - \bar{x})}{\left[\left(\sum_{i=1}^{i=n}(x_i - x)^2\right)\left(\sum_{i=1}^{i=n}(y_i - y)^2\right)\right]^{1/2}} = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} \tag{6.17}$$

Frecuentemente es útil probar la hipótesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0 \tag{6.18}$$

El estadístico de prueba para esta hipótesis es

$$t_0 = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \tag{6.19}$$

El cual sigue una distribución t con n–2 grados de libertad t_{n-2} si H_0 es verdadera. Esto implica que se rechazaría la hipótesis nula si $t_0 > t_{\alpha/2,n-2}$ o $t_0 < -t_{\alpha/2,n-2}$.

El procedimiento de prueba de hipótesis

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0 \end{aligned} \tag{6.20}$$

Es diferente y se aplica para muestras con $n > 25$. Para ello, se utiliza el estadístico de prueba

$$z = \text{ang tanh}(\rho) = \frac{1}{2} Ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \tag{6.21}$$

El cual presenta la siguiente distribución:

$$z \sim N\left(\mu_z = \text{ang tanh}(\hat{\rho}), \sigma_z = \frac{1}{\sqrt{n-3}}\right) \tag{6.22}$$

Por lo que para probar la hipótesis nula de la expresión 6.20, es necesario calcular el estadístico

$$z_0 = \left(\text{ang tanh}(\hat{\rho}) - \text{ang tanh}(\rho_0)\right)\sqrt{n-3} \tag{6.23}$$

La hipótesis nula de la expresión 6.20 se rechazaría si $z_0 > z_{\alpha/2}$ o $z_0 > z_{\alpha/2}$

También es posible deducir un intervalo de confianza para el coeficiente de correlación, usando la expresión 6.21 y sabiendo del curso de cálculo que

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

El estimador por intervalos para el coeficiente de correlación lineal sería

$$\tanh\left[\text{ang tanh}(\hat{\rho}) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right] \leq p \leq \tanh\left[\text{ang tanh}(\hat{\rho}) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right] \tag{6.24}$$

6.4. Intervalo de confianza para la pendiente y para la ordenada al origen de la recta de regresión lineal

Suponga que se pretende probar la siguiente hipótesis estadística:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0} \qquad (6.25)$$

Para probar esta hipótesis se debe hacer la siguiente suposición adicional:

$$\varepsilon \sim N\left(\mu_\varepsilon = 0, \sigma_\varepsilon = \sigma\right) \qquad (6.26)$$

De esta última expresión se desprende entonces que las observaciones yi también deben suponerse normales:

$$y_i \sim N\left(\hat{\mu}_{y_i} = \beta_0 + \beta_1 x_1, \sigma\right) \qquad (6.27)$$

Por otra parte

$$\sigma^2_{\hat{\beta}_1} = var\left\{\hat{\beta}_1\right\} = \frac{1}{S^2_{xx}} \text{ var}\left\{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)\right\} = \frac{\sigma^2}{S_{xx}} \qquad (6.28)$$

Recuerde que el cociente de una normal estándar entre la raíz cuadrada de una ji cuadrada, entre sus grados de libertad, da origen a una variable aleatoria t de Student, lo que implica que se puede usar el estadístico:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{MS_E}{S_{xx}}}} \qquad (6.29)$$

Para probar la hipótesis de la expresión 6.25, se rechazaría la hipótesis nula si se cumple que $t_0 > t_{\alpha/2,n-2}$ o $t_0 < t_{\alpha/2,n-2}$.

De la misma forma, se usa un procedimiento similar para probar la hipótesis estadística respecto de la ordenada al origen:

$$H_0 : \beta_0 = \beta_{0,0}$$
$$H_1 : \beta_0 \neq \beta_{0,0} \qquad (6.30)$$

Usando el estadístico:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \tag{6.31}$$

Se rechazaría la hipótesis nula si se cumple que $t_0 > t_{\alpha/2,n-2}$ o $t_0 < t_{\alpha/2,n-2}$.

Si se sustituyen las expresiones 6.12 y 6.15 en la expresión 6.14, se puede apreciar que la suma de cuadrados Syy se puede descomponer en una suma de cuadrados debida al error aleatorio $SS_E = (n-2)\sigma^2$, y una suma de cuadrados del ajuste que se hace de los datos dados a una recta teórica de regresión $SS_R = S_{xy}^2/S_{xx}$. Se puede demostrar que S_{yy} es una variable aleatoria χ^2 con $n-1$ grados de libertad, SS_E es una variable aleatoria χ^2 con $n-2$ grados de libertad y SS_R es una variable aleatoria $\chi 2$ con 1 grado de libertad; recordando la definición de una variable aleatoria tipo F de Fisher-Snedecor, con k_1 grados de libertad en el numerador y k_2 grados de libertad en el denominador, entonces se puede afirmar que el siguiente cociente presenta distribución F:

$$F_0 = \frac{SS_R / 1}{SS_E / (n-2)} = \frac{MS_R}{MS_E} \sim F_{1,n-2} \tag{6.32}$$

El estadístico anterior puede servir para probar si existe o no una recta de regresión lineal con pendiente diferente de cero que se ajuste a los datos, a través de la siguiente prueba de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0 \tag{6.33}$$

La hipótesis nula H_0 se rechazaría si $F_0 > F_{\alpha,1,n-2}$.

Para llevar a cabo esta prueba se utiliza un arreglo tabular al cual se le conoce como tabla ANOVA (Analysis of Variance) para la regresión, cuyo formato se muestra a continuación:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0	p
Regresión	SS_R	1	$MS_R = SS_R$	MS_R / MS_E	$p(F_{1,n-2} > F_0)$
Error	SS_E	n-2	$MS_E = SS_E/(n-2)$		
Total	S_{yy}	n-1			

De tal manera que para probar la hipótesis de la expresión 6.33 se pueden utilizar dos caminos diferentes, utilizando la expresión 6.29 con el estadístico t o utilizando la expresión 6.32 con el estadístico F, a través de la tabla ANOVA anterior, ambos métodos conducen al mismo resultado.

A partir de los estadísticos de las expresiones 6.29 y 6.31, se pueden deducir los estimadores por intervalos para la pendiente y la ordenada al origen de la recta de regresión lineal, los cuales estarían dados por las siguientes fórmulas:

$$\hat{\beta}_1 - t_{\alpha/2,1,n-2}\sqrt{\frac{MS_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,1,n-2}\sqrt{\frac{MS_E}{S_{xx}}} \tag{6.34}$$

$$\hat{\beta}_0 - t_{\alpha/2,1,n-2}\sqrt{MS_E\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,1,n-2}\sqrt{MS_E\left[\frac{1}{n}+\frac{x^2}{S_{xx}}\right]} \tag{6.35}$$

6.5. Bandas de confianza para la recta de regresión

El estimador puntual para un valor y_0, correspondiente a un valor x_0 de la tabla, está dado en la misma. En este caso, su estimador por intervalos sería:

$$\hat{y}_0 - t_{\alpha/2,\, n-2}\sqrt{MS_E\left[1+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le y_0 \le \hat{y}_0 + t_{\alpha/2,\, n-2}\sqrt{MS_E\left[1+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (6.36)$$

El estimador puntual para un valor y_0, para un valor x_0 no presente en la tabla, pero sí entre los valores de la misma como interpolación, o fuera del intervalo que forman los valores de la tabla como extrapolación, estaría dado por:

$$\hat{y}_0 = \hat{\beta}_1 x_0 + \hat{\beta}_0 \qquad (6.37)$$

En este caso, su estimador por intervalos sería:

$$\hat{y}_0 - t_{\alpha/2,\, n-2}\sqrt{MS_E\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le y_0 \le \hat{y}_0 + t_{\alpha/2,\, n-2}\sqrt{MS_E\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (6.38)$$

Ejercicio 6.5

Se sospecha que la estatura de los hijos varones depende de la estatura de los padres. La siguiente tabla muestra las respectivas estaturas en pulgadas, del padre *x*, y del hijo *y*, de una muestra de 12 padres y sus hijos mayores.

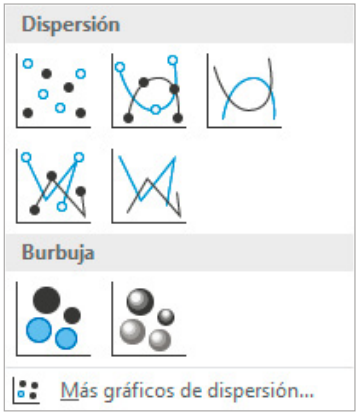
Estatura del padre	**Estatura del hijo**
x	*y*
65	68
63	66
67	68
64	65
68	69
62	66
70	68
66	65
68	71
67	67
69	68
71	70

a. Trace un diagrama de dispersión del conjunto de datos dados y determine si visualmente presentan una relación aproximadamente lineal.

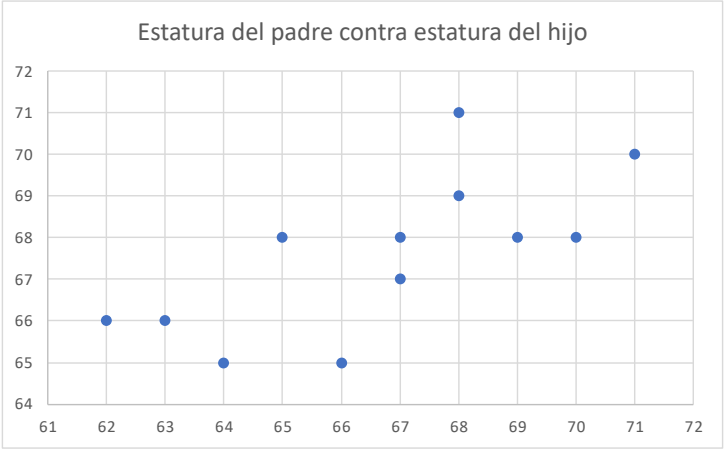
Para trazar el diagrama de dispersión con Excel, se realizan los siguientes pasos:

i. Se capturan las dos columnas de datos dados en Excel.

ii. En la parte superior de la hoja de cálculo de Excel, se elije el menú Insertar, luego se le da un click al submenú  situado en la parte central, encima de donde dice Gráficos, donde aparece la siguiente pantalla:



En esta pantalla se elige la primera opción, dando click en donde aparecen solo puntos en el ícono, entonces aparece la figura siguiente:



En esta figura se percibe una tendencia a crecer a medida que se avanza a la derecha, pero no está muy bien definida la relación lineal.

b. Ajuste un modelo de regresión lineal a los datos dados, estimando puntualmente la pendiente y la ordenada al origen de la recta de regresión.

Si se utiliza Excel manualmente, se forma la siguiente tabla:

No.	x	y	x^2	y^2	xy
1	65	68	4225	4624	4420
2	63	66	3969	4356	4158
3	67	68	4489	4624	4556
4	64	65	4096	4225	4160
5	68	69	4624	4761	4692
6	62	66	3844	4356	4092
7	70	68	4900	4624	4760
8	66	65	4356	4225	4290
9	68	71	4624	5041	4828
10	67	67	4489	4489	4489
11	69	68	4761	4624	4692
12	71	70	5041	4900	4970
Suma =	**800**	**811**	**53418**	**54849**	**54107**

Con esta tabla se calculan *Sxx*, *Syy* y *Sxy*:

$$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 84.6667$$

$$Syy = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 38.9167$$

$$Sxy = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 40.3333$$

De tal manera que los coeficientes de la recta de regresión estimada son:

$\beta_1 = Sxy/Sxx = 40.3333/84.6667 = 0.47638$

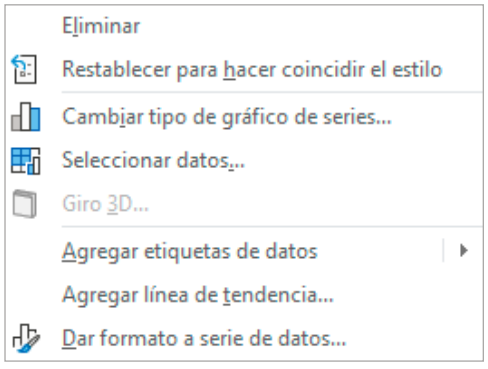
$\beta_0 = \bar{y} - \beta_1 \bar{x} = 811/12 - 0.47638(800/12) = 35.8248$

Por lo cual, la recta de regresión lineal tiene la siguiente expresión

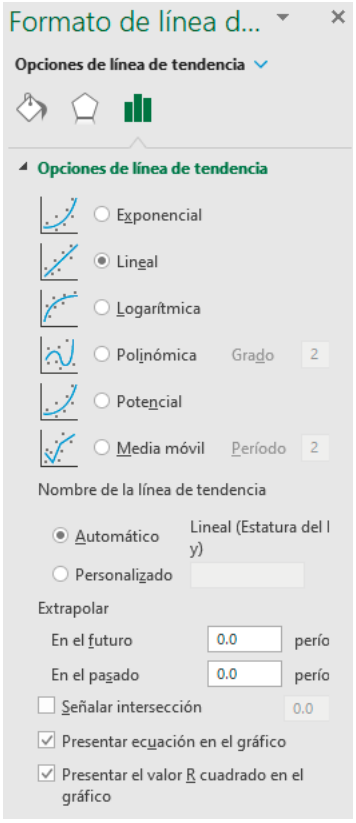
$y = 0.47638x + 35.8248$

Esta recta se puede obtener automáticamente, sin necesidad de calcular la tabla dada anteriormente en Excel, si en el diagrama de dispersión trazado se realizan los siguientes pasos:

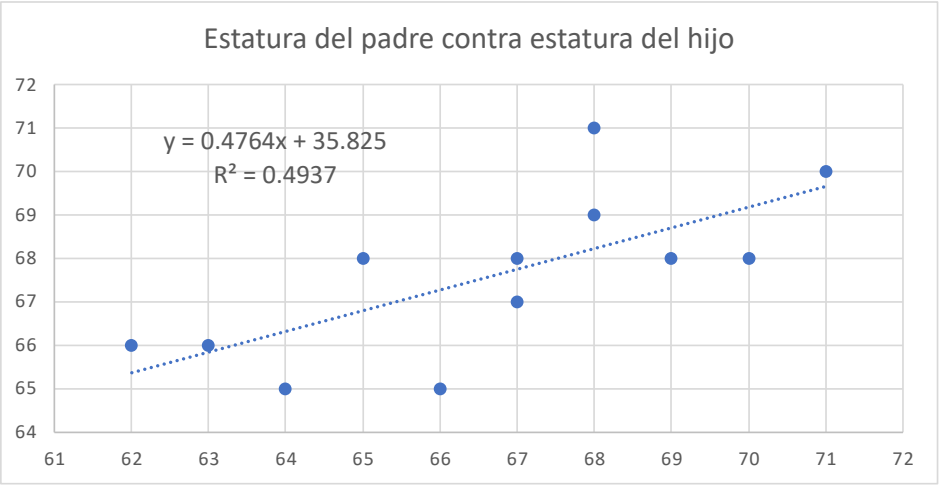
i. Dar un click con el botón derecho del ratón, dentro del diagrama de dispersión en Excel. Aparece la siguiente pantalla:



En esta pantalla se selecciona el menú Agregar línea de tendencia, apareciendo la siguiente pantalla:



Aquí se selecciona la opción de línea de tendencia que se pretende hacer, en este caso lineal (la cual se marca como predeterminada), y se marcan las opciones Presentar ecuación en el gráfico y Presentar el valor R cuadrado en el gráfico, obteniéndose la siguiente figura:



c. Obtenga un estimador de la varianza σ^2

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2} = \frac{S_{yy} - \beta_1 S_{xy}}{n-2} = 1.97028$$

d. Obtenga intervalos de confianza de la pendiente y de la ordenada al origen.

Primero se obtendrá el valor de $t\alpha/2, n-2$ usando Excel con el comando

$$t_{\alpha/2,\, n-2} = \text{INV.T}\left(1-\alpha/2, n-2\right)$$

$t_{0.025,\,10} = 2.22814$ *al 95%_bilateral*

$t_{0.005,10} = 3.16927$ *al 99%_bilateral*

$$\hat{\beta}_1 - t_{\alpha/2,\, n-2}\sqrt{\frac{MS_E}{Sxx}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,\, n-2}\sqrt{\frac{MS_E}{Sxx}}$$

$0.13648 \le \beta_1 \le 0.81628$ *al* 95%
$-0.0071 \le \beta_1 \le 0.95984$ *al* 99%

De la misma forma se obtiene β_0

$$\hat{\beta}_0 - t_{\alpha/2,\, n-2}\sqrt{\frac{MS_E}{Sxx}\left[\frac{1}{n} + \frac{x}{Sxx}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,\, n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{x}{Sxx}\right]^2}$$

$13.1469 \leq \beta_0 \leq 58.5027$ *al* 95%
$3.56809 \leq \beta_0 \leq 68.0815$ *al* 99%

e. Pruebe la hipótesis estadística de que la recta de regresión tiene pendiente cero.

Se hará por los tres métodos ya vistos en el subtema de pruebas de hipótesis.

i. Ya se tienen los intervalos de confianza de la pendiente de la recta de regresión, nótese que, según la hipótesis nula, el valor de $\beta_1 = 0$ no está contenido en el intervalo bilateral al 95% de nivel de confianza, por lo cual se rechazaría la hipótesis nula; sin embargo, al 99% sí está contenido, por lo cual no se podría rechazar la hipótesis nula. Como cae en zona de duda, lo prudente es tomar una muestra de mayor tamaño, ya que $n = 12$ se considera pequeña.

ii. Se calcula el estadístico t_0

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1H_0}}{\sqrt{\frac{MSE}{Sxx}}} = 3.1228$$

Nótese que $t_{0.025,10} < t_0 < t_{0.005,10}$, por lo que cae en zona de duda, lo prudente es tomar una muestra de mayor tamaño, ya que $n = 12$ se considera pequeña.

iii. Se calcula la probabilidad p = p(t > t0)

p = p(t > t0) = DISTR.T.CD(t0,n-2) = DISTR.T.CD(3.1228,10) = 0.00541

Nuevamente, se aprecia que 0.005 < p < 0.025, por lo que cae en zona de duda, lo prudente es tomar una muestra de mayor tamaño, ya que n = 12 se considera pequeña.

Por cualquiera de los tres métodos se puede afirmar que no se puede rechazar la hipótesis nula de que la pendiente de la recta de regresión es cero, para tener mayor confianza se sugiere aumentar el tamaño de la muestra, ya que $n = 12$ se considera insuficiente.

f. Estime los indicadores de correlación: coeficiente de determinación y coeficiente de correlación.

Coeficiente de Determinación:

$$R^2 = 1 - \frac{SSE}{S_{yy}} = 0.5115$$

Coeficiente de Correlación ρ:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = 0.70265$$

g. Obtenga un intervalo de confianza del coeficiente de correlación.

Primero se calcula el valor de $z_{\alpha/2}$

$z_{0.025} = 1.959964$
$z_{0.005} = 2.575829$

$$\tanh\left(\text{ang tanh}\left(\hat{\rho}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\text{ang tanh}\left(\hat{\rho}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

$0.215753 \leq p \leq 0.90971 \quad al \qquad 95\%\_bilateral$
$0.013908 \leq p \leq 0.93919 \quad al \qquad 99\%\_bilateral$

h. Probar la hipótesis unilateral de que $\rho > 0.95$

Se calculará un intervalo unilateral inferior de confianza del coeficiente de correlación.

Primero se calcula el valor de z_{α}

$z_{0.025} = 1.959964$
$z_{0.005} = 2.575829$

$$\rho < \tanh\left(\text{ang tanh}\left(\hat{\rho}\right) + \frac{z_{\alpha}}{\sqrt{n-3}}\right)$$

$\rho \leq 0.88977 \quad al \quad 95\%\_confianza$
$\rho \leq 0.92858 \quad al \quad 99\%\_confianza$

Nótese que se rechaza la hipótesis nula de que el coeficiente de correlación es mayor a 0.95

i. Obtenga el valor estimado puntual de y cuando $x = 70$ pulgadas y determine un intervalo de confianza para el mismo.

De la tabla de datos se observa que $y(x = 70) = 68$

$t_{\alpha/2,n-2} = \text{INV.T}(1-\alpha/2,n-2)$

$t_{0.025,10} = 2.22814 \quad al \quad 95\%\_bilateral$
$t_{0.005,10} = 3.16927 \quad al \quad 99\%\_bilateral$

$$\hat{y}_0 - t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{(x_0-\bar{x})^2}{Sxx}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{(x_0-\bar{x})^2}{Sxx}\right)}$$

$64.67354 \leq y\,(x = 70) \leq 71.3265 \quad al \quad 95\%$
$63.2685 \leq y\,(x = 70) \leq 71.3265 \quad al \quad 99\%$

j. Obtenga el valor estimado puntual de y cuando $x = 68$ pulgadas y determine un intervalo de confianza para el mismo.

De la tabla de datos se observa que existen dos valores de y para $x = 68$: $y_1(x = 68) = 69$ y $y_2(x = 68) = 71$, por lo cual se toma el promedio de ambos: $y = (69 + 71)/2 = 70$.

$t_{\alpha/2,n-2} = \text{INV.T}(1-\alpha/2,n-2)$

$t_{0.025,10} = 2.22814 \quad al \quad 95\%\_bilateral$
$t_{0.005,10} = 3.16927 \quad al \quad 99\%\_bilateral$

$$\hat{y}_0 - t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{(x_0-\bar{x})^2}{Sxx}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{(x_0-\bar{x})^2}{Sxx}\right)}$$

$66.8398 \leq y\,(x = 68) \leq 73.1602 \quad al \quad 95\%$
$65.5049 \leq y\,(x = 68) \leq 74.4951 \quad al \quad 99\%$

k. Obtenga el valor estimado puntual de y cuando $x = 66.5$ pulgadas y determine un intervalo de confianza para el mismo.

De la tabla de datos se observa que no existe una medición para $x = 66.5$ por lo cual este valor debe estimarse usando la recta de regresión lineal.

$$y(x = 66.5) = 0.47638(66.5) + 35.8248 = 67.5039$$

$$t_{\alpha/2,n-2} = \text{INV.T}(1-\alpha/2,n-2)$$

$t_{0.025,10} = 2.22814 \quad al \quad 95\%\_bilateral$
$t_{0.005,10} = 3.16927 \quad al \quad 99\%\_bilateral$

$$\hat{y}_0 - t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{Sxx}\right)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{Sxx}\right)}$$

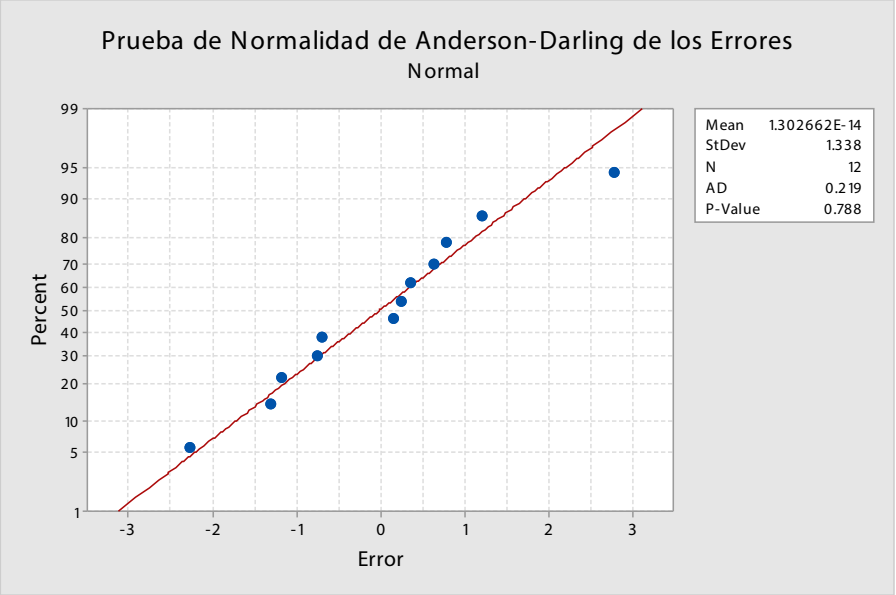
$64.24818 \le y\,(x = 68) \le 70.7597 \quad al \quad 95\%$
$62.87299 \le y\,(x = 66.5) \le 72.1349 \quad al \quad 99\%$

l. Realice un análisis de sensibilidad sobre el cumplimiento de las condiciones iniciales para la regresión lineal.

Lo primero a realizar es probar la hipótesis de normalidad del error aleatorio, para lo cual se elabora la siguiente tabla:

No.	Estatura del padre x	Estatura del hijo y	Estatura del hijo y^	Error y-y^
1	65	68	66.78937008	1.210630
2	63	66	65.83661417	0.163386
3	67	68	67.74212598	0.257874
4	64	65	66.31299213	-1.312992
5	68	69	68.21850394	0.781496
6	62	66	65.36023622	0.639764
7	70	68	69.17125984	-1.171260
8	66	65	67.26574803	-2.265748
9	68	71	68.21850394	2.781496
10	67	67	67.74212598	-0.742126
11	69	68	68.69488189	-0.694882
12	71	70	69.6476378	0.352362
Media=	66.666667	67.583333	67.583333	0.000000
Varianza=	7.696970	3.537879	1.746719	1.791160

Utilizando Minitab y aplicando la prueba de Anderson-Darling, se obtiene la siguiente figura, se puede notar que *p* es mayor de 0.05 y mayor de 0.01, por lo que no existe evidencia estadística suficiente para rechazar la hipótesis nula, la cual afirma que los errores aleatorios se distribuyen normalmente con media cero.



Ejercicios del Capítulo 6

1. Los datos siguientes corresponden al cloro residual en una alberca, medido a diversos tiempos luego de haber sido tratada químicamente.

Tiempo (horas)	Cloro (ppm)
1	1.81
1	1.68
2	1.52
2	1.93
2	1.66
3	1.67
3	1.69
4	1.61
4	1.52
4	1.6
5	1.43
5	1.38
5	1.57
6	1.43
6	1.5

Tiempo (horas)	Cloro (ppm)
7	1.43
7	1.24
7	1.35
8	1.24
8	1.25
9	1.16
9	1.17
9	1.06
10	1.04
10	1.12
11	1.02
11	1.07
11	0.82
12	0.66
12	0.92

   a. Trace un diagrama de dispersión para percibir si existe alguna posible relación entre el tiempo en que se mide el cloro que existe en la alberca y el contenido de cloro residual en la misma.
   b. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza, la pendiente y la ordenada al origen de la recta de regresión.
   c. Pruebe la hipótesis de que la pendiente de la recta de regresión es cero.
   d. Calcule la varianza del error de ajuste.

e. Obtenga el coeficiente de determinación del ajuste a una recta de regresión.
f. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza, al coeficiente de correlación lineal.
g. Pruebe la hipótesis de que el coeficiente de correlación es menor a -0.9.
h. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza el nivel de cloro residual cinco horas después de haber depositado el cloro.
i. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza el nivel de cloro residual seis y media horas después de haber depositado el cloro.
j. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza el nivel de cloro residual 15 horas después de haber depositado el cloro. ¿Qué condiciones debe cumplir la extrapolación?

k. Realice un análisis de sensibilidad sobre el cumplimiento de las condiciones iniciales para la regresión lineal.

2. El peso y la presión arterial de 26 personas de sexo masculino, seleccionadas al azar, en un grupo con edades de 25 a 30 años, se muestra en la siguiente tabla.

Sujeto	Peso	Presión
1	74.9	130
2	75.8	133
3	81.7	150
4	70.3	128
5	96.2	151
6	79.4	146
7	86.2	150
8	95.3	140
9	90.8	148
10	67.6	125
11	71.7	133
12	76.7	135
13	77.1	150

Sujeto	Peso	Presión
14	78	153
15	72.1	128
16	76.2	132
17	78.9	149
18	83	158
19	97.6	150
20	88.5	163
21	81.7	156
22	64.9	124
23	108.9	170
24	106.6	165
25	87.1	160
26	84.8	159

a. Trace un diagrama de dispersión para percibir si existe alguna posible relación entre el peso y la presión interna máxima de una persona.
b. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza, la pendiente y la ordenada al origen de la recta de regresión.
c. Pruebe la hipótesis de que la pendiente de la recta de regresión es uno.
d. Calcule la varianza del error de ajuste.
e. Obtenga el coeficiente de determinación del ajuste a una recta de regresión.
f. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza, al coeficiente de correlación lineal.
g. Pruebe la hipótesis de que el coeficiente de correlación vale 0.7.
h. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza la presión de una persona con 83 kg de peso.
i. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza la presión arterial de una persona con 100 kg de peso.
j. Estime puntualmente y por intervalos de confianza, al 95% y al 99% de nivel de confianza la presión arterial de una persona con 150 kg de peso. ¿Qué condiciones debe cumplir la extrapolación?
k. Realice un análisis de sensibilidad sobre el cumplimiento de las condiciones iniciales para la regresión lineal.

Bibliografía de referencia

1. William W. Hines, Douglas C. Montgomery, David M. Goldsman, Connie M. Borror. *Probabilidad y Estadística para Ingeniería.* Editorial Patria. Cuarta Edición. México, 2013.
2. Maria Dolores Ugarte, Ana F. Militino, Alan T. Arnholt. *Probability and Statistics with R.* Chapman and Hall/CRC, 2015.
3. Alfredo H. S. Ang and Wilson H. Tang. *Probability Concepts in Engineering Planning and Design.* John Wiley & Sons, Inc, 1975.
4. Bernard Ostle, "Estadística Aplicada". Editorial LIMUSA, 1983.
5. Jack R. Benjamin & C. Allin Cornell. *Probability, Statistics and Decision for Civil Engineers.* McGraw-Hill Book Company, 1970.
6. Jerome L. Myers, Arnold D. Well, Robert F. Lorch Jr. *Research Design and Statistical Analysis.* Ed. Routledge. New York. Third edition. 2010.
7. Jean Dickinson Gibbons, Subhabrata Chakraborti. *Nonparametric Statistical Inference.* CRC Press Fith Edition, 2011.
8. Norman R. Draper, Harry Smith. *Applied Regression Analysis.* Wiley Interscience Publication. Third Edition, 1998.
9. Canavos, G. *Probabilidad y Estadística. Aplicaciones y Métodos.* México: McGraw-Hill. 2003.
10. Mendenhall, W., et al. *Estadística Matemática con Aplicaciones.* México: Grupo Editorial Iberoamérica. 2008.
11. Ronald E. Walpole et al. *Probabilidad y Estadística para Ingeniería y Ciencias.* Editorial Pearson 2012.

Fundamentos de Estadística y Aplicaciones, con R, Minitab y Excel
se publicó digitalmente en el repositorio de la Facultad de Ingeniería el8 de mayo de 2023.
Primera edición electrónica de un ejemplar (14 MB) en formato PDF.

El cuidado de la edición y diseño estuvieron a cargo de la Unidad de Apoyo Editorial de la Facultad de Ingeniería. La familia tipográfica utilizada fueron Minion Pro y Chivo con sus respectivas variantes.