ESTADÍSTICA

Profesores:
A. Leonardo Bañuelos Saucedo
Nayelli Manzanarez Gómez

NOTAS

TEMA 1

ESTADÍSTICA DESCRIPTIVA

TEMA I
ESTADÍSTICA DESCRIPTIVA

INTRODUCCIÓN

La *Probabilidad* y la *Estadística* son herramientas muy importantes en el desarrollo de cualquier ingeniería. Sus aplicaciones van desde los juegos de azar hasta la confiabilidad de sistemas, estimaciones de datos para variables inciertas, toma de decisiones en situaciones de incertidumbre, estudio de los efectos del ruido en sistemas electrónicos, el diseño de centrales telefónicas, etc.

Actualmente, los conceptos y métodos básicos de la estadística son indispensables para describir, comprender e intentar predecir el comportamiento del mundo que nos rodea. La estadística nos proporciona los elementos para comprender la información y poder obtener conclusiones con un soporte matemático. El presente curso es un acercamiento al manejo de datos estadísticos, junto con las bases probabilísticas para desarrollar el potencial de la inferencia estadística.

INVESTIGACIÓN

La humanidad, desde siempre, ha estado expuesta a innumerables problemas, lo que la ha llevado a buscar la mejor forma de resolverlos, dando origen a la investigación, y la forma científica de validar las investigaciones es a través de la probabilidad y la estadística. El término investigar proviene del vocablo griego *investigare* que significa "hacer diligencias para descubrir algo", la investigación está encaminada a conseguir estos descubrimientos. En la actualidad se habla de dos tipos de investigación: la básica y la aplicada.

La investigación básica tiene por objeto mismo la investigación, es decir, sólo persigue el conocimiento de las cosas. Por otro lado, como su nombre lo indica la investigación aplicada tiene como propósito el resolver problemas específicos. Es común que la investigación que comienza como básica termine siendo aplicada. Por ejemplo, en la antigüedad los griegos desarrollaron y estudiaron la geometría (medida de la tierra), y la formalizaron a través de la lógica deductiva; sin embargo, la estudiaban principalmente por sus valores estéticos y por considerarla símbolo de cultura. En la actualidad la geometría elemental encuentra sus aplicaciones en los objetos que nos rodean, en las edificaciones y en la forma en que se miden los objetos.

El ingeniero, como profesional, debe estar capacitado para desarrollar investigación aplicada, dejando la investigación básicas a investigadores de carrera.

EL MÉTODO CIENTÍFICO

Los científicos e investigadores, con el objeto de lograr sus propósitos deben seguir una metodología. Sin la intención de entrar en una discusión detallada del método científico, se pueden resaltar en él los siguientes puntos:

1. Parte de una revisión de hechos, teorías y propuestas.
2. Se formula una hipótesis lógica sujeta a prueba mediante métodos experimentales.
3. Se evalúa la hipótesis de manera objetiva con base en los resultados experimentales.

Los científicos, que deben ser capaces de observar y clasificar los hechos que resultan de un diseño experimental, tienen que recurrir a la estadística.

LA ESTADÍSTICA EN LA INVESTIGACIÓN

Para entender el papel de la estadística en la investigación es conveniente tener una idea de lo que es la *estadística*. La palabra estadística significa literalmente "ciencia del estado", debido a que en sus inicios la estadística servia para proporcionar datos que fueran de interés para los gobernantes de una nación. En la actualidad la estadística es mucho más que eso.

La estadística no sólo proporciona información o datos; sino que los agrupa, analiza, interpreta y permite generar inferencias o conclusiones de una población a partir de los datos de una muestra. Son muchas las aplicaciones de la estadística en la investigación. Por ejemplo; en política, es deseable saber que en porcentaje de una población votará en favor de un candidato, sin tener que entrevistar a todos los posibles votantes; en la industria, es conveniente determinar si un lote de productos cumple con ciertos estándares de calidad o deben reprocesar las piezas; en la educación, qué tanto afecta la escuela de procedencia en el aprovechamiento de un grupo de alumnos que ingresaron al nivel superior; en biología, los resultados sobre el crecimiento de vegetales en función de determinadas variables controlables, etc., Todas estas son interrogantes o predicciones que contesta la estadística, por ello la gran relación entre la estadística y la investigación.

LA ESTADÍSTICA Y SUS CLASIFICACIONES

La estadística es la rama de las matemáticas que se encarga de la selección de datos, su organización, presentación y realización de las conclusiones que se pueden obtener de dichos datos.

La estadística puede clasificarse en: **univariable** y **multivariable**, dependiendo de la cantidad de variables que se estén registrando. Si sólo interesa el peso de las personas, entonces se desarrollará estadística univariable, si por el contrario se pretende estudiar la relación entre el peso y la estatura, entonces se estará desarrollando estadística multivariable.

Otra clasificación está basada en la aplicación de la estadística. La estadística **descriptiva** (o deductiva) tiene como propósito la recopilación, organización y presentación de datos para su estudio, mientras que la estadística **inferencial** (o inductiva) tiene como objetivo obtener conclusiones con respecto a una población a partir de la información contenida en una muestra, cuantificando de manera probabilística el grado de certeza de la afirmación. A la estadística descriptiva se le llama también deductiva, mientras que a la estadística inferencial se le llama inductiva.

La estadística descriptiva utiliza gráficas, tablas y parámetros numéricos para la presentación de la información. La estadística inferencial utiliza técnicas de probabilidad para cuantificar el grado de certidumbre de las conclusiones.

Otra clasificación de la estadística está basada en la información que se posee. La **estadística paramétrica** es la rama de la estadística que estudia las pruebas y modelos en los que se conoce la distribución de la población bajo estudio, o que por las condiciones del muestreo, se sabe la distribución que se debe utilizar para el análisis. La **estadística no-paramétrica** estudia las pruebas y modelos cuando la distribución no puede ajustarse mediante la estadística paramétrica, esto ocurre generalmente cuando no se conoce la distribución poblacional.

LA POBLACIÓN Y LA MUESTRA

Para comprender la naturaleza de la inferencia estadística deben distinguirse dos grandes conjuntos: la *población* y la *muestra*. La **Población** es el conjunto de todos los posibles resultados de un experimento, mientras que la **Muestra** es un subconjunto de la población que contiene los resultados observados de un experimento. Debe entenderse que el principal objetivo de la estadística es realizar inferencias (o predicciones) de la población a partir de los datos observados en la muestra; la importancia de esto puede verse en los siguientes ejemplos.

Si se desea conocer el porcentaje de la población que votará por un candidato en particular, el hecho de entrevistar a todos los posibles votantes requeriría de un gran esfuerzo además de un gran costo, por lo que debe realizarse la entrevista solo a un grupo de los votantes (muestra).

Si se desea conocer el tiempo promedio de vida de un foco, el probar toda la población sería el equivalente a prender todos los focos y medir el tiempo que tardan en fundirse, lo cual no permitiría tener artículos para la venta, por lo cual la prueba de vida útil sólo se realiza a un grupo de focos (muestra).

Los diseños muestrales (muestreo) son los procedimientos utilizados para extraer muestras de una población. La forma en la que se extrae la muestra es muy importante, debido a que una mala muestra arrojará conclusiones equivocadas. La probabilidad proporciona las herramientas para realizar un muestreo justo; es decir, representativo de la población.

MUESTREO

Al recordar que la estadística es parte de las matemáticas que se encarga de obtener información y conclusiones acerca de una población tomando para ello datos de una muestra, deberá investigarse la mejor manera de seleccionar dichos datos, es decir, debe buscarse una técnica adecuada para realizar el muestreo, a lo que se llama *diseño del experimento*.

Siempre se desea que la muestra sea representativa de la población, para lo cual se debe tener una muestra aleatoria. Es claro que el término muestra aleatoria sugiere la forma en la que se deberán seleccionar los elementos de la muestra, es decir, en forma aleatoria, pero ¿qué se debe entender por muestra aleatoria en este momento? De una forma sencilla, puede decirse que se tiene una muestra aleatoria si todos los elementos de la población pudieron ser seleccionados.

Para estudiar con profundidad el diseño del experimento se requeriría de un curso especial, por lo cual, en este curso se considerará sólo el *muestreo aleatorio simple*, en el cual, todos los elementos de la población tienen la misma posibilidad de ser seleccionados.

Siempre se desea que la muestra sea representativa de la población, para lo cual se debe tener una muestra aleatoria. Es claro que el término muestra aleatoria sugiere la forma en la que se deberán seleccionar los elementos de la muestra, los cuales se deberán seleccionar en forma aleatoria.

Definición 1.1

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes, cada una con la misma distribución de probabilidad $f_X(x)$. Entonces se define $X_1, X_2, \ldots, X_n$ como la muestra aleatoria de tamaño n de la población $f_X(x)$ y tiene, por independencia, la distribución de probabilidad conjunta:

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

Definición 1.2

Si N y n son el número de elementos de la población y de la muestra, respectivamente, entonces si se hace el muestreo de tal manera que cada una de las $\binom{N}{n}$ muestras tiene la misma probabilidad de ser escogida, el muestreo se denomina aleatorio simple.

Se sabe que en la práctica es bastante difícil realizar un muestreo aleatorio simple; sin embargo, debe ponerse especial cuidado en que lo sea a fin de evitar sesgos. Un procedimiento sesgado es aquel en el cual se producen inferencias que en forma constante sobreestiman o subestiman alguna característica de la población.

En forma sencilla, en el muestreo aleatorio simple, las muestras aleatorias de igual tamaño, tienen la misma probabilidad de ser seleccionados, y por consiguiente, cada elemento de la población tiene la misma probabilidad de ser seleccionado.

Otros tipos de muestreo se mencionan a continuación:

El *muestreo sistemático* consiste en seleccionar una muestra tomando cada k-ésima unidad de la población un elemento. Las unidades elementales de la población se ordenan (o numeran) y cada k elementos se toma uno. La constante k recibe el nombre de razón de muestreo.

El *muestreo estratificado* consiste en dividir a la población en grupos denominados estratos. Los estratos contienen elementos homogéneos (de características similares). De cada estrato se toma un subconjunto mediante el muestreo aleatorio simple y la muestra total se obtiene uniendo los subconjuntos. El muestreo puede ser proporcional o no.

El *muestreo por conglomerados* es lo contrario del muestreo estratificado. Consiste en seleccionar conglomerados, los cuales se forman con elementos con las mayores diferencias posibles. Posterirmente se selecciona un conglomerado por muestro aleatorio simple de ahí se selecciona la muestra, nuevamente por muestreo aleatorio simple.

ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva es la parte de la estadística que tiene como propósito organizar y presentar los datos de una población o de una muestra para su análisis e interpretación. Es a partir de la estadística descriptiva, que la disciplina tomó el nombre de estadística, puesto que en la antigüedad, los gobernantes deseaban conocer la información de sus reinos.

En la estadística descriptiva existen básicamente tres técnicas:

- Distribución de Frecuencias
- Gráficas
- Medidas numéricas

Las técnicas no son independientes, por el contrario, deben complementarse. La distribución de frecuencias es la forma en la que se agrupan los datos cuando se tiene una cantidad considerable de ellos. Las gráficas sirven para visualizar rápidamente la forma en la que se agrupan los datos, y los parámetros numéricos son el resumen de los datos en forma cuantitativa. Cada técnica es una huella de la información que se estudia, pero no debe olvidarse que para realizar una mejor interpretación, deben combinarse las técnicas.

DISTRIBUCIÓN DE FRECUENCIAS

Es una técnica de agrupación usada en estadística cuando se tiene un conjunto muy grande de datos, de tal forma que el análisis posterior: gráficas y parámetros numéricos; se puede realizar de forma más rápida. La tabla de distribución de frecuencias puede usarse para datos **cuantitativos** y para datos **cualitativos**. Los datos cuantitativos son aquellos que indican cantidad, 5 personas, 2.5 Newtones, etc.; mientras que los datos cualitativos expresan cualidades: azul, alto, sano, etc.

Existe una gran diversidad de tablas de distribución de frecuencias; sin embargo, aquí se estudiará una tabla teórica completa. Para resumir los datos, se utilizan intervalos, clases o categorías y posteriormente se indica la frecuencia de cada uno de ellos. Las columnas que forman una tabla completa son:

Límites de clase. Son los valores menor y mayor que de encontrarse como datos en la muestra pertenecen a la clase en cuestión. Los límites de una clase tienen la misma aproximación que los datos en la muestra o de la población, esto es, si los datos son enteros, entonces los límites son enteros. Si los datos tienen aproximación a décimas, entonces los límites tendrán décimas, y así sucesivamente. Se denotan por L_i.

Fronteras de clase. Las fronteras o límites verdaderos de una clase, son los puntos medios entre los límites de intervalos consecutivos. Se denota por L_{R_i}, por límite real. No se acostumbra usar F porque esa letra se reserva para las frecuencias. La distancia entre la frontera inferior y el límite inferior de una misma clase, así como la existente entre el límite superior y la frontera superior de una misma clase es igual a media unidad de aproximación, esto es, si en una tabla de distribución de frecuencias, los límites de clase son: 1-3 , 4-6, 7-9 (con aproximación entera); entonces las fronteras serían 0.5-3.5, 3.5-6.5, 6.5-9.5 (con media unidad de aproximación entera, es decir, 0.5) ; como se muestra en la siguiente tabla (entre paréntesis se indica la operación que se realizó con los respectivos límites de clase para obtener las fronteras de clase).

Tabla 1.1. Límites y Fronteras de clase

Límites de clase	Fronteras de clase
1 - 3	0.5 *(1-0.5)* - 3.5 *(3+0.5)*
4 - 6	3.5 *(4-0.5)* - 6.5 *(6+0.5)*
7 - 9	6.5 *(7-0.5)* - 9.5 *(9+0.5)*
⋮	⋮

Marcas de clase. Es el punto medio de una clase. Se considera como el valor representativo de un intervalo. Las marcas de clase se obtienen promediando los límites de un intervalo, o bien, las fronteras. Se denota por x_i.

Tabla 1.2. Marcas de clase

Límites de clase	Fronteras de clase	Marca de clase, x_i
1 - 3	0.5 - 3.5	2
4 - 6	3.5 - 6.5	5
7 - 9	6.5 - 9.5	8
⋮	⋮	⋮

Frecuencia. Es el número de elementos en la muestra o en la población que pertenecen a la clase en cuestión. Se denota por f_i. Si los datos de una muestra son: 1, 9, 5, 8, 4, 1, 2, 7, 6, 3, 3, 2, 7, 9; entonces al agrupar por intervalos se obtienen las siguientes frecuencias.

Tabla 1.3. Frecuencia

Límites de clase	Fronteras de clase	Marca de clase, x_i	Frecuencia f_i
1 - 3	0.5 - 3.5	2	6
4 - 6	3.5 - 6.5	5	3
7 - 9	6.5 - 9.5	8	5
⋮	⋮	⋮	⋮

Frecuencia acumulada. Es el número de datos en la muestra o población, que son menores o iguales que el límite superior del intervalo en cuestión. Se denota por F_i, y se obtiene sumando la frecuencia del intervalo actual y de los anteriores intervalos.

Tabla 1.4. Frecuencia acumulada

Límites de clase	Fronteras de clase	Marca de clase, x_i	frecuencia f_i	Frecuencia acumulada F_i
1 - 3	0.5 - 3.5	2	6	6
4 - 6	3.5 - 6.5	5	3	6+3=9
7 - 9	6.5 - 9.5	8	5	6+3+5=14
⋮	⋮	⋮	⋮	⋮

Frecuencia Relativa. Es la proporción de datos que pertenecen a la clase en cuestión. Se denota por $f_i^{/}$ o por f_i^*. Es el cociente de la frecuencia entre el número total de datos, esto es: $f_i^* = \frac{f_i}{n}$. Para la tabla del ejemplo, si el total de datos es n=100, entonces:

Tabla 1.5. Frecuencia relativa

Límites de clase	Fronteras de clase	Marca de clase, x_i	frecuencia f_i	Frecuencia acumulada F_i	frecuencia relativa f_i^*
1 - 3	0.5 - 3.5	2	6	6	6/100=0.06
4 - 6	3.5 - 6.5	5	3	9	3/100=0.03
7 - 9	6.5 - 9.5	8	5	14	5/100=0.05
⋮	⋮	⋮	⋮	⋮	⋮

Frecuencia Acumulada Relativa. Es la proporción de los datos en la muestra o población que son menores o iguales que el límite superior de la clase en cuestión.

Se denota por $F_i^{/}$ o por F_i^*. Matemáticamente se define como el cociente de la frecuencia acumulada entre el número de datos, esto es: $F_i^* = \dfrac{F_i}{n}$. Considerando nuevamente que n=100, entonces:

Tabla 1.6. Tabla de distribución de frecuencias teórica completa

Límites de clase	Fronteras de clase	Marca de clase, x_i	frecuencia f_i	Frecuencia acumulada F_i	frecuencia relativa f_i^*	Frecuencia acumulada relativa F_i^*
1 - 3	0.5 - 3.5	2	6	6	0.06	0.06
4 - 6	3.5 - 6.5	5	3	9	0.03	0.09
7 - 9	6.5 - 9.5	8	5	14	0.05	0.014
⋮	⋮	⋮	⋮	⋮	⋮	

La tabla anterior, ya es una tabla de distribución de frecuencias teórica completa, pero debe observarse que para tener una tabla de distribución de frecuencias, basta con tener dos columnas, una que indique la clase (Límites, fronteras o marcas) y una que indique la frecuencia (Frecuencia, frecuencia acumulada, frecuencia relativa o frecuencia acumulada relativa).

Existen otras tablas de distribución de frecuencias, basadas en intervalos, en donde se utiliza la notación del Cálculo para los intervalos abiertos y cerrados, por ejemplo:

Tabla 1.7. Tabla de distribución de frecuencias con intervalos

Intervalo	Marca de clase, x_i	frecuencia f_i
[1 , 4)	2.5	2
[4 , 7)	5.5	5
[7 , 10)	8.5	8
⋮	⋮	⋮

En donde la marca de clase es nuevamente el punto medio del intervalo y no puede existir traslape en ningún intervalo.

Algunos valores que ayudan en la construcción de la tabla son:

Longitud de la clase. Es la diferencia entre la frontera superior y la inferior de una misma clase. Se denota por ***c***. Así para la tabla 1.6, la longitud del intervalo es: $c = 3$; lo mismo que para la tabla 1.7.

Cualquier tabla que contenga una columna de clase o intervalo y una columna de frecuencias, es una tabla de distribución de frecuencias; sin embargo, en este momento deberán de construirse tablas completas atendiendo a las siguientes recomendaciones.

Recomendaciones para la construcción de una tabla de frecuencias

- El número de clases estará entre 5 y 20, inclusive. La primera aproximación del número de clases se obtendrá con $\sqrt{n}$.

- Todas las clases serán de la misma longitud (***c***).

- La longitud de clase se aproxima mediante

  $$c = \frac{Rango}{número\ de\ clases}$$, donde el Rango es:

  $$Rango = Mayor\ valor\ en\ los\ datos\ -\ menor\ valor\ en\ los\ datos$$

  Posteriormente se ajusta de manera conveniente, de forma que el primer límite inferior sea ligeramente menor o igual que el menor valor, y el último límite superior sea ligeramente mayor o igual que el mayor dato.

- Tratará de evitarse que haya clases con frecuencia cero.

- La primera y la última clase nunca tendrán frecuencia cero.

Ejemplo 1.1
Los siguientes valores representan el tiempo diario de transporte de una muestra de 50 alumnos de cierta universidad al sur de Copilco.

69	56	73	66	64	44	36	69	76	53
79	72	82	77	71	48	49	49	60	67
73	70	64	56	31	62	56	55	51	45
30	40	80	49	59	60	76	67	30	72
45	43	77	49	46	42	63	41	64	79

Construir una tabla de distribución de frecuencias teórica completa.

Resolución
Puesto que no se proporciona ninguna indicación con respecto a los intervalos, se realiza la primera aproximación del número de intervalos con $\sqrt{n} = \sqrt{50} = 7.07$, por lo que se utilizarán 7 intervalos. El menor de los datos es 30 y el mayor de los datos es 82, por lo que el rango de los datos es $R = 82 - 30 = 52$, entonces la longitud del intervalo aproximada es $\frac{52}{7} = 7.428$, por lo que se utilizará una longitud de $c = 8$.
Puesto el menor valor es 30, se toma la decisión de iniciar en 29, teniéndose:

Tabla 1.8. Tabla de distribución del ejemplo 1.1

Límites de clase	Fronteras de clase	Marca de clase, x_i	Frecuencia f_i	Frecuencia acumulada F_i	Frecuencia relativa f_i^*	Frecuencia acumulada relativa F_i^*
29 - 36	28.5 -36.5	32.5	4	4	0.08	0.08
37 - 44	36.5-44.5	40.5	5	9	0.1	0.18
45 - 52	44.5 -52.5	48.5	9	18	0.18	0.36
53 - 60	52.5-60.5	56.5	8	26	0.16	0.52
61 - 68	60.5-68.5	64.5	8	34	0.16	0.68
69 - 76	68.5-76.5	72.5	10	44	0.2	0.88
77 - 84	76.5-84.5	80.5	6	50	0.12	1

GRÁFICAS

Cuando se desea dar un mayor impacto de la forma en la que se distribuyen los datos, éstos se presentan en una o varias gráficas. Son muchas las gráficas que se pueden utilizar en la estadística descriptiva, destacando el histograma, el polígono y la ojiva. Otras gráficas usadas son la de sectores circulares (también llamadas pastel o *pie*), la de tallos y hojas y el diagrama de caja (que se estudiará después de las medidas numéricas).

HISTOGRAMA DE FRECUENCIAS

El histograma es una gráfica de barras rectangulares cuyas bases están centradas en la marca de clase del intervalo, y sus áreas proporcionales a la frecuencia del intervalo. Es evidente que para un histograma bien construido, las distancias entre marcas de clase son siempre las mismas, por lo que la condición de que las áreas de los rectángulos sean proporcionales a las frecuencias, se convierte en la altura proporcional a la frecuencia. No es necesario dibujar el eje de las ordenadas; sin embargo, puede hacerse sin ningún conflicto.

Con los datos de la tabla 1.8, el histograma de frecuencias es:

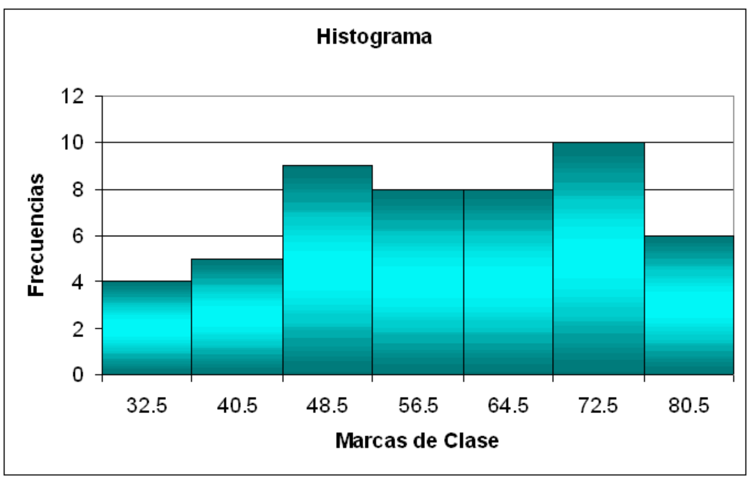


Fig. 1.1. Histograma con frecuencias en el eje

O bien, las frecuencias pueden colocarse sobre los rectángulos o dentro de ellos.

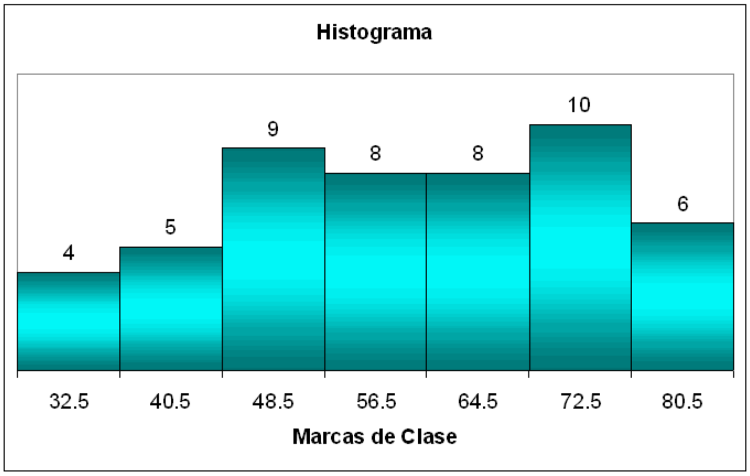


Fig. 1.2. Histograma con frecuencias en los rectángulos

POLÍGONO DE FRECUENCIAS

El polígono de frecuencias es una gráfica poligonal o de líneas rectas que indica para cada marca de clase la frecuencia. Se obtiene uniendo los puntos medios de las partes superiores de las barras del histograma. Para que la línea no se dibuje "flotando", se puede dibujar una marca de clase antes de la primera y una marca de clase posterior a la última, cada una con frecuencia cero, de esta forma la gráfica poligonal parte del eje de las abscisas y termina en él.

Tanto el histograma como la ojiva, se dibujan generalmente con las frecuencias absolutas, pero también pueden dibujarse con las frecuencias relativas, con las acumuladas o con las acumuladas relativas.

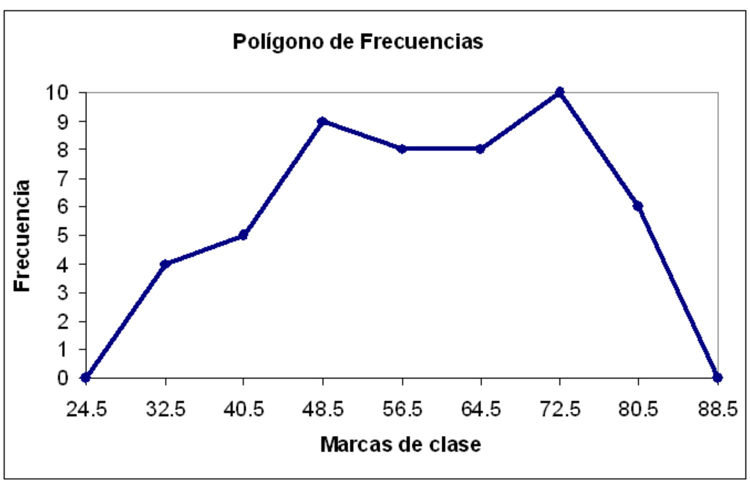


Fig. 1.3. Polígono de Frecuencias

OJIVA

La ojiva es también una gráfica poligonal, pero se dibuja utilizando las fronteras contra las frecuencias acumuladas (o acumuladas relativas). La ojiva indica, para cada frontera, los elementos (o proporción de elementos), que son menores o iguales que dicha frontera. Si se utiliza la frecuencia acumulada relativa se llama ojiva porcentual. A la ojiva también se le llama en ocasiones *polígono de frecuencias acumuladas*

La ojiva para los datos del ejemplo 1.1 se muestra en la siguiente figura.

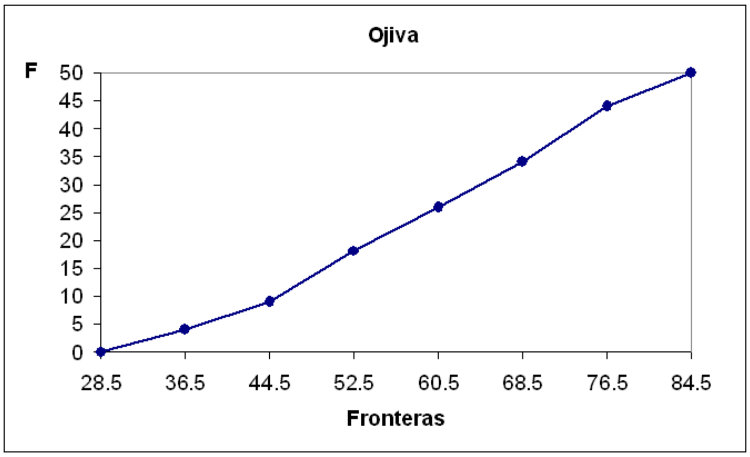


Fig. 1.4. Ojiva

Con la tabla de distribución de frecuencias y con las gráficas, se describe el comportamiento de un conjunto de datos; sin embargo, para no caer en subjetividades, o errores por la escala, se utilizan también las medidas numéricas.

DIAGRAMA DE TALLO Y HOJAS

El diagrama de tallo y hojas es una representación semi-gráfica, utilizada para datos cuantitativos, utilizada ampliamente en la década de los ochenta, cuando las computadoras no realizaban gráficas más sofisticadas como los histogramas. Consiste en separar los números en dos partes, por ejemplo decenas y unidades, de esta forma los números: 21, 33, 22, 37, 41, 9, 12, 38, 28, 15, 25; se ordenan para tener 9, 12, 15, 21, 22, 25, 28, 33, 37, 38,41; y posteriormente se separan en decenas y unidades, para formar el gráfico.

0 | 9
1 | 2 5
2 | 1 2 5 8
3 | 3 7 8
4 | 1

Este diagrama se recomienda cuando se tienen entre 20 y 40 datos; y tiene la ventaja de que muestra con cierta facilidad la forma de la distribución. La separación de los datos puede hacerse de distintas formas, por ejemplo los números de tres dígitos pueden separarse en dos para el tallo y uno para las hojas.

MEDIDAS NUMÉRICAS

Las medidas numéricas, por el tipo de información que proporcionan se clasifican en medidas de tendencia central, medidas de dispersión y medidas de forma.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son valores representativos de un conjunto de datos, que se sitúan en la parte central de los mismos. Las medidas de tendencia central más conocidas son la media, la mediana y la moda.

Media

Media aritmética

La media aritmética es más conocida simplemente como media, y es el promedio de un conjunto de valores. Es sin duda la medida de tendencia central más utilizada, y por lo general es la más representativa. Se denota por $\bar{x}$.

$$\bar{x} = \begin{cases} \dfrac{1}{n} \sum_{i=1}^{n} x_i & ; \ \textit{Datos no agrupados} \\ \dfrac{1}{n} \sum_{i=1}^{m} x_i f_i & ; \ \textit{Datos agrupados} \end{cases}$$

Debe observarse que para datos agrupados la suma va desde *1* hasta *m*, donde ***m* es el número de intervalos**, y x_i y f_i son la marca de clase y la frecuencia del intervalo, respectivamente.

Para el ejemplo 1.1, se pueden obtener las medias de los datos sin agrupar y agrupados, teniéndose:
Para datos sin agrupar:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{50} (54 + 70 + 65 + . . . + 66) = 58.7$$

Para datos agrupados se tiene

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{m} x_i f_i=\frac{1}{50}\left[\,32.5(4)\;+\;40.5(5)+\ldots+80.5(6)\,\right]=58.9$$

En resumen:

$$\bar{x}\;=\begin{cases}58.7 & ;\;\; \textit{Datos no agrupados}\\ 58.9 & ;\;\; \textit{Datos agrupados}\end{cases}$$

y debe observarse que las medias obtenidas son muy parecidas, pero en lo general diferentes.

<u>Media geométrica</u>

La media geométrica de un conjunto de valores positivos se calcula con la raíz *n*-ésima del producto de las *n* observaciones. Se denota por G.

$$G\;=\begin{cases}\sqrt[n]{x_1\;x_2\;.\;.\;.\;x_n} & ;\;\; \textit{Datos no agrupados}\\ \sqrt[n]{x_1^{f_1}\;x_2^{f_2}\;.\;.\;.\;x_m^{f_m}} & ;\;\; \textit{Datos agrupados}\end{cases}$$

Para los datos del ejemplo 1.1 se tiene:

$$G\;=\begin{cases}59.7967 & ;\;\; \textit{Datos no agrupados}\\ 56.9836 & ;\;\; \textit{Datos agrupados}\end{cases}$$

<u>Media armónica</u>

La media armónica de un conjunto de datos se denota por H, y es el recíproco de la media aritmética de los recíprocos de cada uno de los valores.

$$H\;=\begin{cases}\dfrac{1}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{1}{x_i}}\;=\;\dfrac{n}{\displaystyle\sum_{i=1}^{n}\dfrac{1}{x_i}} & ;\;\; \textit{Datos no agrupados}\\[2ex] \dfrac{1}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{m}\dfrac{f_i}{x_i}}\;=\;\dfrac{n}{\displaystyle\sum_{i=1}^{m}\dfrac{f_i}{x_i}} & ;\;\; \textit{Datos agrupados}\end{cases}$$

Para los datos del ejemplo 1.1 se tiene:

$$H = \begin{cases} 54.7191 & ; \; \textit{Datos no agrupados} \\ 54.9336 & ; \; \textit{Datos agrupados} \end{cases}$$

Las medias aritmética, geométrica y armónica, para un conjunto de valores positivos están relacionadas mediante

$$\boxed{H \le G \le \bar{x}}.$$

Con la hoja de cálculo Excel, es muy fácil calcular las medias para datos sin agrupar utilizando los comandos *promedio(Rango de celdas), media.geom(Rango de celdas)* y *media.armo(Rango de valores)*. Para datos agrupados deben utilizarse el comando *sumaproducto(Rango de celdas 1, Rango de celdas 2)*, para facilitar las operaciones.

Mediana

La mediana de un conjunto de datos ordenados, es el valor que divide al conjunto en dos conjuntos de igual tamaño, o bien, es el promedio de los dos valores centrales. Se denota por $\tilde{x}$.

Cuando los datos no están agrupados, se deben ordenar en forma ascendente o descendente y seleccionar el valor central. Si los datos son pares, entonces se toma el promedio de los dos valores centrales; si los datos son impares entonces se toma el dato central.

Cuando los datos están agrupados, entonces se realiza una interpolación lineal utilizando las fronteras y la frecuencia acumulada (es decir, los datos de la ojiva), para encontrar el valor de x en el cual la frecuencia acumulada es de $\frac{n}{2}$.

Con los datos del ejemplo 1.1, y al ordenar los valores se tiene:

30, 30, 31, 36, 40, 41, 42, 43, 44, 45,
45, 46, 48, 49, 49, 49, 49, 51, 53, 55,
56, 56, 56, 59, $\boxed{60, 60}$, 62, 63, 64, 64,
64, 66, 67, 67, 69, 69, 70, 71, 72, 72,
73, 73, 76, 76, 77, 77, 79, 79, 80, 82.

puesto que el número de datos es par, se toman los 2 valores centrales (60 y 60), y de ellos se obtiene el promedio, finalmente, $\tilde{x} = 60$.

Con la distribución de frecuencias obtenida en el ejemplo 1.1, se utilizan las columnas de fronteras y de frecuencia acumulada

Fronteras de clase	Frecuencia acumulada F_i
28.5 - 36.5	4
36.5 - 44.5	9
44.5 - 52.5	18
52.5 - 60.5	26
60.5 - 68.5	34
68.5 - 76.5	44
76.5 - 84.5	50

y se realiza una interpolación para obtener el valor de x, para el cual la frecuencia acumulada sea de $\frac{n}{2} = \frac{50}{2} = 25$.

Frontera	Frecuencia acumulada
52.5	18
$\tilde{x}$	$\leftarrow$ 25
60.5	26

Interpolando se obtiene: $\tilde{x} = 59.5$.

Finalmente, se tiene, para los datos del ejemplo 1.1

$$\tilde{x} = \begin{cases} 60 & ; \ Datos \ no \ agrupados \\ 59.5 & ; \ Datos \ agrupados \end{cases}$$

Moda

La moda de un conjunto de datos es el valor que se repite con mayor frecuencia. Se denota por x_{mo}, o en ocasiones mo. Si existe más de una moda, entonces se dice que los datos tienen distribución bimodal.

Para datos sin agrupar, se deben contar las repeticiones que puedan existir, y el que se repita mayor número de veces será la moda. Si todos los datos aparecen el mismo número de veces, entonces se dice que no existe moda.

Para datos agrupados, la moda se aproxima con la marca de clase del intervalo con mayor frecuencia, o bien, utilizando la fórmula:

$$x_{mo} = Frontera\ inferior + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$$

Donde:

$Frontera\ inferior \triangleq$ Es la frontera inferior del intervalo con mayor frecuencia.
$\Delta_1 \triangleq$ Es el exceso de la frecuencia modal sobre la frecuencia de la clase inmediata anterior.
$\Delta_2 \triangleq$ Es el exceso de la frecuencia modal sobre la frecuencia de la clase inmediata posterior.
$c \triangleq$ Es la longitud de la clase.

Para los datos del ejemplo 1.1, sin agrupar, el valor que más se repite es el 49, con 4 repeticiones, por lo que la moda es 49, esto es: $x_{mo} = 49$.

Para los datos agrupados la moda puede obtenerse con la marca de clase del intervalo modal, por lo que se obtiene: $x_{mo} = 72.5$. Debe observarse que 72.5 es la marca de clase del intervalo con límites 69-76, y con la máxima frecuencia observada, $f = 10$.

Utilizando la fórmula para la moda se tiene:

$$x_{mo} = 68.5 + \left(\frac{2}{2 + 4} \right) (8)$$

$$x_{mo} = 71.1667$$

Relación entre la media, la mediana y la moda
Para un conjunto de datos con distribución de frecuencia unimodal y poca simetría se tiene la siguiente relación empírica:

$$\bar{x} - x_{mo} = 3 (\bar{x} - \tilde{x})$$

Si la distribución es simétrica y unimodal, entonces se tiene la siguiente relación:

$$\bar{x} = \tilde{x} = x_{mo}$$

Cuartiles, Deciles y Percentiles

Así como la mediana es el valor que divide a una conjunto de datos ordenados en dos conjuntos de igual tamaño, los datos pueden dividirse en cuatro conjuntos de igual tamaño (cuartiles), en 10 conjuntos de igual tamaño (deciles) y en 100 conjuntos de igual tamaño (percentiles).

Los cuartiles se denotan generalmente por Q_1, Q_2 y Q_3, y el segundo cuartil coincide con la mediana. Los deciles se denotan $D_1, D_2, \ldots, D_9$, y el quinto decil D_5 coincide con la mediana y con el segundo cuartil. Los percentiles se denotan $P_{01}, P_{02}, \ldots, P_{99}$, el El percentil 50 coincide con la mediana, con el segundo cuartil y con el quinto decil. El percentil 10 coincide con el primer decil, y así se pueden encontrar muchas otras relaciones.

MEDIDAS DE DISPERSIÓN

Las medidas de dispersión proporcionan un indicador del alejamiento de los datos. También se les llama medidas de variación. Las medidas más comunes son: Rango, desviación media, variancia, desviación estándar, el rango semi-intercuartil y el coeficiente de variación.

Rango

El rango de un conjunto de datos es la diferencia entre el mayor valor menos el menor valor. Se denota por $Rango$, o simplemente por R. Para datos agrupados se utilizan los límites mayor y menor. Es común no realizar la operación de resta y solamente indicarla.

Para los datos del ejemplo 1.1 se tienen los siguientes resultados.

$$R = \begin{cases} 82 - 30 & ; \textit{ Datos no agrupados} \\ 84 - 29 & ; \textit{ Datos agrupados} \end{cases}$$

En Excel se pueden utilizar los comandos *max(Rango de celdas)* y *min(Rango de celdas)* para obtener los valores mayor y menor de un conjunto de datos.

Desviación Media

La desviación media o desviación promedio de un conjunto de datos es el promedio de las distancias de cada valor con respecto a la media. Se denota por DM.

$$DM = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left| x_i - \bar{x} \right| & ; \ \textit{Datos no agrupados} \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{m} \left| x_i - \bar{x} \right| f_i & ; \ \textit{Datos agrupados} \end{cases}$$

Para los datos del ejemplo 1.1 se tiene:
Para datos no agrupados se utiliza la media de datos no agrupados, que es, $\bar{x} = 58.7$, por lo que se tiene:

$$DM = \frac{1}{50} \sum_{i=1}^{50} \left| x_i - 58.7 \right|$$

$$DM = \frac{1}{50} \left(\left| 54 - 58.7 \right| + \left| 70 - 58.7 \right| + \ . \ . \ . \ + \left| 66 - 58.7 \right| \right) = 12.244$$

y para datos agrupados, se utiliza la media de datos agrupados, las marcas de clase y la frecuencia, por lo que:

$$DM = \frac{1}{50} \sum_{i=1}^{7} \left| x_i - 58.9 \right| f_i$$

$$DM = \frac{1}{50} \left(\left| 32.5 - 58.9 \right| (4) + \left| 40.5 - 58.9 \right| (5) + . . . + \left| 80.5 - 58.9 \right| (6) \right) = 12.416$$

En resumen

$$DM = \begin{cases} 12.244 & ; \ \textit{Datos no agrupados} \\ 12.416 & ; \ \textit{Datos agrupados} \end{cases}$$

Desviación Mediana

Una variación de la desviación media es la *desviación mediana*, la cual consiste en tomar el promedio de las distancias con respecto a la mediana. Se denota DMd.

$$DMd = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left| x_i - \tilde{x} \right| & ; \ \textit{Datos no agrupados} \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{m} \left| x_i - \tilde{x} \right| f_i & ; \ \textit{Datos agrupados} \end{cases}$$

Tanto la desviación media como la desviación mediana son poco utilizadas en la práctica por lo difícil de manejar el valor absoluto. Para eliminar el signo de las diferencias $x_i - \bar{x}$ y evitar el cálculo del valor absoluto, se define la *variancia* o *varianza* de un conjunto de datos utilizando el cuadrado de la diferencia.

No puede utilizarse solamente la suma de desviaciones, porque ésta da como resultado siempre cero, esto es:

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

es por eso que se obtiene el valor absoluto en el caso de la desviación media y que se eleva al cuadrado para la variancia.

Varianza o Variancia

La variancia de un conjunto de datos es el promedio de las distancias cuadradas de cada valor con respecto a su media. Se denota por s_n^2 o por s_{n-1}^2 dependiendo del valor que se utilice para promediar. Se divide entre n (se promedia entre n), cuando se considera que se tienen todos lo datos posibles (población), y se divide entre $n-1$ cuando se tiene solo una fracción de los datos (muestra). La fórmula para la variancia es:

$$s_n^2 = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 & ; \; \textit{Datos no agrupados} \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2 f_i & ; \; \textit{Datos agrupados} \end{cases}$$

o bien:

$$s_{n-1}^2 = \begin{cases} \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 & ; \; \textit{Datos no agrupados} \\ \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2 f_i & ; \; \textit{Datos agrupados} \end{cases}$$

En el ejemplo 1.1, se tiene una muestra de 50 alumnos, por lo que se debe de obtener s_{n-1}^2, aunque para valores mayores o iguales que 30 el resultado es muy parecido, y por eso algunos autores siguieren utilizar s_n^2, cuando $n \geq 30$. Con los datos del ejemplo se tiene:

$$s_{n-1}^2 = \begin{cases} 206.94898 & ; \; \textit{Datos no agrupados} \\ 212.2449 & ; \; \textit{Datos agrupados} \end{cases}$$

El inconveniente de utilizar la variancia como medida de dispersión se encuentra en sus unidades, puesto que queda en unidades cuadradas. Para evitar esta complejidad, la medida de dispersión más utilizada es la *desviación estándar*.

Desviación estándar

La desviación estándar de un conjunto de datos es la raíz cuadrada de la variancia. Se denota por s_{n-1} o por s_n, dependiendo de si se obtiene la desviación estándar de una muestra o de toda la población.

Es claro que para calcular la desviación estándar debe calcularse la variancia primero, de forma que:

$$s_{n-1} = \sqrt{s_{n-1}^2}$$

$$s_n = \sqrt{s_n^2}$$

Cuando los datos agrupados tienen una distribución que se aproxima a una campana, entonces la desviación estándar puede aproximarse mediante el rango dividido entre 4, esto es:

$$\boxed{s \approx \frac{R}{4}}$$

Para los datos del ejemplo 1.1 se tiene:

$$s_{n-1} = \begin{cases} 14.3857 & ; \quad \textit{Datos no agrupados} \\ 14.5686 & ; \quad \textit{Datos agrupados} \end{cases}$$

Coeficiente de Variación

El coeficiente de variación de un conjunto de datos es una medida de la dispersión en relación con la media de los datos, no tiene unidades y se define mediante el cociente de la desviación estándar entre la media., cuando la media es positiva.

$$cv = \frac{s}{\bar{x}}$$

se utiliza la desviación estándar adecuada, s_{n-1} o s_n, y puede utilizarse un subíndice en el coeficiente de variación para reconocer entre qué se promedió para obtener la desviación estándar.

Para los datos del ejemplo 1.1 se tiene:

$$cv_{n-1} = \begin{cases} 0.2451 & ; \quad Datos\ \ no\ \ agrupados \\ 0.2473 & ; \quad Datos\ \ agrupados \end{cases}$$

El coeficiente de variación tiene su principal uso al comparar muestras y también se utiliza para medir la homogeneidad de los datos. Si $cv \leq 0.15$ indica datos homogéneos, mientras que $cv > 0.15$ se tienen datos heterogéneos.

Rango intercuartílico

El rango intercuartílico de un conjunto de datos es la diferencia entre el tercer y el primer cuartil.

$$R_Q = Q_3 - Q_1$$

Rango semi-intercuartílico

El Rango semi-intercuartílico es el promedio del rango intercuartílico, esto es:

$$Q = \frac{Q_3 - Q_1}{2}$$

Rango Percentil

El rango percentil de un conjunto de datos, es la diferencia entre el percentil 90 y el percentil 10, esto es:

$$R_P = P_{90} - P_{10}$$

MEDIDAS DE FORMA

Las medidas de forma de un conjunto de datos son el sesgo y la curtosis. Para poder definir a las medidas de forma, es necesario definir primero los momentos.

Momentos con respecto al origen.

El r-ésimo momento con respecto al origen se define mediante:

$$m_r' = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^r & ; \textit{ Datos no agrupados} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{m} x_i^r f_i & ; \textit{ Datos agrupados} \end{cases}$$

Y el r-ésimo momento con respecto a la media se define mediante:

$$m_r = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^r & ; \textit{ Datos no agrupados} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{m} \left(x_i - \bar{x}\right)^r f_i & ; \textit{ Datos agrupados} \end{cases}$$

Debe observarse que el primer momento con respecto al origen m_1' es la media $\bar{x}$, mientras que el segundo momento con respecto a la media m_2 es la variancia s_n^2. Los momentos con respecto a la media pueden calcularse con momentos con respecto al origen, al desarrollar las sumas. Las primeras relaciones son:

$$m_2 = m_2' - m_1'^{\,2}$$

$$m_3 = m_3' - 3\,m_1'\,m_2' + 2\,m_1'^{\,3}$$

$$m_4 = m_4' - 4\,m_1'\,m_3' + 6\,m_1'^{\,2}\,m_2' - 3\,m_1'^{\,4}$$

Sesgo

El sesgo de un conjunto de datos es una medida del grado de simetría (o asimetría) de los datos. Se denota por a_3 o por α_3, y se define mediante:

$$a_3 = \frac{m_3}{s^3}$$

donde m_3 es el tercer momento con respecto a la media y s es la desviación estándar.

El sesgo se compara con cero. Cuando el coeficiente de sesgo es menor que cero se dice que los datos tienen una distribución sesgada a la izquierda o con sesgo negativo. Cuando el coeficiente de sesgo es positivo, se dice que los datos tienen una distribución sesgada a la derecha o con sesgo positivo. Si el coeficiente de sesgo es cero, entonces los datos tienen una distribución simétrica o insesgada.

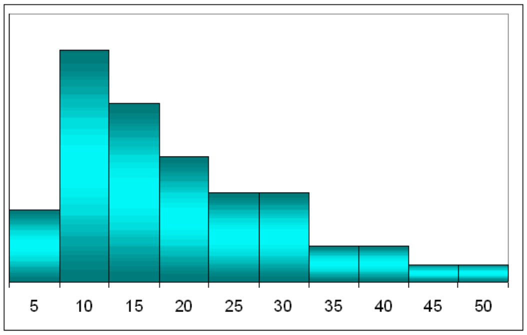


Fig. 1.5. Sesgo positivo

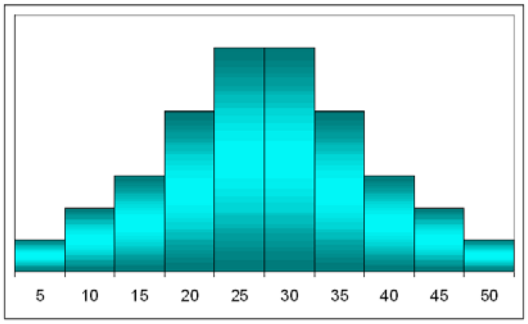


Fig. 1.6. Distribución simétrica

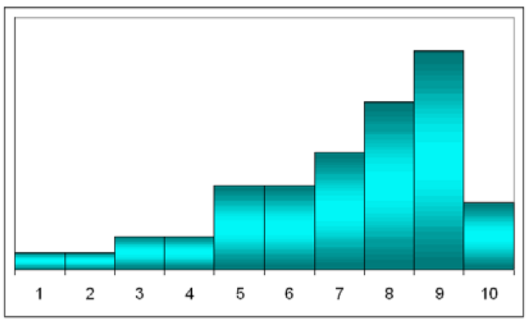


Fig. 1.7. Sesgo negativo

Debe observarse la relación de la medidas de tendencia central: media, mediana y moda, con el signo del sesgo cuando los datos se encuentran agrupados.

Sesgo positivo: $x_{mo} < \tilde{x} < \bar{x}$

Sesgo negativo: $\bar{x} < \tilde{x} < x_{mo}$

Insesgado: $x_{mo} = \tilde{x} = \bar{x}$

Para los datos de ejemplo 1.1 se tiene, para datos sin agrupar:

$$a_3 = \frac{m_3}{s^3} = \frac{-740.784}{(14.3857)^3} = -0.2488$$

y para los datos agrupados se tiene

$$a_3 = \frac{m_3}{s^3} = \frac{-559.104}{(14.5686)^3} = -0.1808$$

En resumen:

$$a_3 = \begin{cases} -0.2488 \;\; ; \;\; \textit{Datos no agrupados} \\ -0.1808 \;\; ; \;\; \textit{Datos agrupados} \end{cases}$$

Al utilizar la relación de las medidas de tendencia central con los datos agrupados del ejemplo 1.1, se observa que $\bar{x} \le \tilde{x} \le x_{mo}$, puesto que $\bar{x} = 58.9 \le \tilde{x} = 59.5 \le x_{mo} = 72.5$, por lo que se tiene un sesgo negativo, como ya se había calculado.

En la práctica cuando se requiere saber el signo del sesgo, pero no es determinante su magnitud, basta con realizar la comparación de las medidas de tendencia central.

Curtosis

El coeficiente de curtosis de un conjunto de datos mide el grado de aplanamiento relativo de la distribución de los datos. Se denota mediante a_4, o bien, α_4. Se define mediante la expresión:

$$a_4 = \frac{m_4}{s^4}$$

donde m_4 es el cuarto momento con respecto a la media y s es la desviación estándar.

La curtosis se compara contra tres, porque tres es la curtosis de la distribución normal, que se estudiará en el tema 4 y es ampliamente utilizada en la probabilidad y la estadística.

Si los datos tienen una distribución más puntiaguda que la distribución normal $a_4 > 3$, entonces se dice que los datos tienen una distribución *leptocúrtica*. Si los datos tienen una distribución como la normal, $a_4 = 3$, entonces se dice que la distribución es *mesocúrtica*. Si los datos tienen una distribución aplanada,

$a_4 < 3$, entonces se les llama *platicúrticos*.

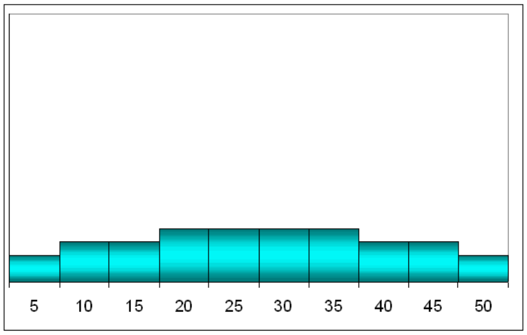


Fig. 1.8. Distribución platicúrtica

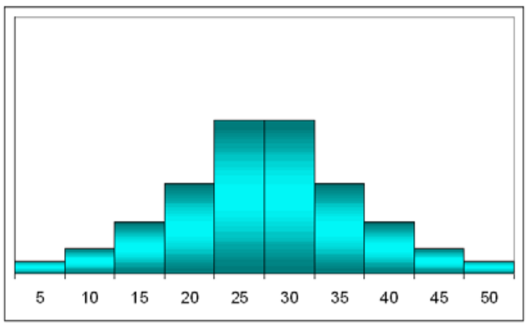


Fig. 1.9. Distribución mesocúrtica

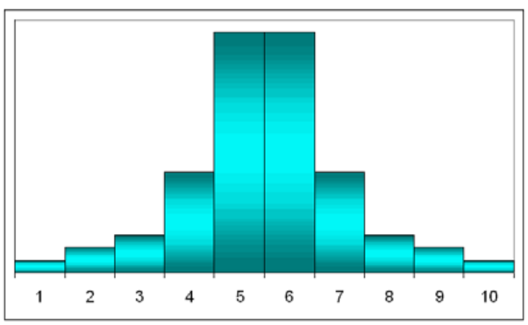


Fig. 1.10. Distribución leptocúrtica

Con los valores del ejemplo 1.1 se tiene, para datos sin agrupar:

$$a_4 = \frac{m_4}{s^4} = \frac{84669.8737}{(14.3857)^4} = 1.97698$$

y para datos agrupados:

$$a_4 = \frac{m_4}{s^4} = \frac{85554.3808}{(14.5686)^4} = 1.8992$$

en resumen:

$$a_4 = \begin{cases} 1.9769 & ; \ \textit{Datos no agrupados} \\ 1.8992 & ; \ \textit{Datos agrupados} \end{cases}$$

Puede observarse que los datos del ejemplo 1.1 tienen una distribución mesocúrtica.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 1.2

Los datos siguientes señalan el tiempo de funcionamiento (en días) hasta que se presenta la primera falla de $n = 88$ radio transmisores-receptores:

16	224	16	80	96	536	400	80
392	576	128	56	656	224	40	32
358	384	256	246	328	464	448	716
304	16	72	8	80	72	56	608
108	194	136	224	80	16	424	264
156	216	168	184	552	72	184	240
438	120	308	32	272	152	328	480
60	208	340	104	72	168	40	152
360	232	40	112	112	288	168	352
56	72	64	40	184	264	96	224
168	168	114	280	152	208	160	176

a) Utilizar el rango para aproximar el valor de la desviación estándar.
b) Obtener una distribución de frecuencias con 15 clases de longitud 50 comenzando con 0.5

Resolución

a) Utilizando la fórmula par aproximar el rango se tiene:

$$s \approx \frac{R}{4} = \frac{716 - 8}{4} = 177$$

b) La tabla queda:

Límites	Fronteras	Marcas de clase	Frecuencia	Frecuencia relativa
1 - 50	0.5 - 50.5	25.5	11	0.125
51 - 100	50.5 - 100.5	75.5	16	0.182

S))))))))))))))))))))))))))))))))))))))

101 - 150	100.5 - 150.5	125.5	8	0.091
151 - 200	150.5 - 200.5	175.5	15	0.170
201 - 250	200.5 - 250.5	225.5	10	0.114
251 - 300	250.5 - 300.5	275.5	6	0.068
301 - 350	300.5 - 350.5	325.5	5	0.057
351 - 400	350.5 - 400.5	375.5	6	0.068
401 - 450	400.5 - 450.5	425.5	3	0.034
451 - 500	450.5 - 500.5	475.5	2	0.023
501 - 550	500.5 - 550.5	525.5	1	0.011
551 - 600	550.5 - 600.5	575.5	2	0.023
601 - 650	600.5 - 650.5	625.5	1	0.011
651 - 700	650.5 - 700.5	675.5	1	0.011
701 - 750	700.5 - 750.5	725.5	1	0.011

Debe observarse que el valor en el que se inicia la tabla es una frontera, puesto que los datos no tiene el valor 0.5.

S))

Ejemplo 1.3

De los resultados en un examen de antecedentes de probabilidad, aplicado a los alumnos que cursan estadística, se obtuvo la siguiente tabla de distribución de frecuencias

Calificación	*Frecuencia*
$[0, 2)$	37
$[2, 4)$	198
$[4, 6)$	138
$[6, 8)$	31
$[8, 10]$	7

Obtener:

a) La media, la mediana y la moda.
b) La variancia.
c) Con los resultados obtenidos en el inciso (a), indicar si la distribución de las calificaciones tiene un sesgo positivo, negativo o no tiene sesgo.

Resolución

a) La media es $\bar{x} = \frac{1}{n} \sum_{i=1}^{m} x_i f_i$

$$\bar{x} = \frac{1}{411}\left[1(37) + 3(198) + 5(138) + 7(31) + 9(7)\right]$$

$$= 3.895$$

La mediana se obtiene mediante interpolación, por lo que se tiene:

$\tilde{x} = 3.7$

La moda se calcula con la expresión

$$x_{mo} = \textit{Frontera inferior} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c$$

de donde

$$x_{mo} = 2 + \left[\frac{(198 - 37)}{(198 - 37) + (198 - 138)}\right](2)$$

$$= 2 + \left[\frac{161}{161 + 60}\right](2)$$

$$= 3.45$$

o bien, se puede aproximar con la marca de clase del intervalo modal, con lo que

$x_{mo} = 3$

b) La variancia está dada por

$$s^2 = \frac{1}{n}\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i$$

por lo que

$$s^2 = \frac{1}{411}\Big[(1 - 3.895)^2(37) + (3 - 3.895)^2(198) + \;. \;. \;.$$
$$+ (9 - 3.895)^2(7)\Big]$$

$$= 2.7214$$

c) Puesto que la $\textit{media} > \textit{mediana} > \textit{moda}$ se tiene un sesgo positivo.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 1.4

En la siguiente tabla, se tienen los tiempos medidos en horas con un decimal que necesitó un transbordador para cruzar de la Ciudad de Mazatlán a La Paz, en 60 viajes sucesivos.

8.7	8.4	9.3	8.7	8.3	9.0
9.2	8.2	8.6	8.8	9.0	8.6
9.4	8.3	8.2	8.3	9.1	8.9
8.5	8.7	8.5	9.5	8.4	8.6

8.6	8.7	9.2	8.5	8.1	9.8
8.9	9.6	8.8	8.6	8.2	8.5
8.6	8.8	8.8	8.7	8.7	8.5
9.0	8.5	8.9	9.3	8.3	8.7
9.2	8.5	8.6	8.5	9.1	8.5
9.0	8.7	9.2	9.0	8.4	8.9

a) Construir una tabla de frecuencias de la duración de los viajes, con 6 intervalos.
b) Dibujar el histograma de frecuencias relativas.
c) Calcular la media.
d) Calcular la mediana.
e) Calcular la moda.
f) Calcular la desviación estándar.
g) Investigar si la distribución empírica es o no simétrica.
h) Clasificar la distribución empírica por su grado de aplanamiento.

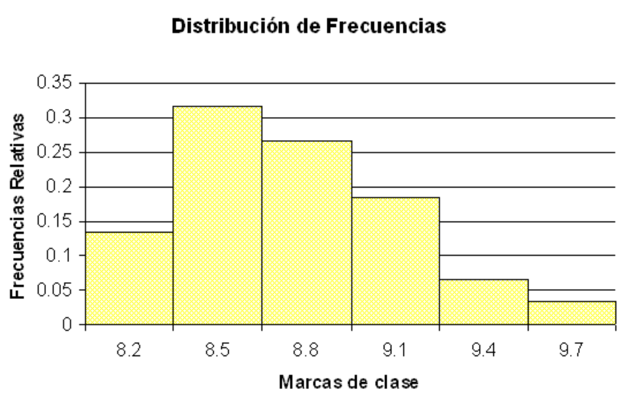
Resolución

a) El rango de los datos es: 9.8 - 8.1 = 1.7

Dividiendo en 6 intervalos, con $c = 0.3$ y comenzando en 8.05, se tiene:

Límites	Fronteras	Marca de clase x_i	frecuencia f_i	frecuencia relativa f'_i	Frec. relat. acum. $\left(F'_i\right)$
8.1 - 8.3	8.05 - 8.35	8.2	8	0.1333	0.1333
8.4 - 8.6	8.35 - 8.65	8.5	19	0.3166	0.4499
8.7 - 8.9	8.65 - 8.95	8.8	16	0.2666	0.7165
9.0 - 9.2	8.95 - 9.25	9.1	11	0.1833	0.8998
9.3 - 9.5	9.25 - 9.55	9.4	4	0.0666	0.9664
9.6 - 9.8	9.55 - 9.85	9.7	2	0.0334	1

b)



c) Datos sin agrupar: $\bar{x} = \dfrac{1}{60} \sum_{i=1}^{60} x_i = 8.7516$

Datos agrupados: $\bar{x} = \dfrac{1}{60} \sum_{i=1}^{6} x_i f_i = 8.75$

d) Datos sin agrupar: $\tilde{x} = 8.7$

Datos agrupados (interpolando):

8.65	27
$\tilde{x}$	30
8.95	43

De donde $\tilde{x} = 8.7$

e) Datos sin agrupar: $m_o = 8.5$

Datos agrupados: $m_o = 8.7$

f) Datos sin agrupar: $s_{n-1} = \sqrt{\dfrac{1}{59} \sum_{i=1}^{60} \left(x_i - \bar{x}\right)^2} = 0.3712$

Datos agrupados: $s_{n-1} = \sqrt{\dfrac{1}{59} \sum_{i=1}^{6} \left(x_i - \bar{x}\right)^2 f_i} = 0.3793$

g) Para determinar si la distribución es simétrica o no, se calcula el sesgo.

Datos sin agrupar: $a_3 = \dfrac{m_3}{s_{n-1}^3} = \dfrac{0.03117668}{(0.3712)^3} = 0.61$

Datos agrupados: $a_3 = \dfrac{m_3}{s_{n-1}^3} = \dfrac{0.02765}{(0.3793)^3} = 0.507$

La distribución tiene un ligero sesgo positivo.

h) Para determinar el aplanamiento se calcula la curtosis.

Para datos sin agrupar: $a_4 = \dfrac{m_4}{s^4} = \dfrac{0.051792411}{(0.3712)^4} = 2.73$

Ligeramente platicúrtica.

Para datos agrupados: $a_4 = \dfrac{m_4}{s^4} = \dfrac{0.05524125}{(0.3793)^4} = 2.67$

Ligeramente platicúrtica.

S))))))))))))))))))))))))))))))))))))

Ejemplo 1.5

Determinar cómo se relacionan la media y la mediana muestrales de las x_i con las y_i para cada uno de los siguientes casos.

a) Si se agrega una constante c a cada una de las x_i en una muestra, dando $y_i = x_i + c$.

b) Si cada x_i se multiplica por una constante c, dando $y_i = c\,x_i$.

Resolución

a) Para la media

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (x_i + c)}{n}$$

$$= \bar{x} + c$$

Para la mediana

$$\tilde{y} = \tilde{x} + c$$

b) Para la media

$$\bar{y} = c\,\bar{x}$$

Para la mediana

$$\tilde{y} = c\,\tilde{x}$$

S))))))))))))))))))))))))))))))))))))

Ejemplo 1.6

Los valores observados de las cantidades $\sum_{i=1}^{n} x_i$ y $\sum_{i=1}^{n} x_i^2$ en el estudio de la vida útil, en horas, de las baterías de litio para cierta calculadora son:

$$\sum_{i=1}^{50} x_i = 63\,707 \qquad \text{y} \qquad \sum_{i=1}^{50} x_i^2 = 154\,924\,261\,.$$

a) ¿Sorprendería la afirmación de que la duración media de las baterías de litio usadas en esa calculadora es de 1270 horas? Responder y explicar utilizando solamente estadística descriptiva.

b) Calcular la variancia y la desviación estándar muestrales de estos datos.

Resolución

S))))))))))))))))))))))))))))))))))))

a) De los datos se obtiene $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$= 1274.14$,

No sorprendería la afirmación debido a que el valor es muy cercano al observado en la muestra.

b) $s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$

$= \frac{154\,924\,261}{50} - (1274.14)^2$

$= 1\,475\,052.48$

Y la variancia muestral es $s_{n-1}^2 = \frac{n s_n^2}{n-1}$

de donde

$s_{n-1}^2 = 1\,505\,155.592$ y

$s_{n-1} = 1\,226.8478$

S))

DIAGRAMA DE CAJA

Finalmente, después de estudiar medidas como la media, la mediana, el rango y los cuartiles es posible estudiar el diagrama de caja. El diagrama de caja esta compuesto por una figura rectangular, cuyos lados proporcionan los cuartiles Q_1 y Q_3, y en la parte intermedia de la caja se dibuja una línea que representa el segundo cuartil o la mediana $\tilde{x} = Q_2$, a los extremos de la caja se agregan líneas (llamadas bigotes), para se prolongan hasta los valores mínimo y máximo respectivamente. Es común localizar la media de los datos con un punto. Los diagramas de caja proporcionan entonces una forma visual de identificar el rango y los cuartiles. Es especialmente útil para comparar muestras o poblaciones y permite con gran facilidad determinar si los datos son simétricos.

S))

Ejemplo 1.7

Empleando los datos :
-1,0, 1, 2, 3, 4, 5, 7, 6, 5, 4, 2, 4, 6, 8, 6, 4, 2, 4, 6, 4, 3, 2, 2, 3, 4, 5, 6, 4, 2
trazar el diagrama de caja.

Resolución

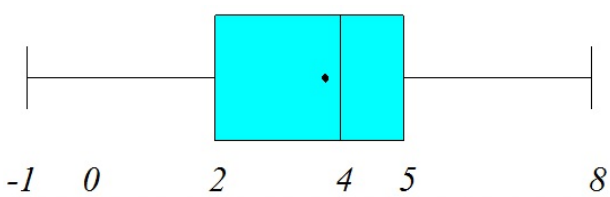
Para obtener los cuartiles se requiere ordernar los datos:

-1, 0, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 8

de donde

$$Q_1 = 2$$
$$\tilde{x} = Q_2 = 4$$
$$Q_3 = 5$$
$$\bar{x} = 3.7667$$
$$x_{\min} = -1$$
$$x_{\max} = 8$$

Por lo que la gráfica es:



S)))))))))))))))))))))))))))))))))))))))

Ejercicios de autoevaluación

1.- La investigación básica busca: $\square$

A) Reconstruir hechos a través de cifras.
B) el conocimiento de los fenómenos naturales.
C) La aplicación práctica del conocimiento.
D) La recopilación de la información.
E) Explicar la información de una muestra.

2.- La inferencia estadística es la parte de la estadística que: $\square$

A) Ordena y presenta los datos para su interpretación.
B) Estudia las pruebas y modelos en los que se conoce la distribución de la población bajo análisis.
C) Busca obtener conclusiones con respecto a la población a partir de la información contenida en una muestra.
D) Asigna probabilidades *a priori* a lo eventos.
E) Determina la relación entre poblaciones.

3.- El muestreo por conglomerados consiste en: $\square$

A) Seleccionar elementos de tal forma que todos tengan la misma probabilidad de ser seleccionados.
B) Seleccionar una muestra tomando cada k-ésima unidad de la población un elemento.
C) Seleccionar grupos con características semejantes y de un grupo se selecciona la muestra.
D) Dividir a la población en conglomerados y de cada conglomerado algunos elementos mediante muestreo aleatorio simple .
E) Dividir en grupos con las mayores diferencias posibles y de un grupo realizar un muestreo aleatorio simple.

4.- Dada la muestra:

$$1 \ , \ 3 \ , \ 5 \ , \ 3 \ , \ 2 \ , \ 1 \ , \ 6 \ , \ 5 \ , 3$$

La mediana es: $\square$

A) 1 B) 2 C) 3 D) 3.25 E) Ninguna de las anteriores

5.- Si al construir una tabla de distribución de frecuencias los límites del primer intervalo son $3.5 \ - \ 8.9$, entonces la maraca de clase del tercer intervalo es: $\square$

A) 14.5 B) 14.45 C) 17.2 D) 11.7 E) Ninguna de las anteriores

6.- El muestreo aleatorio simple se caracteriza por: $\boxed{}$

A) El investigador realiza la elección por conveniencia.
B) La elección se hace de manera aleatoria.
C) Cada posible muestra tiene la misma posibilidad de ser seleccionada.
D) Se utiliza un razón de muestreo.
E) Ninguna de las anteriores.

7.- Es una técnica de la estadística descriptiva: $\boxed{}$

A) Elección del tamaño de muestra.
B) Muestreo por conglomerados.
C) Cálculo de la probabilidad de los datos.
D) Cálculo de la correlación de los datos.
E) Distribución de frecuencias.

8.- Es la proporción de datos en la muestra que son menores o iguales que el límite superior de una clase $\boxed{}$:

A) Marca de clase.
B) Frecuencia.
C) Frecuencia acumulada .
D) Frecuencia relativa.
E) Frecuencia acumulada relativa.

9.- Al completar la siguiente tabla de distribución de frecuencias para 25 datos, el valor f_3 es: $\boxed{}$

Límites	Fronteras	Marca de clase x_i	Frecuencia f_i	Frecuencia relativa f'_i	Frec. relat. acum. $\left(F'_i\right)$
1-			8		
-	-	6.5		0.20	
-	-				0.72
-	-			0.16	
- 20	-				

A) 2 B) 3 C) 4 D) 5 E) 6

BIBLIOGRAFÍA

Spiegel, Murray R.- Estadística.- McGraw-Hill.- Segunda edición.- México, 1991.

Bonilla, Gildaberto,- Métodos Prácticos de Inferencia Estadística.- Trillas.- México, 1991.

Chou, Ya-Lun .- Análisis Estadístico .-McGraw-Hill.- Segunda Edición.- México, 1990.

Weimer, Richard C.- Estadística.- CECSA.- México, 1996.

Hines, William W. y Montgomery, Douglas C. - Probabilidad y Estadística para ingeniería, cuarta edición..- CECSA.- México, 2005.

Milton, Susan J. Y Arnold, Jesse C.- Probabilidad y Estadística para con aplicaciones para ingeniería y ciencias computacionales, cuarta edición.- McGraw-Hill.- México, 2004.

Aguilar Juárez, Isabel Patricia.- Apuntes de clase.

Devore, Jay L.- Probabilidad y Estadística para ingeniería y ciencias, séptima edición.- Cengage Learning.- México, 2008.

Mendenhall, William III. et al.- Introducción a la Probabilidad y Estadística.- Décimo cuarta edición.- Cengage Learning.- México 2015.

Wackerly Dennis D.- Mendenhall, William, *et al*.- Estadística Matemática con Aplicaciones, sexta edición.- Editorial Thomson.- México, 2002.

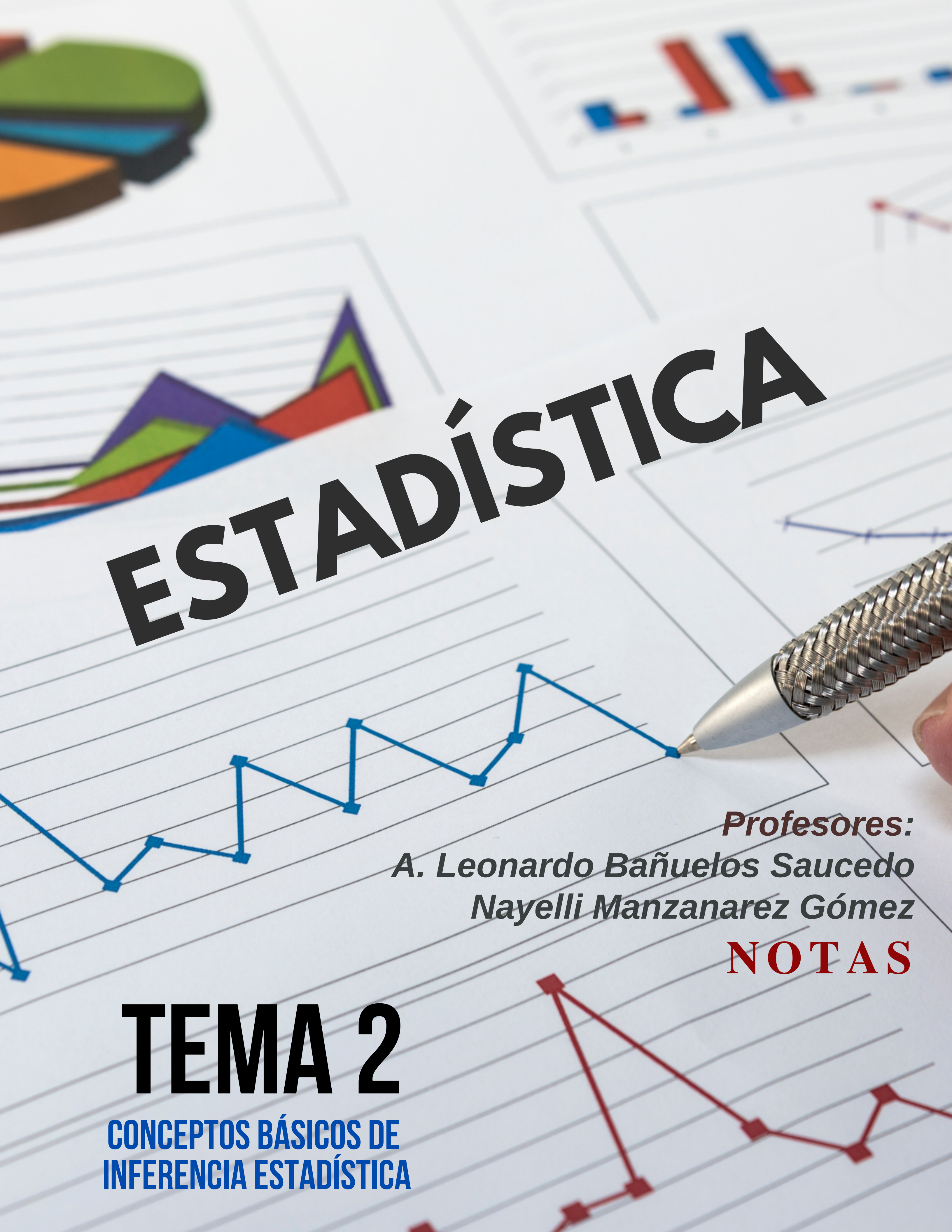
Walpole, Ronald E., *et al*.- Probability and Statistics for Engineers and Scientists.- Pearson.- USA, 2007.

Montgomery, Douglas C. y Runger, George C.-Probabilidad y Estadística aplicadas a la Ingeniería, segunda edición.- Limusa-Wiley.- México, 2002.

Scheaffer, Richard L. y McClave, James T.- Probabilidad y Estadística para Ingeniería.- Grupo Editorial Iberoamérica.- México, 1993.

Canavos, George C.- Probabilidad y Estadística Aplicaciones y Métodos.- McGraw-Hill.- México, 1988.

Borras García, Hugo E., *et al*.- Apuntes de Probabilidad y Estadística.-Facultad de Ingeniería.- México, 1985.



ESTADÍSTICA

Profesores:
A. Leonardo Bañuelos Saucedo
Nayelli Manzanarez Gómez

NOTAS

TEMA 2

CONCEPTOS BÁSICOS DE INFERENCIA ESTADÍSTICA

TEMA II
CONCEPTOS BÁSICOS DE INFERENCIA ESTADÍSTICA

La inferencia estadística es la parte de la estadística que tiene por objeto obtener conclusiones acerca de toda una población a partir de la información contenida en una muestra, cuantificando en forma probabilística el grado de certidumbre de dichas conclusiones. Y para obtener la información que permite generar las conclusiones utiliza, como se explicó en el capítulo uno, el muestreo aleatorio; del cual se obtienen muestras representativas.

Si bien, se dice que se tiene una muestra aleatoria cuando todos los elementos de la población tienen cierta probabilidad de ser seleccionados, es necesario definir las características que deben cumplir ciertas variables aleatorias para que puedan, en conjunto, generar una muestra aleatoria.

Definición 2.1 Parámetro

Un *parámetro* estadístico es un número que resume el comportamiento de una variable aleatoria y que describe parcial o completamente su distribución de probabilidad.

La Media y la variancia son parámetros de cualquier variable aleatoria. En el curso de probabilidad se estudiaron variables aleatorias así como sus parámetros de tendencia central, de dispersión y de forma.

Definición 2.2

Las variables aleatorias $X_1, X_2, \ldots, X_n$ forman una *muestra aleatoria* de tamaño n, si son independientes y tienen la misma distribución de probabilidad.

En términos sencillos la independencia de las variables aleatorias significa que el conocimiento del valor que toma una de las variables no afecta el valor que podrán tomar el resto de las variables; mientras que cuando se dice que tienen la misma distribución de probabilidad, debemos entender que son variables extraídas de la misma población.

En adelante, cuando se hable de una muestra aleatoria, o de las variables aleatorias de muestreo, deberán tenerse presentes las características de independencia e idéntica distribución.

Definición 3.3

Un *estadístico*[1] es una función de las variables aleatorias que se pueden observar en una muestra y que no depende de parámetros desconocidos.

[1] También se llama estadística o estadígrafo.

Cuando se utilizan las variables aleatorias de muestro para formar con ellas una función, se obtiene una estadístico, pero debe vigilarse que no dependa de ningún parámetro desconocido. Así, si se construye la función: $Y = X_1 + X_2$, se tiene un estadístico. Otro ejemplo es: $\bar{X} = \sum_{i=1}^{n} X_i$, donde aparece el parámetro n, pero es conocido, por lo que también es un estadístico.

Definición 2.4

Si X es una variable aleatoria con función de densidad o de probabilidad $f_X(x;\theta)$, donde θ es un parámetro desconocido, y si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de tamaño n, entonces el estadístico

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

recibe el nombre de *estimador* de θ .

Debe observarse que el estimador $\hat{\Theta}$ del parámetro θ , es una variable aleatoria porque es una función de los datos de muestreo. Cuando se sustituyen las variables aleatorias $X_1, X_2, \ldots, X_n$ por sus valores observados $x_1, x_2, \ldots, x_n$, entonces se tiene una estimación $\hat{\theta}$ del parámetro θ.

En términos sencillos, puede decirse que un estimador es un estadístico que tiene como propósito definido el "aproximar" un parámetro desconocido. Evidentemente, el estimador no depende de parámetros desconocidos, de hecho, "despeja" al parámetro desconocido.

Con el propósito de relacionar los valores observados en una muestra, con las variables aleatorias de muestreo y los estadísticos, deberá interpretarse la tabla 2.1, en la cual se muestran por renglón las distintas muestras que se pueden observar, siendo x_{ij} , valores de la muestra i , y dentro de esa muestra, el elemento j ; con $1 \le i \le m$, $1 \le j \le n$. Adicionalmente, para cada muestra i ; se obtiene el promedio de los datos de la muestra $\bar{x}_i$, $1 \le i \le m$. Finalmente, antes de tomar o extraer la muestra, no se sabe cual será el valor de la primera observación, por lo que se tiene una variable aleatoria, y así para el resto de las observaciones, esto es: X_i , $1 \le i \le n$ son las variables aleatorias que representan el valor que podrá observarse en la muestra en la i-ésima observación y $\bar{X}$ es el promedio de las X_i.

	X_1	X_2		X_n	$\bar{X}$
muestra 1	x_{11}	x_{12}	. . .	x_{1n}	$\bar{x}_1$
muestra 2	x_{12}	x_{22}	. . .	x_{2n}	$\bar{x}_2$
.	.	.	. . . .	.	.
.	.	.	. . . .	.	.
.	.	.	. . . .	.	.
muestra m	x_{m1}	x_{m2}	. . .	x_{mn}	$\bar{x}_m$

Tabla 2.1 Variables de muestreo

Evidentemente, si las X_i son variables aleatorias, cualquier función que se genere de ellas, por ejemplo $\bar{X}$, es también una variable aleatoria. La distribución de las variables aleatorias generadas a partir de las variables de muestreo, se estudiará en el capítulo siguiente.

El objetivo de la inferencia estadística será el de aproximar los parámetros de la población; por ejemplo, la media de la población μ se aproximará mediante el estadístico $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, que recibe el nombre de *media muestral*, el cual a su vez se valuará para una muestra en particular a través de $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, que recibe el nombre de *media de la muestra*.

DISTRIBUCIÓN MUESTRAL DE LA MEDIA: DISTRIBUCIÓN NORMAL Y TEOREMA CENTRAL DEL LÍMITE

La distribución normal servirá para caracterizar la distribución muestral de la media. Con el propósito de introducir la forma en la que se utilizarála distribución normal, se recordará la propiedad de aditividad de la distribución normal y otro teoremas importantes.

Teorema 2.1

Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes, y todos con distribución normal con media μ_i y variancia σ_i^2, entonces la variable aleatoria Y definida como

$$Y = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

tiene distribución normal con media $\sum_{i=1}^{n} \mu_i$ y variancia $\sum_{i=1}^{n} \sigma_i^2$.

Demostración

Si $X \sim N(\mu, \sigma^2)$, entonces su función generadora de momentos está dada por

$$M_X(\theta) = e^{\mu\theta + \frac{1}{2}\sigma^2\theta^2}$$

y si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y $Y = \sum_{i=1}^{n} X_i$

$$\begin{aligned} M_Y(\theta) &= M_{X_1}(\theta)\, M_{X_2}(\theta) \ldots M_{X_n}(\theta) \\ &= \left(e^{\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2}\right)\left(e^{\mu_2\theta + \frac{1}{2}\sigma_2^2\theta^2}\right)\ldots\left(e^{\mu_n\theta + \frac{1}{2}\sigma_n^2\theta^2}\right) \\ &= e^{\mu_1\theta + \mu_2\theta + \ldots + \mu_n\theta + \frac{1}{2}\sigma_1^2\theta^2 + \frac{1}{2}\sigma_2^2\theta^2 + \ldots + \sigma_n^2\theta^2} \\ &= e^{(\mu_1 + \mu_2 + \ldots + \mu_n)\theta + \frac{1}{2}(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)\theta^2} \end{aligned}$$

que es la función generadora de momentos de la distribución normal con media $\sum_{i=i}^{n} \mu_i$ y variancia $\sum_{i=1}^{n} \sigma_i^2$

$$Y \sim N\left(\sum_{i=1}^{n} \mu_i\,,\ \sum_{i=1}^{n} \sigma_i^2\right)$$

■

La característica de aditividad de la distribución normal se extiende a combinaciones lineales, teniéndose los siguientes teoremas.

Teorema 2.2

Si $X_i, X_2, \ldots, X_n$ son variables aleatorias independientes y todas con distribución normal con media μ_i y variancia σ_i^2 entonces la variable aleatoria Y_c definida como

$$Y_c = c_1 X_1 + c_2 X_2 + \ldots + c_n X_n = \sum_{i=1}^{n} c_i X_i$$

tiene distribución normal con media $\sum_{i=1}^{n} c_i \mu_i$ y variancia $\sum_{i=1}^{n} c_i^2 \sigma_i^2$.

Teorema 2.3

Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas, y $X_i \sim N(\mu, \sigma^2)$ entonces la variable aleatoria

$$Y = \sum_{i=1}^{n} X_i$$

tiene distribución normal. $Y \sim N(n\mu,\ n\sigma^2)$ por lo que $\dfrac{Y - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$

Además de los teoremas para las distribuciones de sumas de variables aleatorias, un caso particular se tiene cuando se desea obtener la distribución del estadístico $\bar{X}$ definido como $\frac{1}{n}\sum_{i=1}^{n} X_i$, que se puede estudiar como un caso particular de $Y = \sum_{i=1}^{n} c_i X_i$ en el cual $c_i = \frac{1}{n}$ para $i = 1, 2, \ldots, n$, obteniéndose el siguiente teorema.

Teorema 2.4

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria con distribución normal y media μ y variancia σ^2. Entonces la variable aleatoria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tiene distribución normal con media μ y variancia $\frac{\sigma^2}{n}$.

$$\bar{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$$

Para ilustrar la propiedad de aditividad de la distribución normal, considérese el siguiente ejemplo, en el cual, las medias se restan, pero las variancias se siguen sumando, debido a las propiedades de la suma de variancias.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.1

Un eje de diámetro exterior distribuido normalmente con media 1.2 cm y variancia 0.0016 cm^2 se inserta en un cojinete que tiene un diámetro interior distribuido normalmente con media 1.25 cm y variancia $0.0009\ \mathrm{cm}^2$. Determinar la probabilidad de que no embonen las piezas.

Resolución

Sea X_1 la variable aleatoria que representa el diámetro exterior del eje y X_2 la variable aleatoria que representa el diámetro interior del cojinete.

$$X_1 \sim N(1.2, 0.0016)$$
$$X_2 \sim N(1.25, 0.0009)$$

Para que no embonen $X_1 > X_2$ o bien $Y = X_1 - X_2 > 0$

$$Y \sim N(-0.05, 0.0025)$$

Por lo que

$$P(Y > 0) = P\left(\frac{Y - (-0.05)}{\sqrt{0.0025}} > \frac{0 - (-0.05)}{\sqrt{0.0025}} \right)$$
$$= P(Z > 1) = 0.1587$$

$\therefore$ La probabilidad de que no embonen las piezas es 0.1587

S))

Los teoremas anteriores son útiles si se desea obtener la distribución de una suma de variables aleatorias distribuidas normalmente; sin embargo, esto no siempre ocurre. En general las variables aleatorias pueden tener cualquier distribución y no sólo la normal. Cuando se presentan estos casos se utiliza el teorema central de límite[1].

Teorema 2.5 (Central del límite)

Sean $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias independientes e idénticamente distribuidas con parámetros $\mathrm{E}(X_i) = \mu_X$ y $Var(X_i) = \sigma_X^2$ para $i = 1, 2, \ldots, n$.

Entonces la variable aleatoria $\bar{X}$ definida como $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

tiene una distribución que converge a la normal, con parámetros $\mu_{\bar{X}} = \mu_X$ y $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$

estándar cuando $n \to \infty$, esto es:

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X\ ,\ \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \right)$$

En general, si las n variables aleatorias independientes tienen parámetros $\mathrm{E}(X_i) = \mu_i$ y

[1]Es muy común estudiar este teorema como Teorema Central del Límite, pero este nombre proviene de una falla en la traducción de la idea original expresada por el matemático húngaro George Polya (1887-1985), que intentaba expresar la importancia del teorema, de ahí el nombre de Central: Teorema Central.

Este teorema tiene sus inicios en el libro publicado por Abraham de Moivre, *The Doctrine of Chances*, y a partir de ahí, son un conjunto de teoremas los que han recibido el distintivo de Teoremas Centrales.

$\text{Var}(X_i) = \sigma_i^2$ y se define la variable aleatoria $Y = \sum_{i=1}^{n} X_i$ entonces $\dfrac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$ tiene distribución que converge a la normal estándar.

En la práctica, el teorema central del límite (TCL) proporciona una buena aproximación cuando n es mayor o igual que 30, sin importar la distribución de las variables aleatorias de muestro. En el caso particular de que las variables provengan de una distribución uniforme, una muestra de tamaño 12 produce ya una buena aproximación.

Si la población es normal, entonces puede utilizarse la distribución normal por la característica de aditividad.

S))))))))))))))))))))))))))))))))))))Q

Ejemplo 2.2

El consumo de cierto tipo de focos led, tiene un promedio de población de 9.5 [W] y una desviación estándar de 0.5, según las especificaciones de producción. Si se instalan ocho de estos focos, calcular la probabilidad de que el consumo promedio sea mayor a 10 [W], suponiendo que las mediciones del consumo tienen distribución normal.

Resolución

Sea $\bar{X}$ el consumo promedio de los ocho focos, $\bar{X} \sim N(9.5, \frac{0.25}{8})$

$$P(\bar{X} > 10) = P\left(Z > \frac{10 - 9.5}{\frac{0.5}{\sqrt{8}}} \right) = P(Z > 2.83)$$

$$P(\bar{X} > 10) = 0.0023$$

S))))))))))))))))))))))))))))))))))))Q

Pero cuando la población es desconocida y la muestra es grande, se utiliza el teorema central del límite.

S))))))))))))))))))))))))))))))))))))Q

Ejemplo 2.3

La resistencia a la ruptura de un remache especial tiene una valor medio de 10 000 kilogramos por centímetro cuadrado y una desviación estándar de 500.

a) ¿Cuál es la probabilidad de que la resistencia media a la ruptura de la muestra, para una muestra aleatoria de 40 remaches, sea entre 9900 y 10200?

b) Si el tamaño muestral hubiera sido 15 en lugar de 40, ¿podría calcularse la probabilidad pedida en el inciso (a)?

Resolución

a) Puesto que $n = 40$ se puede utilizar el teorema central del límite, por lo que

$$P(9900 \le \bar{X} \le 10200) = P\left(\frac{9900 - 10000}{\frac{500}{\sqrt{40}}} \le Z \le \frac{10200 - 10000}{\frac{500}{\sqrt{40}}} \right)$$

$$= P(-1.26 \le Z \le 2.53) = 0.8905$$

b) Si la muestra es de tamaño 15, se requiere conocer la distribución de la población, puesto que el TCL se utiliza a partir de 30, por lo que, con la información proporcionada NO puede calcularse la probabilidad.

S)))))))))))))))))))))))))))))))))))))Q

Como se comentó antes, la distribución normal sirve para caracterizar la media muestral, y deben resaltarse los siguientes resultados relativos a la media y a la variancia.

Si se extrae una muestra aleatoria de una población infinita (o finita pero el muestreo es con reemplazo), con parámetros μ_X y σ_X^2, entonces el estadístico media muestral $\bar{X}$, tiene los siguientes parámetros:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)\right) \\ &= \frac{1}{n}\left(\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n) \right) \\ &= \frac{1}{n}(n\mu) = \mu \end{aligned}$$

lo que significa que el valor esperado de la media muestral es la media de la población y,

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)\right) \\ &= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) \right) \\ &= \frac{1}{n^2}\left(n\sigma_X^2\right) = \frac{\sigma_X^2}{n} \end{aligned}$$

la variancia de la media muestral es la variancia de la población dividida entre el tamaño de la muestra.

Por supuesto, estos resultados son independientes de la población.

A la desviación estándar de una distribución de muestreo se le suele llamar *error estándar*, puesto que mide la variabilidad del muestreo debida a casualidad o a fuerzas aleatorias. El error estándar de la media es entonces:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Obsérvese que el error estándar es menor que la desviación estándar de la población para muestras de tamaño dos o mayores, y que cuando n tiende a infinito el error estándar tiende a cero, es decir:

$$\lim_{n \to \infty} \frac{\sigma_X}{\sqrt{n}} = 0$$

Esto es, mientras mayor sea el tamaño de la muestra menores serán las fluctuaciones entre la media de una muestra y otra.

Cuando el muestreo se realiza en una población finita y sin reemplazo, debe introducirse el factor que se conoce como *Factor de Corrección por Población Finita*, el cual se denota ***FCPF***.

$$FCPF = \frac{N - n}{N - 1}$$

donde N es el tamaño de la población. Para la variancia se tiene: $\sigma^2_{\bar{X}} = \dfrac{\sigma^2_X}{n}\, FCPF$

Por lo que el error estándar cuando la población es finita queda:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Las diferencias entre los muestreos con y sin reemplazo; y sus afectaciones en la variancia muestral se observan en el siguiente ejemplo.

S)))))))))))))))))))))))))))))))))))Q

Ejemplo 2.4

Considérese una población que sólo contiene a los números 0, 1, 2 y 3.

a) Obtener la media y la variancia de la población.

b) Si se seleccionan muestras de tamaño dos con reemplazo, obtener la media y la variancia del promedio muestral.

c) Obtener la media y la variancia del promedio muestral a partir de la media y la variancia de la población, para el caso de muestreo con reemplazo.

d) Si se seleccionan muestras de tamaño dos sin reemplazo, obtener la media y la variancia del promedio muestral.

e) Obtener la media y la variancia del promedio muestral a partir de la media y la variancia de la población, para el caso de muestreo sin reemplazo.

Resolución

a) Puesto que la población está formada por cuatro elementos y todos tienen la misma posibilidad de ser seleccionados, la distribución de probabilidad es:

x	0	1	2	3
$f_X(x)$	0.25	0.25	0.25	0.25

Entonces:

$$\mu_X = \sum_{\forall x} x\, f_X(x) = \frac{1}{4}\left(0 + 1 + 2 + 3\right) = 1.5$$

$$\sigma^2_X = \sum_{\forall x} x^2 f_X(x) - \mu^2_X$$

$$\sigma^2_X = \frac{1}{4}\left(0^2 + 1^2 + 2^2 + 3^2\right) - (1.5)^2 = 1.25$$

b) Las muestras de tamaño 2 con reemplazo y considerando el orden son:

$$(0,0),(0,1),(0,2),(0,3)$$
$$(1,0),(1,1),(1,2),(1,3)$$
$$(2,0),(2,1),(2,2),(2,3)$$
$$(3,0),(3,1),(3,2),(3,3)$$

Se tienen 16 muestras.

Las medias son:

$$\begin{array}{cccc} 0\,, & 0.5\,, & 1\,, & 1.5 \\ 0.5\,, & 1\,, & 1.5\,, & 2 \\ 1\,, & 1.5\,, & 2\,, & 2.5 \\ 1.5\,, & 2\,, & 2.5\,, & 3 \end{array}$$

La función de probabilidad para las medias muestrales es:

$\bar{x}$	0	0.5	1	1.5	2	2.5	3
$f_{\bar{X}}(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

De donde la media es:

$$\mu_{\bar{X}} = E(\bar{X}) = \sum_{\forall \bar{x}} \bar{x} f_{\bar{X}}(\bar{x})$$

$$\mu_{\bar{X}} = (0)\left(\frac{1}{16}\right) + (0.5)\left(\frac{2}{6}\right) + \,.\,.\,.\, + (3)\left(\frac{1}{16}\right)$$

$$\mu_{\bar{X}} = 1.5$$

$$\sigma^2_{\bar{X}} = \sum_{\forall \bar{x}} \bar{x}^2 f_{\bar{X}}(\bar{x}) - \mu^2_{\bar{X}}$$

$$\sigma^2_{\bar{X}} = (0)^2\left(\frac{1}{16}\right) + (0.5)^2\left(\frac{2}{6}\right) + \,.\,.\,.\, + (3)^2\left(\frac{1}{16}\right) - (1.5)^2$$

$$\sigma^2_{\bar{X}} = 0.625$$

c) $\mu_{\bar{X}} = \mu_X = 1.5$

$$\sigma^2_{\bar{X}} = \frac{\sigma_X}{n} = \frac{1.25}{2} = 0.625$$

Los resultados coinciden con los obtenidos con los valores muestrales.

d) Las muestras de tamaño 2 sin reemplazo son:

$$\begin{array}{c} (0,1),(0,2),(0,3) \\ (1,0),(1,2),(1,3) \\ (2,0),(2,1),(2,3) \\ (3,0),(3,1),(3,2) \end{array}$$

Se tienen 12 muestras.

Las medias son:

$$\begin{array}{ccc} 0.5\,, & 1\,, & 1.5 \\ 0.5 & 1.5\,, & 2 \\ 1\,, & 1.5\,, & 2.5 \\ 1.5\,, & 2\,, & 2.5 \end{array}$$

de donde:

$\bar{x}$	0.5	1	1.5	2	2.5
$f_{\bar{X}}(\bar{x})$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{2}{12}$

$$\mu_{\bar{X}} = E(\bar{X}) = \sum_{\forall \bar{x}} \bar{x} f_{\bar{X}}(\bar{x})$$

$$\mu_{\bar{X}} = (0.5)\left(\frac{2}{12}\right) + (1)\left(\frac{2}{12}\right) + \ldots + (2.5)\left(\frac{2}{12}\right)$$

$$\mu_{\bar{X}} = 1.5$$

$$\sigma^2_{\bar{X}} = \sum_{\forall \bar{x}} \bar{x}^2 f_{\bar{X}}(\bar{x}) - \mu^2_{\bar{X}}$$

$$\sigma^2_{\bar{X}} = (0.5)^2\left(\frac{2}{12}\right) + (1)^2\left(\frac{2}{12}\right) + \ldots + (2.5)^2\left(\frac{2}{12}\right) - (1.5)^2$$

$$\sigma^2_{\bar{X}} = 0.41666$$

e) $\mu_{\bar{X}} = \mu_X = 1.5$

Para la variancia, como la población es finita y el muestreo se hace sin reemplazo, se debe de utilizar el factor de corrección población finita, por lo que

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}\,\frac{N-n}{N-1} = \frac{1.25}{2}\left(\frac{4-2}{4-1}\right) = 0.41666$$

Los resultados coinciden con los obtenidos anteriormente.

S))))))))))))))))))))))))))))))))))Q

DISTRIBUCIÓN PARA CARACTERIZAR LA DIFERENCIA DE MEDIAS, VARIANZAS CONOCIDAS: NORMAL

La distribución normal, se puede utilizar también para caracterizar la distribución de muestreo de la diferencia de medias muestrales. Si las muestras son normales y se conoce la variancia, entonces la diferencia de medias $\bar{X} - \bar{Y}$ tiene una distribución normal, por lo que

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma^2_X}{n_X} + \dfrac{\sigma^2_Y}{n_Y}}}$$

Si se desconocen las variancias de las poblaciones σ^2_X y σ^2_Y; y las muestras son grandes, entonces se pueden sustituir por sus estimadores puntuales $S^2_{n_X-1}$ y $S^2_{n_Y-1}$, y se utiliza la expresión anterior debido al TLC.

La caracterización de la media se utilizará para intervalos de confianza y pruebas de hipótesis relacionadas con la media de una población.

S)))

Ejemplo 2.5

La gerente de una planta de una fábrica enlatadora de jugo de naranja está interesada en comparar el rendimiento de dos diferentes líneas de producción. Como la línea 1 es relativamente nueva, sospecha que el número de cajas que se producen al día es mayor que el correspondiente a la vieja línea 2. Se toman datos al azar durante diez días para cada línea, encontrándose que $\bar{x}_1 = 824.9$ cajas por día y $\bar{x}_2 = 818.6$ cajas por día. Se sabe por experiencia que $\sigma_1^2 = 40$ y $\sigma_2^2 = 50$.¿Qué tan probables son las sospechas de la gerente?

Resolución

Se desea calcular

$$P\left(\bar{X}_1 - \bar{X}_2 \geq \bar{x}_1 - \bar{x}_2 \;\middle|\; \mu_1 = \mu_2\right)$$

El estadístico está dado por

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \quad ; \qquad \text{donde } Z \sim N(0\,,1)$$

$$z = \frac{(824.9 - 818.6) - 0}{\sqrt{\dfrac{40}{10} + \dfrac{50}{10}}} = 2.1$$

$$\begin{aligned} P\left(\bar{X}_1 - \bar{X}_2 \geq 6.3\right) &= P(Z \geq 2.1) \\ &= 1 - P(Z \leq 2.1) \\ &= 1 - 0.982 \\ &= 0.018 \end{aligned}$$

Dado que la probabilidad es baja, podemos decir que las sospechas de la gerente son acertadas y que la línea 1 produce más que la línea 2.

S)))

Por otro lado, cuando las poblaciones son normales, las muestras son pequeñas, con variancias desconocidas pero iguales, entonces se puede utilizar la distribución t para caracterizar la diferencia de medias muestrales, la cual se estudiará más adelante.

Si las muestras son pequeñas, de poblaciones normales, con variancias desconocidas y diferentes se puede utilizar una aproximación mediante la distribución t , la cual se estudiará en temas posteriores. Cualquier otro caso queda fuera del alcance de este curso.

DISTRIBUCIÓN PARA CARACTERIZAR A LA VARIANCIA: DISTRIBUCIÓN Ji CUADRADA

Por las características especiales que presenta la distribución normal, se han estudiado distribuciones que puedan generarse a partir de ella, tal es el caso de la distribución Ji cuadrada, también llamada en ocasiones chi-cuadrada, del inglés *chi-square*.

Definición 2.5

Sean $Z_1, Z_2, \ldots, Z_\nu$; ν variables aleatorias independientes con distribución normal estándar, entonces:

$$X^2 = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$$

es una variable aleatoria que recibe el nombre de Ji cuadrada con ν grados de libertad y se denota mediante el símbolo $\chi^2_{(\nu)}$.

Su función de densidad es

$$f_{X^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{\nu}{2}}\,\Gamma\left(\dfrac{\nu}{2}\right)}\, x^{\left(\frac{\nu}{2}-1\right)}\, e^{-\frac{x}{2}} & x > 0 \\ 0 & en\ otro\ caso \end{cases}$$

Debe recordarse que la distribución Ji cuadrada es un caso particular de la distribución Gamma, en donde $\lambda = \frac{1}{2}$ y $r = \frac{\nu}{2}$.

La distribución Ji cuadrada, al igual que la distribución normal, presenta la característica de aditividad.

Teorema 2.6

Sean $X_1^2, X_2^2, \ldots, X_n^2$ variables aleatorias independientes con distribución Ji cuadrada y $\nu_1, \nu_2, \ldots, \nu_n$ grados de libertad, respectivamente. Entonces la variable

$Y = X_1^2 + X_2^2 + \ldots + X_n^2 = \sum_{i=1}^{n} X_i^2$ tiene una distribución Ji cuadrada con $\nu = \sum_{i=1}^{n} \nu_i$ grados de libertad.

La demostración del teorema anterior resulta como consecuencia directa de la definición de la distribución Ji cuadrada, que es una suma de variables aleatorias con distribución normal estándar al cuadrado; por lo que una suma de Ji cuadradas sigue siendo una suma de normales estándar al cuadrado lo que sigue siendo una Ji cuadrada.

Para introducir el uso de la distribución Ji cuadrada para caracterizar a la variancia muestral, considérese el siguiente caso.

Si $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria con distribución normal con media μ y variancia σ^2 conocidas, entonces la distribución de la variable aleatoria Y definida por $Y = \dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2}$, se obtiene directamente de la definición de la variable aleatoria Ji-cuadrada, es decir:

Del enunciado $X_i \sim N(\mu, \sigma^2)$ por lo que $\dfrac{X_i - \mu}{\sigma} = Z_i \sim N(0, 1)$ y la variable aleatoria Y se puede reescribir como $Y = \sum_{i=1}^{n} Z_i^2$ de donde se observa que Y tiene una distribución Ji cuadrada con n grados de libertad.

$$Y \sim \chi^2_{(n)}$$

Sin embargo, la variable Y, no es una variancia muestral, para seguir buscando la relación entre la variancia muestral y la variable Ji-cuadrada, considérese ahora la variable S_μ^2, definida como $S_\mu^2 = \dfrac{\sum_{i=1}^{n} \left(X_i - \mu\right)^2}{n}$, entonces de la explicación anterior para la variable Y se observa que

$$\frac{n S_\mu^2}{\sigma^2} \sim \chi^2_{(n)}$$

puesto que $\dfrac{n S_\mu^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma} = Y$

Este resultado sirve para estudiar la distribución de muestreo de S_n^2, definida como

$$S_n^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n}$$

que es la variancia muestral.

O bien, para caracterizar a la variancia muestral definida como:

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Considérese en particular la variancia muestral dividida entre $n - 1$

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

para obtener la distribución de S_{n-1}^2, se realiza el siguiente análisis.

Recordando que la extracción es de una población normal, $X_i \sim N(\mu, \sigma^2)$ se tiene que

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

y manipulando la expresión de la variancia,

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

se tiene
$$(n-1)\frac{S_{n-1}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2}$$

Trabajando exclusivamente con la suma:

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i - \mu - \bar{X} + \mu)^2$$

$$= \sum_{i=1}^{n} [(X_i - \mu) - (\bar{X} - \mu)]^2$$

$$= \sum_{i=1}^{n} [(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2]$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu)\, n(\bar{X} - \mu) + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

por lo que

$$(n-1)\frac{S_{n-1}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2}{\sigma^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} - \frac{n(\bar{X} - \mu)^2}{\sigma^2}$$

Despejando

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} = \frac{(\bar{X} - \mu)^2}{\left(\frac{\sigma}{\sqrt{n}}\right)^2} + \frac{(n-1) S_{n-1}^2}{\sigma^2}$$

y del teorema 2.6 y la definición 2.5 se observa que $\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2$ tiene distribución Ji cuadrada con n grados de libertad; mientras que $\dfrac{(\bar{X}-\mu)^2}{\left(\frac{\sigma}{\sqrt{n}}\right)^2}$ tiene también distribución Ji cuadrada con un grado de libertad, por lo cual puede concluirse que $\dfrac{(n-1)S^2}{\sigma^2}$ tiene distribución Ji cuadrada con $n-1$ grados de libertad. Finalmente:

$$\frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

De manera similar, para $S_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$, se tiene que:

$$\frac{n\,S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Puesto que, quien realiza la aportación de los grados de libertad es la suma, también se obtiene una distribución Ji-cuadrada con $n-1$ grados de libertad. El grado de libertad se pierde debido al desconocimiento de la media de la población, en el cálculo de la variancia.

Teorema 2.7

Si X^2 es una variable aleatoria que tiene una distribución Ji cuadrada con ν grados de libertad, $X^2 \sim \chi^2_{(\nu)}$, entonces

$$E(X^2) = \nu \quad , \quad \mathrm{Var}(X^2) = 2\nu$$

y la función generadora de momentos de X^2 es:

$$M_{X^2}(\theta) = (1-2\theta)^{-\frac{\nu}{2}}$$

La distribución Ji cuadrada presenta un sesgo positivo, según se muestra en la siguiente figura.

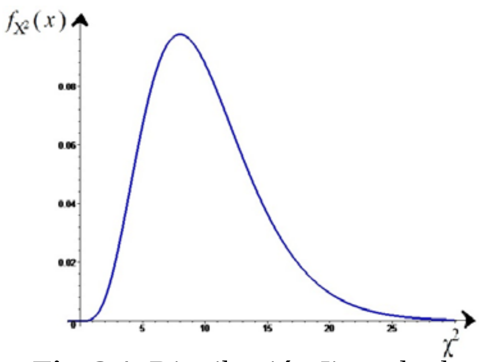


Fig. 2.1 Distribución Ji cuadrada

Para calcular probabilidades de variables aleatorias con distribución Ji cuadrada se utilizan tablas. Las tablas de la distribución Ji-cuadrada generalmente proporcionan el valor de χ^2 en función del área de la "cola derecha" α, y el valor de Ji-cuadrada se denota χ^2_α cuando $P\left(X^2 \geq \chi^2\right) = \alpha$, como se observa en la siguiente gráfica.

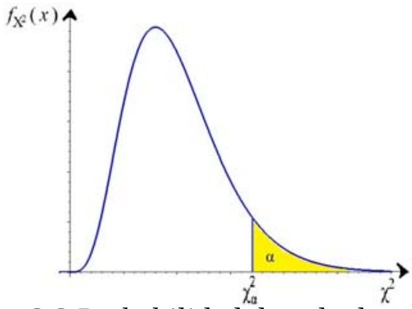


Fig. 2.2 Probabilidad de cola derecha

Resumen de notación

X^2 Variable aleatoria ji cuadrada.
χ^2 Valor que toma la variable aleatoria ji cuadrada.
$\chi^2_{(\nu)}$ Distribución ji cuadrada con ν(ni) grados de libertad.
χ^2_α Valor de ji-cuadrada para el cual la probabilidad de cola derecha es α.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.6

Utilizar las tablas de la distribución Ji-cuadrada para obtener las siguientes probabilidades

a) Si $X^2 \sim \chi^2_{(15)}$, obtener $P\left(X^2 \geq 15.73\right)$

b) Si $X^2 \sim \chi^2_{(20)}$, obtener $P\left(X^2 < 12.44\right)$

Resolución

a) De tablas de la distribución Ji-cuadrada, y observando que las tablas proporcionan probabilidades de "cola derecha" o de lado derecho, entonces, se busca en el renglón el número de grados de libertad y posteriormente el valor de $\chi^2 = 15.73$, es decir:

Distribución Ji cuadrada

	α					
ν	0.9	0.5	0.4	0.3	0.2	0.1
12	6.304	11.340	12.584	14.011	15.812	18.549
13	7.042	12.340	13.636	15.119	16.985	19.812
14	7.790	13.339	14.685	16.222	18.151	21.064
15	8.547	14.339	$\boxed{15.733}$	17.322	19.311	22.307
16	9.312	15.339	16.780	18.418	20.465	23.542
17	10.085	16.338	17.824	19.511	21.615	24.769
18	10.865	17.338	18.868	20.601	22.760	25.989
19	11.651	18.338	19.910	21.689	23.900	27.204
20	12.443	19.337	20.951	22.775	25.037	28.412
21	13.240	20.337	21.992	23.858	26.171	29.615
22	14.041	21.337	23.031	24.939	27.301	30.813

por lo que la probabilidad es

$$P\left(X^2 \geq 15.73 \right) = 0.4$$

b) De forma similar, pero utilizando el complemento

$$P\left(X^2 < 12.44 \right) = 1 - P\left(X^2 \geq 12.44 \right) = 1 - 0.9 = 0.1$$

))))))))))))))))))))))))))))))))))))))

Después de aprender a utilizar las tablas de la distribución Ji-cuadrada, es posible realizar cálculos de probabilidad que involucren a las variancias muestrales S^2_{n-1} o S^2_n.

))))))))))))))))))))))))))))))))))))))

Ejemplo 2.7

Sea una muestra aleatoria de tamaño 20 tomada de una población con media 8 y variancia 4. Obtener la probabilidad de que la variancia muestral S^2_{n-1} sea mayor o igual a 5.7.

Resolución

Puesto que la población es normal,

$$P\left(S^2_{n-1} \geq 5.7 \right) = P\left(\frac{n-1}{\sigma^2} S^2_{n-1} \geq \frac{19}{4}(5.7) \right)$$

$$= P\left(X^2 \geq 27.08 \right)$$

Y de tablas, con $X^2 \sim \chi^2_{(19)}$ se tiene:

$$P\left(S^2_{n-1} \geq 5.7 \right) \approx 0.1$$

S))))))))))))))))))))))))))))))))))))))

La distribución Ji cuadrada se utilizará en la construcción de intervalos de confianza y en pruebas de hipótesis relacionadas con la variancia de una población normal.

DISTRIBUCIÓN PARA CARACTERIZAR A LA MEDIA MUESTRAL, MUESTRA PEQUEÑA, VARIANCIA DESCONOCIDA: DISTRIBUCIÓN *t*-STUDENT

Definición 2.6

Sean Z y X^2 dos variables aleatorias independientes con distribuciones normal estándar y Ji-cuadrada respectivamente, es decir:

$$Z \sim N(0,1) \quad , \quad X^2 \sim \chi^2_{(\nu)}$$

entonces la variable aleatoria T definida como $T = \dfrac{Z}{\sqrt{\dfrac{X^2}{\nu}}}$

tiene una distribución t de Student con ν grados de libertad y función de densidad dada por

$$f_T(t) = \frac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\dfrac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < t < \infty$$

$$\nu > 0$$

Como puede observarse la distribución t de Student posee el parámetro ν, que al igual que para la distribución ji-cuadrada, recibe el nombre de grados de libertad.

Teorema 2.8

Si T es una variable aleatoria que tiene distribución t de Student con ν grados de libertad, entonces

$$E(T) = 0$$

$$Var(T) = \frac{\nu}{\nu - 2} \qquad \nu > 2$$

La función de densidad t de Student es simétrica y unimodal al igual que la normal, pero siempre está centrada en cero; es muy parecida a la distribución normal estándar.

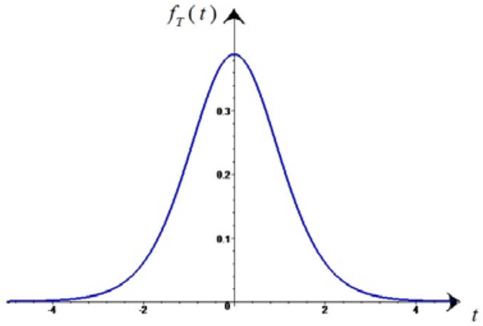


Fig 2.3. Distribución t de Student

La distribución t converge a la distribución normal estándar cuando el número de grados de libertad tiende a infinito.

La principal aplicación de la distribución t radica en la obtención de la distribución del estadístico

$$\frac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}}$$

que se utiliza para hacer inferencias con respecto a la media μ_X cuando el muestreo se lleva a cabo sobre una distribución normal con variancia desconocida.

Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con distribución normal con media μ_X y variancia σ_X^2 y se definen los estadísticos

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Entonces, para obtener la distribución de $\dfrac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}}$ se realiza el siguiente procedimiento:

Puesto que $\bar{X} \sim N\left(\mu_X, \dfrac{\sigma_X^2}{n}\right)$ y $\dfrac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2_{(n-1)}$ entonces, de la definición 2.6 se tiene:

$$\frac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} = \frac{\bar{X} - \mu_X}{\sqrt{\frac{(n-1)}{(n-1)}} \frac{S_{n-1}\ \sigma_X}{\sqrt{n}\ \sigma_X}} = \frac{\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S_{n-1}^2}{(n-1)\sigma_X^2}}}$$

de donde se observa que

$$\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0, 1)$$

$$\frac{(n-1)\,S_{n-1}^{2}}{\sigma^{2}} \sim \chi^{2}_{(n-1)}$$

entonces se tiene el cociente de una normal estándar entre la raíz cuadrada de una Ji cuadrada entre ν, por lo que

$$\frac{\bar{X}-\mu_X}{\frac{S_{n-1}}{\sqrt{n}}} = T \sim t_{(n-1)}$$

Esto es, la distribución de $\dfrac{\bar{X}-\mu_X}{\frac{S_{n-1}}{\sqrt{n}}}$ es una distribución t, con $n-1$ grados de libertad.

El estadístico $\dfrac{\bar{X}-\mu_X}{\frac{S_{n-1}}{\sqrt{n}}}$ se utiliza para construir intervalos de confianza y realizar pruebas de hipótesis con respecto a la media de una distribución normal.

Es importante observar que hasta ahora se han estudiado dos estadísticos relacionados con la distribución muestral de la media:

Del teorema central del límite

$$\frac{\bar{X}-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0,1)$$

cuando σ_X es conocida, o bien, cuando σ_X se aproxima mediante S_{n-1} para muestras grandes.

Y de la definición de la distribución t de Student

$$\frac{\bar{X}-\mu_X}{\frac{S_{n-1}}{\sqrt{n}}} \sim t_{(n-1)}$$

cuando los datos provienen de una distribución normal con σ_X desconocida y $n < 30$.

))

Ejemplo 2.8

Sea T una variable aleatoria con distribución t de Student, con 19 grados de libertad, obtener:

a) $P(T > 1.33)$

b) El valor de t, tal que $P(T < t) = 0.01$

Resolución

a) Del enunciado $T \sim t_{(19)}$, por lo que, directamente de tablas en el renglón de 19 grados de libertad se busca el valor más cercano a 1.33 y se lee en la parte superior el valor de α, teniéndose:
$P(T > 1.33) = 0.1$

b) De tablas, y recordando que la distribución t es simétrica, se tiene que
$0.01 = P(T < t) = P(T > -t)$
Al localizar $\alpha = 0.01$ y $\nu = 19$ se obtiene que: $-t = 2.54$
Finalmente: $t = -2.54$

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.9

Los siguientes seis datos son los tiempos de permanencia (espera y atención) en un banco:
15 , 32 , 18 , 26 , 27 y 20
Si el banco afirma que el tiempo promedio de permanencia es de 20 minutos o menos, determinar si la afirmación es razonable.

Resolución

De los datos $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \bar{x} = \frac{1}{6} \sum_{i=1}^{6} x_i = 23$

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$s_{n-1}^2 = \frac{1}{6-1} \sum_{i=1}^{6} (x_i - \bar{x})^2 = (6.387)^2$$

Las probabilidades se calculan para variables aleatorias, en este caso la variable aleatoria es la media muestral, la cual se tiene que comparar contra algún valor que tome (o que pueda tomar). Debe calcularse la probabilidad para un intervalo, puesto que la probabilidad de que una variable aleatoria continua tome un valor puntual es cero. La probabilidad, además, es condicional, puesto que se considera como verdadero el valor del parámetro objetivo.
De manera que pueden plantearse las probabilidades:

$$P(\bar{X} \geq \bar{x} \mid \mu_X = \mu_0) \quad \text{o} \quad P(\bar{X} \leq \bar{x} \mid \mu_X = \mu_0)$$

Con ambas probabilidades puede concluirse; sin embargo, es más conveniente utilizar aquella probabilidad en la cual se plantee que la variable aleatoria tome valores iguales o más alejados con respecto a la media hipotética, de esta manera se tiene:

$$P(\bar{X} \geq 23 \mid \mu_X = 20) = P\left(\frac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} \geq \frac{23 - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} \right)$$

pero $\dfrac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} = T \sim t_{(n-1)}$

valuando

$$P\left(\bar{X} \geq 23\right) = P\left(T \geq \frac{23 - 20}{\frac{6.387}{\sqrt{6}}} \right) = P\left(T \geq 1.15 \right)$$

y de tablas, con 5 grados de libertad

$$0.1 < P\left(T \geq 1.15\right) < 0.2$$

por lo que los datos no proporcionan una fuerte evidencia de que el banco no tenga razón, es decir, la afirmación del banco puede ser razonable.

S))))))))))))))))))))))))))))))))))))))

DISTRIBUCIÓN PARA CARACTERIZAR LA DIFERENCIA DE MEDIAS, VARIANZAS DESCONOCIDAS: T

Cuando se realizan muestreos sobre poblaciones diferentes, las poblaciones son normales, las muestras son pequeñas, con variancias desconocidas pero iguales, entonces se utiliza la distribución t para caracterizar la diferencia de medias muestrales, teniéndose el estadístico:

$$T = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

donde

$$S_p^2 = \frac{\left(n_X - 1\right) S_X^2 + \left(n_Y - 1\right) S_Y^2}{n_X + n_Y - 2}$$

y

$$T \sim t_{\left(n_1 + n_2 - 2\right)}$$

el estadístico S_P^2 recibe el nombre de *estimador combinado de la variancia* (o estimador ponderado de la variancia). Evidentemente S_P^2 proporciona un valor entre las variancias muestrales de las poblaciones S_X^2 y S_Y^2, pero no es el punto medio entre ambas.

Si las muestras son pequeñas, extraídas de poblaciones normales, con variancias desconocidas y diferentes puede utilizarse una aproximación mediante la distribución t, la cual se estudiará en temas posteriores. Cualquier otro caso queda fuera del alcance de este curso.

S))Q

Ejemplo 2.10

Una organización independiente está interesada en probar la distancia de frenado a una velocidad de 50 [km/h] para dos marcas distintas de automóviles. Para la primera marca se seleccionaron nueve automóviles y se probaron en un medio controlado. La media muestral y la desviación estándar fueron de 145 [m] y 8 [m], respectivamente. Para la segunda marca se seleccionaron 12 automóviles y la distancia promedio resultó ser de 132 [m] y una desviación estándar de 10 [m]. Con base en esta evidencia, ¿existe alguna razón para creer que la distancia de frenado para ambas marcas, es la misma? Supóngase que las distancias de frenado son variables aleatorias independientes normalmente distribuidas con variancias iguales.

Resolución

De los datos del enunciado se tiene:

$\bar{x}_1 = 145 \quad , \quad \bar{x}_2 = 132$

$s_1 = 8 \quad , \quad s_2 = 10$

$n_1 = 9 \quad , \quad n_2 = 12$

$$P(\bar{X}_1 - \bar{X}_2 \geq 13 \mid \mu_1 = \mu_2, \sigma_1 = \sigma_2)$$

Y puesto que X_1 y X_2 tienen distribuciones normales con la misma variancia, entonces:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 13) = P(T \geq 3.2) \approx 0.002$$

puesto que $t = \dfrac{13 - 0}{9.211\sqrt{\frac{1}{9} + \frac{1}{12}}} = 3.2$

con $s_p^2 = \dfrac{8(64) + 11(100)}{19} = 84.8421, \qquad s_p = 9.21$

Y puesto que $P(T \geq 3.2) \approx 0.002$, es muy poco probable suponer que las distancias de frenado sean iguales.

S))Q

Si las muestras son pequeñas, extraídas de poblaciones normales, con variancias desconocidas y diferentes puede utilizarse una aproximación mediante el siguiente estadístico:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(\nu)}$$

el cual tiene una distribución aproximadamente t, con ν grados de libertad, los cuales se aproximan mediante:

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1}\right] + \left[\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

aproximando al máximo entero, es decir, hacia el entero inferior más cercano.

O bien mediante

$$\nu \approx \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\left[\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 + 1}\right] + \left[\dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 + 1}\right]} - 2$$

aproximando al entero más cercano.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 1.11

Un fabricante de unidades de pantallas de video prueba dos diseños de microcircuitos para determinar si ellos producen flujos de corriente equivalentes. Ingeniería de desarrollo ha obtenido los siguientes datos:

Diseño 1	$n_1 = 15$	$\bar{x}_1 = 24.2$	$s_1^2 = 10$
Diseño 2	$n_2 = 10$	$\bar{x}_2 = 23.9$	$s_2^2 = 20$

Si se supone que ambas poblaciones son normales, pero no estamos dispuestos a considerar que las variancias desconocidas σ_1^2 y σ_2^2 son iguales. ¿Qué tan probable es el resultado muestral observado?

Resolución

Se desea calcular

$$P\left(\bar{X}_1 - \bar{X}_2 \geq \bar{x}_1 - \bar{x}_2 \;\middle|\; \mu_1 = \mu_2, \sigma_1 \neq \sigma_2\right)$$

El estadístico está dado por

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t_{(\nu)}$$

$$t = \frac{(0.3) - 0}{\sqrt{\dfrac{10}{15} + \dfrac{20}{10}}} = 0.1837$$

$$P\left(\bar{X}_1 - \bar{X}_2 \geq 0.3\right) = P\left(T \geq 0.1837\right)$$

Para obtener los grados de libertad

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_1^2}{n_2}\right)^2}{n_2 + 1}\right]} - 2$$

$$\nu \approx \frac{\left(\frac{10}{15} + \frac{20}{10}\right)^2}{\left[\frac{\left(\frac{10}{15}\right)^2}{15 + 1} + \frac{\left(\frac{20}{10}\right)}{10 + 1}\right]} - 2$$

$$\nu \approx = 16.1676 \approx 16$$

Por lo que $T \sim t_{(16)}$

Entonces:

$$P\left(T \geq 0.1837\right) \approx 0.428$$

Por lo que puede considerarse que los flujos de corriente son equivalentes $\left(\mu_1 = \mu_2\right)$

DISTRIBUCIÓN PARA CARACTERIZAR A LA RAZÓN DE VARIANZAS: DISTRIBUCIÓN F

La distribución F (de Fisher) se utiliza cuando se desean hacer inferencias con respecto a las variancias de dos distribuciones normales independientes a partir de muestras aleatorias de cada distribución.

Definición 2.7

Si X y Y son dos variables aleatorias independientes con distribuciones ji cuadrada con parámetros u y v, es decir:

$$X \sim \chi^2_{(u)} \quad , \quad Y \sim \chi^2_{(v)}$$

Entonces la variable aleatoria F definida como

$$F = \frac{\frac{X}{u}}{\frac{Y}{v}}$$

tiene una distribución F de Fisher con u grados de libertad en el numerador y v grados de libertad en el denominador.

Se denota $F \sim F_{(u,v)}$

La media y la variancia de la distribución F se proporcionan en el siguiente teorema.

Teorema 2.9

Si F es una variable aleatoria con distribución F de Fisher con u grados de libertad en el numerador y v grados de libertad en el denominador, entonces:

$$\mathrm{E}(F) = \frac{v}{v-2} \qquad v > 2$$

$$\mathrm{Var}(F) = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \qquad v > 4$$

La distribución F es asimétrica y sesgada hacia la derecha (sesgo positivo) según se observa en la siguiente figura

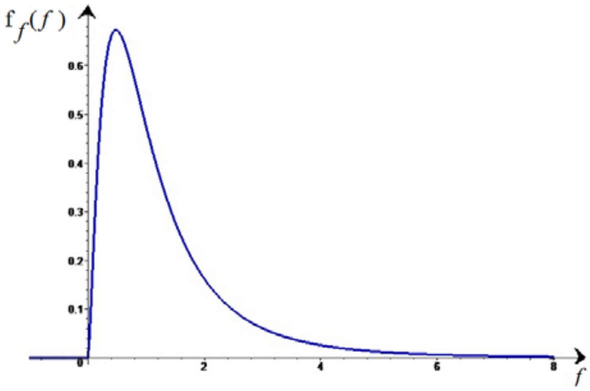


Fig. 2.4. Distribución F de Fisher

Para entender el uso de la distribución F, considérense dos muestras aleatorias independientes de tamaños n_X y n_Y respectivamente, de tal forma que para la muestra $X_1, X_2, \ldots, X_{n_x}$ cada variable tiene distribución normal con parámetros μ_X y σ_X^2 ; y para la muestra $Y_1, Y_2, \ldots, Y_{n_Y}$ cada variable tiene distribución normal con parámetros μ_Y y σ_Y^2 entonces los estadísticos

$$\frac{(n_X-1)\,S_X^2}{\sigma_X^2} \quad \text{y} \quad \frac{(n_Y-1)\,S_Y^2}{\sigma_Y^2}$$

son variables aleatorias Ji cuadradas independientes con $n_X - 1$ y $n_Y - 1$ grados de libertad. Y de la definición 4.3 se tiene que

$$\frac{\dfrac{\dfrac{(n_X-1)\,S_X^2}{\sigma_X^2}}{n_X-1}}{\dfrac{\dfrac{(n_Y-1)\,S_Y^2}{\sigma_Y^2}}{n_Y-1}} = \frac{\dfrac{S_X^2}{\sigma_X^2}}{\dfrac{S_Y^2}{\sigma_Y^2}} \sim F_{(n_X-1,\,n_Y-1)}$$

Es decir, $\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ tiene distribución F con $n_X - 1$ y $n_Y - 1$ grados de libertad en el numerador y denominador, respectivamente.

Un caso particular se tiene cuando se considera que las variancias de las poblaciones son iguales, esto es, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, teniéndose que:

De la expresión $\dfrac{\dfrac{S_X^2}{\sigma_X^2}}{\dfrac{S_Y^2}{\sigma_Y^2}}$ cuando $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

entonces $\dfrac{S_X^2}{S_Y^2} \sim F_{(n_X-1,\, n_Y-1)}$

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.12

Si dos muestras aleatorias independientes de tamaño $n_1 = 9$ y $n_2 = 16$ se extraen de una población normal, determinar la probabilidad de que la variancia de la primera sea al menos cuatro veces más grande que la segunda.

Resolución

Del enunciado se pregunta

$$P\left(\frac{S_1^2}{S_2^2} > 4 \;\middle|\; \sigma_1^2 = \sigma_2^2\right)$$

Se considera la misma variancia puesto que las muestras se extraen de la misma población, y dado que las muestras son independientes y provienen de una población normal, entonces

$$\frac{S_1^2}{S_2^2} \sim F_{(n_1-1,\, n_2-2)}$$

en particular

$$\frac{S_1^2}{S_2^2} \sim F_{(8,15)}$$

Por lo que, de tablas

$$P\left(\frac{S_1^2}{S_2^2} > 4\right) = 0.01$$

La probabilidad es de 0.01

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.13

Si S_1^2 y S_2^2 representan las variancias de muestras aleatorias independientes de tamaño $n_1 = 25$ y $n_2 = 31$, que se toman de poblaciones normales con variancias $\sigma_1^2 = 10$ y $\sigma_2^2 = 15$, respectivamente, encontrar la $P\left(\frac{S_1^2}{S_2^2} > 1.26\right)$.

Resolución

Se sabe que $F = \dfrac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F_{(n_1 - 1, n_2 - 1)}$ y del enunciado $\sigma_1^2 = 10$, $\sigma_2^2 = 15$.

Por lo que

$$P\left(\frac{S_1^2}{S_2^2} > 1.26\right) = P\left(\dfrac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} > \frac{15(1.26)}{10}\right)$$

$$= P(F > 1.89)$$

Y de tablas con $F \sim F_{(24,30)}$

$$P\left(\frac{S_1^2}{S_2^2} > 1.26\right) \approx 0.05$$

S))

DISTRIBUCIÓN MUESTRAL PARA CARACTERIZAR A UNA PROPORCIÓN: DISTRIBUCIÓN NORMAL

Si la muestra proviene de una población Bernoulli, en la cual nos interesa caracterizar la probabilidad de éxito p, llamada proporción, entonces se requiere que el tamaño sea grande para poder utilizar el teorema del límite central, con lo que una proporción se caracterizará con la distribución normal. Esto es, si se desea caracterizar el estadístico para una proporción,

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

donde p es el parámetro de una población Bernoulli, Y es el número de éxitos que se observan en la muestra de tamaño n (variable aleatoria binomial); se utiliza la distribución normal cuando la muestra es grande.

Además, puesto que la distribución Bernoulli es discreta y la normal es continua, debe realizarse un ajuste por continuidad, que es especialmente necesario cuando el tamaño de la muestra no es tan grande. El *factor de corrección por continuidad*, ***FCC***, se define como:

$$FCC = \frac{1}{2n}$$

Obsérvese como el ajuste es mediante el término $\frac{1}{2n}$, en lugar del término $\frac{1}{2}$ utilizado en probabilidad, debido a que una proporción de éxitos es el número de éxitos divididos entre n.

Para calcular una probabilidad utilizando la aproximación normal y el factor de ajuste por continuidad se utiliza entonces la expresión:

$$P(\hat{P} \le \hat{p}) \approx P\left(Z \le \frac{\hat{p} + \frac{1}{2n} - p}{\sigma_{\hat{P}}} \right)$$

donde $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$.

S)))))))))))))))))))))))))))))))))))))))

Ejemplo 2.14

Se procede a detener el funcionamiento de una máquina para repararla si en una muestra aleatoria de 100 artículos de la producción diaria de la máquina se encuentran por lo menos 15% de artículos defectuosos. (Suponer que la producción diaria consta de un gran número de artículos). Si realmente la máquina produce sólo 10% de artículos defectuosos, encontrar la probabilidad de que se pare la máquina un día dado. (Utilizar la corrección por continuidad).

Resolución

Puesto que la producción diaria consta de un gran número de artículos, se considera que la población es infinita, y si Y representa el número de artículos defectuosos en la muestra de 100, entonces:

$$P\left(\frac{Y}{n} \ge 0.15 \right) = P\left(\frac{\left(\frac{Y}{n}\right) - p}{\sqrt{\frac{pq}{n}}} \ge \frac{0.15 - \frac{1}{2(100)} - 0.10}{\sqrt{\frac{(0.1)\,(0.9)}{100}}} \right)$$

$$P\left(\frac{Y}{n} \ge 0.15 \right) \simeq P(Z > 1.50)$$

$$P\left(\frac{Y}{n} \ge 0.15 \right) \approx 0.0668$$

S)))))))))))))))))))))))))))))))))))))))

DIFERENCIA DE PROPORCIONES

Definición 2.8

Si dos muestras independientes de tamaño n_X y n_Y se extraen de poblaciones infinitas con distribuciones binomiales, X representa el número de observaciones de la primera muestra que corresponden a la clase de interés, y Y representa el número de observaciones de la segunda muestra que corresponden a la clase en cuestión, entonces la distribución de muestreo para la diferencia de proporciones está dada por

$$Z = \frac{(\hat{P}_X - \hat{P}_Y) - (p_X - p_Y)}{\sqrt{\frac{p_X(1 - p_X)}{n_X} + \frac{p_Y(1 - p_Y)}{n_Y}}}$$

donde $Z \sim N(0,1)$

S))))))))))))))))))))))))))))))))))))))

Ejemplo 2.15

Se están considerando dos tipos diferentes de computadoras de control de disparo que se utilizaron en baterías de 6 cañones de 105 mm del ejército de los Estados Unidos. Los dos sistemas de computadoras se someten a una prueba operacional en la cual se cuenta el número total de impactos en el blanco. El sistema de computadora 1 produce 250 impactos de 300 descargas, en tanto que el sistema 2 consigue 178 impactos de 260 descargas. ¿Hay alguna razón para pensar que los dos sistemas de computadora difieren?

Resolución

Se desea calcular

$$P\left(\hat{P}_X - \hat{P}_Y \geq p_X - p_Y \;\middle|\; p_X = p_Y\right)$$

$$\hat{p}_X = \frac{250}{300} \approx 0.8333 \; ; \qquad n_1 = 300$$

$$\hat{p}_Y = \frac{178}{260} \approx 0.6846 \; ; \qquad n_1 = 260$$

El estadístico está dado por

$$Z = \frac{\left(\hat{P}_X - \hat{P}_Y\right) - \left(p_X - p_Y\right)}{\sqrt{\frac{\hat{p}_X\left(1 - \hat{p}_X\right)}{n_X} + \frac{\hat{p}_Y\left(1 - \hat{P}_Y\right)}{n_Y}}} \sim N(0,1)$$

$$z = \frac{(0.8333 - 0.6846) - 0}{\sqrt{\frac{0.8333(1 - 0.8333)}{300} + \frac{0.6846(1 - 0.6846)}{260}}}$$

$z = 4.1353$

Entonces:

$$P\left(\hat{P}_X - \hat{P}_Y \geq 0.1487\right) = P(Z \geq 4.1353)$$
$$= 1 - P(Z \leq 4.1353)$$
$$\approx 0$$

Debido a que la probabilidad es prácticamente cero, podemos decir que hay una diferencia significativa en los dos sistemas de computadora.

S))))))))))))))))))))))))))))))))))))Q

Resumen

Distribuciones de X_i , Y_i	Estimador	Distribución del estimador
$X_i \sim N(\mu_X, \sigma_X^2)$ σ_X^2 conocida	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$
$X_i \sim N(\mu_X, \sigma_X^2)$ σ_X^2 desconocida	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\frac{\bar{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} \sim t_{(n-1)}$
Teorema central del límite $n \geq 30$ $X_i \sim$ Cualquiera con parámetros: $E(X_i) = \mu_X$ $Var(X_i) = \sigma_X^2$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$
$X_i \sim N(\mu_X, \sigma_X^2)$	$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	$\frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2_{(n-1)}$
$X_i \sim N(\mu_X, \sigma_X^2)$	$\frac{S_X^2}{S_Y}$	$F_{(n_X-1, n_Y-1)}$

BIBLIOGRAFÍA

Hines, William W. y Montgomery, Douglas C. - Probabilidad y Estadística para ingeniería, cuarta edición..- CECSA.- México, 2005.

Milton, Susan J. Y Arnold, Jesse C.- Probabilidad y Estadística para con aplicaciones para ingeniería y ciencias computacionales, cuarta edición.- McGraw-Hill.- México, 2004.

Devore, Jay L.- Probabilidad y Estadística para ingeniería y ciencias, séptima edición.- Cengage Learning.- México, 2008.

Mendenhall, William III. et al.- Introducción a la Probabilidad y Estadística.- Décimo cuarta edición.- Cengage Learning.- México 2015.

Wackerly Dennis D.- Mendenhall, William, *et al.*- Estadística Matemática con Aplicaciones, sexta edición.- Editorial Thomson.- México, 2002.

Walpole, Ronald E., *et al.*- Probability and Statistics for Engineers and Scientists.- Pearson.- USA, 2007.

Montgomery, Douglas C. y Runger, George C.-Probabilidad y Estadística aplicadas a la Ingeniería, segunda edición.- Limusa-Wiley.- México, 2002.

Scheaffer, Richard L. y McClave, James T.- Probabilidad y Estadística para Ingeniería.- Grupo Editorial Iberoamérica.- México, 1993.

Canavos, George C.- Probabilidad y Estadística Aplicaciones y Métodos.- McGraw-Hill.- México, 1988.

Meyer, Paul L.- Probabilidad y Aplicaciones Estadísticas.- Addison Wesley Iberoamericana.- México, 1992.

Spiegel, Murray R. et al.- Probabilidad y Estadística, cuarta edición.- Mc Graw-Hill.-México 2013.

Borras García, Hugo E., *et al.*- Apuntes de Probabilidad y Estadística.-Facultad de Ingeniería.- México, 1985.

Rosenkrantz, Walter A.- Introduction to Probability and Statistics for Scientists and Engineers.- McGraw-Hill.- EE.UU., 1997.

Ziemer, Rodger E.- Elements of Engineering Probability & Statistics.- Prentice Hall.- USA 1997.

ESTADÍSTICA

Profesores:

A. Leonardo Bañuelos Saucedo

Nayelli Manzanarez Gómez

NOTAS

TEMA 3

ESTIMACIÓN DE PARÁMETROS

TEMA III
ESTIMACIÓN DE PARÁMETROS

La estimación puntual de un parámetro relativo a una población es el valor numérico de un estadístico correspondiente a ese parámetro.

En la elección de un estimador deben tenerse en cuenta las siguientes propiedades: *insesgabilidad, eficiencia, error cuadrático medio, consistencia y suficiencia*.

INSESGABILIDAD

Cuando se obtiene una estimación puntual de un parámetro cualquiera, es deseable que la distribución de dicha estimación se centre en el parámetro real (al cual se le llamará parámetro-objetivo),si se cumple la condición anterior entonces el estimador se llama insesgado.

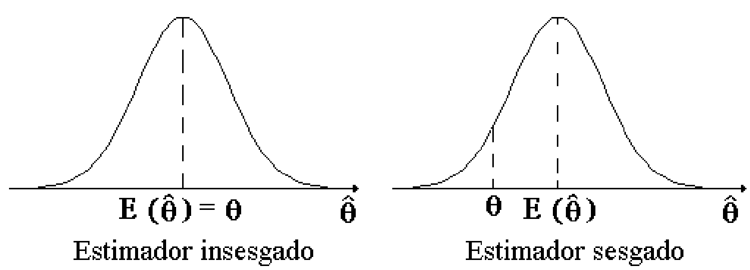


Fig. 3.1 Sesgo en la estimación

Definición 3.1

Sea $\hat{\Theta}$ un estimador puntual del parámetro θ. Entonces si $E(\hat{\Theta}) = \theta$ se dice que $\hat{\Theta}$ es un estimador insesgado de θ , de lo contrario se dice que es sesgado.

S))

Ejemplo 3.1

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n extraída de una población con media μ y variancia σ^2 . Determinar si los siguientes estimadores son sesgados o insesgados.

a) $\bar{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

b) $S_n^2 = \dfrac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

c) $S_{n-1}^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Resolución

a) Para determinar si tiene o no sesgo debe obtenerse $E(\hat{\Theta})$.

$$E(\hat{\Theta}) = E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$E(\hat{\Theta}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu$$

$E(\bar{X}) = \mu$ $\qquad\qquad$ $\therefore$ es insesgado.

b) $$E(S_n^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$E(S_n^2) = \frac{1}{n} E\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right]$$

$$E(S_n^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2)\right]$$

$$E(S_n^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = \frac{1}{n} \mathrm{E}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

$$E(S_n^2) = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right]$$

Recordando que $Var(X) = E(X^2) - [E(X)]^2$

entonces: $\qquad E(X^2) = Var(X) + \mu^2 = \sigma^2 + \mu^2$

y de forma similar

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

$$E(S_n^2) = \frac{1}{n}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right]$$

$$E(S_n^2) = \frac{1}{n}[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2]$$

$$E(S_n^2) = \frac{1}{n}[\sigma^2(n-1)] = \frac{n-1}{n}\sigma^2$$

$E(S_n^2) = \dfrac{n-1}{n}\sigma^2$ $\qquad\qquad$ $\therefore$ Es sesgado.

Otra forma de calcular si el estadístico S_n^2 es insesgado, es a través de la variable aleatoria ji cuadrada.

Se desea obtener $\mathrm{E}\left(S_n^2\right)$,

pero se sabe que, para una v.a. ji cuadrada $\mathrm{E}(\mathrm{X}^2) = \nu$, si $\mathrm{X}^2 \sim \chi^2_{(\nu)}$.

De donde

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

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

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

c) $$E(S_{n-1}^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

Pero

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

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

$$E(S_{n-1}^2) = \sigma^2$$ $\therefore$ Es insesgado.

S))

En la práctica se suelen preferir los estimadores insesgados sobre los sesgados; por ello que cuando se desean hacer estimaciones con respecto a la variancia σ^2 de una población se utiliza el estadístico S_{n-1}^2.

La siguiente tabla muestra algunos de los parámetros objetivos más comunes, juntos con sus valores esperados y sus variancias.

Tabla 3.1. Valores esperados y variancias de estimadores basados en muestras grandes

Parámetro-objetivo θ	Tamaño de las muestras	Estimador puntual $\hat{\Theta}$	$E(\hat{\Theta})$	$Var(\hat{\Theta})$
μ	n	$\bar{X}$	μ	$\frac{\sigma^2}{n}$
p	n	$\hat{p} = \frac{X}{n}$	p	$\frac{p\,q}{n}$
$\mu_1 - \mu_2$	n_1 y n_2	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
$p_1 - p_2$	n_1 y n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

EFICIENCIA

Puesto que es posible obtener más de un estimador insesgado para el mismo parámetro objetivo, deberá utilizarse el de mínima variancia, que recibe el nombre de estimador eficiente.

Definición 3.2

Sean $\hat{\Theta}_1$ y $\hat{\Theta}_2$ dos estimadores insesgados del parámetro θ, con variancias $Var(\hat{\Theta}_1)$ y $Var(\hat{\Theta}_2)$, respectivamente, entonces el la eficiencia relativa de $\hat{\Theta}_1$ con respecto de $\hat{\Theta}_2$, η, se define como:

$$\eta = \frac{\mathrm{Var}(\hat{\Theta}_2)}{\mathrm{Var}(\hat{\Theta}_1)}$$

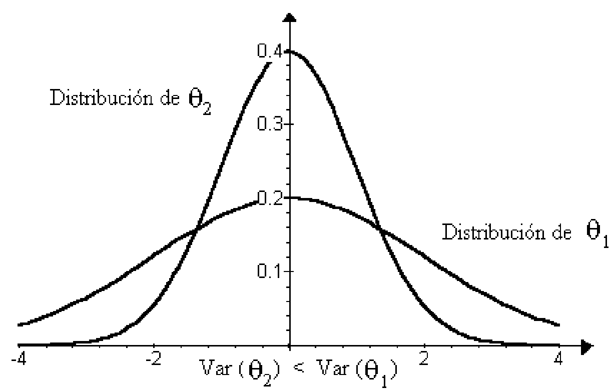


Fig. 3.2

En la figura 3.2 se observan las distribuciones de los estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ del parámetro θ, considerando que ambos estimadores son insesgados, se prefiere al estadístico $\hat{\Theta}_2$ porque tiene menor varianza y esto repercute en estimaciones con menos variabilidad. Al estimador con menor varianza se le llama eficiente.

La eficiencia relativa, no es una eficiencia del tipo mecánico, esto es, η puede ser mayor que uno cuando $Var(\hat{\Theta}_1) < Var(\hat{\Theta}_2)$, evidentemente lo que se busca al calcular una eficiencia entre estimadores es una comparación de la mejoría entre uno y otro.

S))

Ejemplo 3.2

Supóngase que se tiene una muestra aleatoria de tamaño $2n$ de una población denotada por X y $E(X) = \mu$ y $Var(X) = \sigma^2$. Sean

$$\bar{X}_1 = \frac{1}{2n}\sum_{i=1}^{2n} X_i \quad \text{y} \quad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

dos estimadores de μ. Determina cuál es el mejor estimador de μ. Explicar la selección.

Resolución

$\bar{X}_1$ y $\bar{X}_2$ son estimadores insesgados; sin embargo $\text{Var}(\bar{X}_1) = \dfrac{\sigma^2}{2n}$ mientras que $\text{Var}(\bar{X}_2) = \dfrac{\sigma^2}{n}$, puesto que $\text{Var}(\bar{X}_1) < \text{Var}(\bar{X}_2)$ se concluye que $\bar{X}_1$ es un estimador más eficiente que $\bar{X}_2$ por lo que el mejor estimador es $\bar{X}_1$.

S))

ERROR CUADRÁTICO MEDIO

Cuando se desean comparar dos estimadores, de los cuales al menos uno no es insesgado, entonces la eficiencia relativa no se calcula como el cociente de las variancias, sino como el cociente de los errores cuadráticos medios, ECM.

Definición 3.3

El error cuadrático medio de un estimador $\hat{\Theta}$, del parámetro θ, se define como:

$$ECM(\hat{\Theta}) = E\left[\left(\hat{\Theta} - \theta\right)^2\right]$$

El error cuadrático medio también puede escribirse en términos de la variancia y del sesgo.

$$\begin{aligned} \text{ECM}(\hat{\Theta}) &= \text{E}\left[\left(\hat{\Theta} - \theta\right)^2\right] = \text{E}\left(\hat{\Theta}^2 - 2\,\theta\,\hat{\Theta} + \theta^2\right) \\ &= \text{E}(\hat{\Theta}^2) - 2\,\theta\,\text{E}(\hat{\Theta}) + \theta^2 \end{aligned}$$

sumando y restando $\left[\text{E}(\hat{\Theta})\right]^2$, se tiene:

$$\begin{aligned} \text{ECM}(\hat{\Theta}) &= \text{E}(\hat{\Theta}^2) - \left[\text{E}(\hat{\Theta})\right]^2 + \left[\text{E}(\hat{\Theta})\right]^2 - 2\,\theta\,\text{E}(\hat{\Theta}) + \theta^2 \\ &= \text{Var}(\hat{\Theta}) + \left(\text{E}(\hat{\Theta}) - \theta\right)^2 \\ &= \text{Var}(\hat{\Theta}) + \left(\theta - \text{E}(\hat{\Theta})\right)^2 \end{aligned}$$

donde a la cantidad $\theta - \text{E}(\hat{\Theta})$ se le llama ***sesgo***, o bien, error cometido, y se denota mediante la letra B, entonces:

$$\boxed{\text{ECM}(\hat{\Theta}) = \text{Var}(\hat{\Theta}) + \text{B}^2}$$

La eficiencia relativa de $\hat{\Theta}_2$ a $\hat{\Theta}_1$ se define como

$$\eta_{ECM} = \frac{\text{ECM}(\hat{\Theta}_1)}{\text{ECM}(\hat{\Theta}_2)}$$

si $\eta < 1$ entonces $\hat{\Theta}_1$ es mejor estimador que $\hat{\Theta}_2$.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.3

Supóngase que $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$ son estimadores del parámetro θ. Si se sabe que $E(\hat{\Theta}_1) = E(\hat{\Theta}_2) = \theta$, $E(\hat{\Theta}_3) \neq \theta$, $Var(\hat{\Theta}_1) = 12$, $Var(\hat{\Theta}_2) = 10$ y $E[(\hat{\Theta}_3 - \theta)^2] = 6$, utilizando el criterio del error cuadrático medio, determinar el mejor estimador.

Resolución

Para los primeros dos estimadores, se tiene que el error cuadrático medio, ECM, es:

$$\mathrm{ECM}(\hat{\Theta}_1) = \mathrm{Var}(\hat{\Theta}_1) = 12$$

$$\mathrm{ECM}(\hat{\Theta}_2) = \mathrm{Var}(\hat{\Theta}_2) = 10$$

y para el tercer estimador

$$\mathrm{ECM}(\hat{\Theta}_3) = \mathrm{E}[(\hat{\Theta}_3 - \theta)^2] = 6$$

Por lo que el mejor estimador es $\hat{\Theta}_3$, puesto que tiene menor error cuadrático medio.

S))))))))))))))))))))))))))))))))))))))

CONSISTENCIA

Mientras mayor sea el tamaño de la muestra, la estimación deberá ser más precisa. Si el estimador cumple con la característica anterior entonces se llama consistente. Por ejemplo $\bar{X}$ es un estimador consistente de μ.

Definición 3.4

El estimador $\hat{\Theta}_n$ es consistente al estimar a θ si para cualquier $\xi > 0$ se cumple:

$$\lim_{n \to \infty} P(|\hat{\Theta}_n - \theta| \leq \xi) = 1$$

Para un estimador insesgado se puede probar la consistencia evaluando el límite de la variancia cuando $n \to \infty$, i.e. un estimador insesgado de θ es consistente si:

$$\lim_{n \to \infty} Var(\hat{\Theta}_n) = 0$$

Cuando el estimador es sesgado, debe probarse que

$$\lim_{n \to \infty} E[(\hat{\Theta}_n - \theta)^2] = 0$$

para que $\hat{\Theta}$ sea consistente. Esto es, el límite del error cuadrático medio cuando n tiende a infinito debe ser cero. Existen estimadores consistentes que son sesgados.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.4

Sea $Y_1 , Y_2 , . . . , Y_n$ una muestra aleatoria de una población con media μ_Y y variancia σ_Y^2. Considérense los tres estimadores siguientes para μ_Y:

$$\hat{\mu}_1 = \frac{1}{2}(Y_1 + Y_2)$$

$$\hat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + . . . + Y_{n-1}}{2(n - 2)} + \frac{1}{4}Y_n \qquad \text{y}$$

$$\hat{\mu}_3 = \bar{Y}$$

Determinar si los estimadores son consistentes para μ_Y.

Resolución

Puesto que $Y_1 , Y_2 , . . . , Y_n$ forman una muestra aleatoria de una población con media μ_Y y variancia σ_Y^2, entonces $E(Y_i) = \mu_Y \;, \; i = 1 , 2 , . . . , n$, y lo primero que se prueba es el insesgamiento de los estimadores.

$$E(\hat{\mu}_1) = E\left[\frac{1}{2}(Y_1 + Y_2) \right] = \frac{1}{2} [E(Y_1) + E(Y_2)]$$

$$= \frac{1}{2}(\mu_Y + \mu_Y) = \mu_Y \;\therefore\; \hat{\mu}_1 \text{ es insesgado}$$

$$E(\hat{\mu}_2) = E\left[\frac{1}{4}Y_1 + \frac{Y_2 + . . . + Y_{n-1}}{2(n - 2)} + \frac{1}{4}Y_n \right]$$

$$= \frac{1}{4}\mu_Y + \frac{(n - 2)\mu_Y}{2(n - 2)} + \frac{1}{4}\mu_Y = \mu_Y \;\therefore\; \hat{\mu}_2 \text{ es insesgado}$$

$$E(\hat{\mu}_3) = E(\bar{Y})$$

$$= \frac{n\mu_Y}{n} = \mu_Y \;\therefore\; \hat{\mu}_3 \text{ es insesgado.}$$

Los tres estimadores son insesgados.

Para las variancias, puesto que $\text{Var}(Y_i) = \sigma_Y^2 \; i = 1 , 2 , . . . , n$, entonces:

$$\text{Var}(\hat{\mu}_1) = \text{Var}\left[\frac{1}{2}(Y_1 + Y_2) \right] = \frac{1}{4}(\sigma_Y^2 + \sigma_Y^2) = \frac{\sigma_Y^2}{2}$$

$$\text{Var}(\hat{\mu}_2) = \frac{1}{16}\sigma_Y^2 + \frac{(n - 2)\sigma_Y^2}{4(n - 2)^2} + \frac{1}{16}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{8} + \frac{\sigma_Y^2}{4(n-2)}$$

$$\mathrm{Var}(\hat{\mu}_3) = \mathrm{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Por lo que:

$$\lim_{n\to\infty} \mathrm{Var}(\hat{\mu}_1) = \frac{\sigma_Y^2}{2}$$

$$\lim_{n\to\infty} \mathrm{Var}(\hat{\mu}_2) = \frac{\sigma_Y^2}{8}$$

$$\lim_{n\to\infty} \mathrm{Var}(\hat{\mu}_3) = \lim_{n\to\infty} \frac{\sigma_Y^2}{n} = 0$$

Y el único estimador consistente para μ es $\hat{\mu}_3$.

S))))))))))))))))))))))))))))))))))))))

SUFICIENCIA

Un estimador es suficiente si toma en cuenta toda la información de la muestra.

Definición 3.5

Sea θ un parámetro desconocido y $X_1, X_2, \ldots, X_n$ una muestra aleatoria. Entonces el estadístico

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

es suficiente para θ si la distribución condicional de $X_1, X_2, \ldots, X_n$ dado $\hat{\theta}$ no depende de θ.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.5

Considérese una muestra aleatoria de tamaño n : $X_1, X_2, \ldots, X_n$ de una población Bernoulli con parámetro p. Determinar si $Y = \frac{1}{n} \sum_{i=1}^{n} X_i$ es un estimador suficiente para p o no.

Resolución

$$P\left(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n \ \middle|\ \frac{1}{n}\sum_{i=1}^{n} X_i = k\right)$$

$$= \frac{P\left(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n,\ \frac{1}{n}\sum_{i=1}^{n} X_i = k\right)}{P\left(\frac{1}{n}\sum_{i=1}^{n} X_i = k\right)}$$

$$= \frac{p^{nk}(1-p)^{n-nk}}{\binom{n}{nk} p^{nk}(1-P)^{n-nk}} = \frac{1}{\binom{n}{nk}} \qquad \text{No depende de } p$$

$\therefore \quad Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ es un estimador suficiente para p .

S))

MÉTODOS PARA DETERMINAR ESTIMADORES PUNTUALES

Anteriormente se explicaron las características deseables para un estimador, en esta sección se verá la forma de obtenerlas. Como debe intuirse, existen varias formas de estimar un parámetro, por lo que no debe ser una sorpresa el hecho de que existan varios métodos para determinar los estimadores. Dos de los métodos más comunes son: el de los momentos y el de máxima verosimilitud.

MÉTODOS DE LOS MOMENTOS

El método de los momentos sugiere utilizar como estimador de alguno de los momentos de la población, al mismo momento con respecto a la muestra.

Definición 3.6 Método de los momentos

Elegir como estimadores puntuales, a aquellos valores de los parámetros que sean solución de las ecuaciones

$$\mu_k^{\prime} = m_k^{\prime} \qquad ; \qquad k = 1, 2, \ldots, n$$

donde n es igual al número de parámetros a estimar y $\mu_k^{\prime}$ y $m_k^{\prime}$ representan los momentos con respecto al origen de la población y de la muestra, respectivamente.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.6

Sea X una v.a. con distribución normal y parámetros μ y σ^2 desconocidos. Determinar los estimadores de dichos parámetros por el método de los momentos.

Resolución

Puesto que se buscan dos parámetros se requieren dos momentos. La media de μ es el primer momento con respecto al origen, y la variancia σ^2 es el segundo momento con respecto a la media, pero que puede expresarse a través de momentos con respecto al origen.
Para la media.

$$\hat{\mu} = \mu_1^{/} = m_1^{/} \quad \Rightarrow$$

El estimador es

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

y una estimación está dada por $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Para la variancia, se utilizan los segundos momentos con respecto a la media, por lo que

$$\mu_2 = \; = \hat{\sigma}^2 = m_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

El estimador es

$$\hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

y una estimación está dada por

$$s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

S))))))))))))))))))))))))))))))))))))))

Debe observarse que el método de los momentos proporciona el estimador sesgado de la varianza.

En la práctica, es mucho más sencillo igualar momentos con respecto a la media. Es decir, se pueden igualar momentos con respecto al origen, o bien, momentos con respecto a la media, respectivamente, según sea más conveniente.

En el ejemplo anterior basta entonces con igualar los primeros momentos con respecto al origen (poblacional y muestral, respectivamente):

$$\mu_1^{/} = m_1^{/}$$

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} X_i$$

y los segundos momentos con respecto a la media (poblacional y muestral, respectivamente)

$$\mu_2 = m_2$$

$$\hat{\sigma}_X^2 = S_n^{\,2}$$

S))

Ejemplo 3.7

Sea Y una v.a. con distribución de Pascal (binomial negativa) con parámetros r y p desconocidos. Utilizar el método de los momentos para obtener estimadores de dichos parámetros.

Resolución

Puesto que $Y \sim Pascal(r, p)$

entonces los momentos poblacionales son:

$$\mu_1^{/} = \mathrm{E}(Y) = \frac{r}{p}$$

$$\mu_2 = \mathrm{Var}(Y) = \frac{r\,q}{p^2} = \frac{r(1-p)}{p^2}$$

Y los momentos muestrales son:

$$m_1^{/} = \bar{Y}$$

$$m_2 = S_n^2$$

por lo que al igualar con los momentos poblacionales con los muestrales se tiene:

$$\bar{Y} = \frac{\hat{r}}{\hat{p}} \qquad \text{. . . (a)}$$

$$S_n^2 = \frac{\hat{r}(1-\hat{p})}{\hat{p}^2} \qquad \text{. . . (b)}$$

Resolviendo simultáneamente para $\hat{r}$ y $\hat{p}$,
de (b)

$$\hat{p}^2 S_n^2 = \hat{r}(1-\hat{p}) \qquad \text{. . . (c)}$$

pero de (a) $\hat{r} = \hat{p}\,\bar{Y}$
y sustituyendo en (c)

$$\hat{p}^2 S_n^2 = \hat{p}\,\bar{Y}(1-\hat{p})$$

$$\hat{p}\,S_n^2 = \bar{Y}(1-\hat{p}) = \bar{Y} - \bar{Y}\hat{p}$$

$$\hat{p}\,(S_n^2 + \bar{Y}) = \bar{Y}$$

$$\hat{p} = \frac{\overline{Y}}{S^2 + \overline{Y}} \qquad \text{Estimador de } p$$

Y sustituyendo la última expresión en $\hat{r} = \hat{p}\,\overline{Y}$

$$\hat{r} = \frac{\overline{Y}^2}{S^2 + \overline{Y}} \qquad \text{Estimador de } r$$

S))))))))))))))))))))))))))))))))))))))

MÉTODO DE MÁXIMA VEROSIMILITUD

Uno de los mejores métodos para realizar estimación puntual es el de *máxima verosimilitud*, el cual consiste básicamente en obtener una función de verosimilitud (probabilidad conjunta) y maximizarla.

Definición 3.7

Sea $f_X(x;\theta)$ la distribución de una población donde θ es el parámetro a estimar. La función de verosimilitud es una función de las vv.aa. de muestreo y del parámetro θ a estimar definida como sigue:

$$L(x_1, x_2, \ldots, x_n \; ; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i \; ; \theta)$$

Nótese que la función de verosimilitud L es la distribución conjunta de las vv.aa. de muestreo si éstas son independientes.

Definición 3.8

Un estimador de máxima verosimilitud es aquel que maximiza la función de verosimilitud.

En la práctica, para maximizar la función de verosimilitud se utiliza el cambio de variable de L por $\ln L$, como se observa en el siguiente ejemplo.

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.8

Construir un estimador de máxima verosimilitud para el parámetro p de una distribución Bernoulli, utilizando una muestra de tamaño n.

Resolución

La distribución de Bernoulli es

$$f_x(x) = \begin{cases} p^x(1-p)^{1-x} & si\ x = 0,1 \\ 0 & en\ otro\ caso \end{cases}$$

La función de verosimilitud es

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Puesto que si un valor $p = \hat{p}$ maximiza a $\ln L(p)$ también maximiza a $L(p)$ se puede realizar el cambio

$$\ln L(p) = \ln\left[p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right]$$

utilizando las propiedades de los logaritmos

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

Maximizando

$$\frac{d}{dp}\left[\ln L(p)\right] = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$$

$$\frac{1-p}{p} = \frac{n - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$\frac{1}{p} = \frac{n}{\sum_{i=1} x_i} \quad \Rightarrow \quad p = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\therefore$ El estimador de máxima verosimilitud de p es

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

S))))))))))))))))))))))))))))))))))))))

Ejemplo 3.9

Obtener un estimador de máxima verosimilitud para el parámetro p de una distribución geométrica, utilizando una muestra aleatoria de tamaño n.

Resolución

La distribución geométrica es

$$f_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, 3, \ldots \\ 0 & \textit{en otro caso} \end{cases}$$

por lo que la función de verosimilitud es

$$L(p) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum_{i=1}^{n}(x_i - 1)}$$

$$L(p) = \frac{p^n (1-p)^{\sum_{i=1}^{n} x_i}}{(1-p)^n}$$

Tomando logaritmos

$$\ln L(p) = \ln \left[\frac{p^n (1-p)^{\sum_{i=1}^{n} x_i}}{(1-p)^n} \right]$$

$$= n \ln p + \left(\sum_{i=1}^{n} x_i \right) \ln(1-p) - n \ln(1-p)$$

Derivando e igualando a cero

$$\frac{d}{dp}\left[\ln L(p) \right] = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i}{1-p} + \frac{n}{1-p} = 0$$

despejando a p ,

$$\frac{n}{p} + \frac{n}{1-p} = \frac{\sum_{i=1}^{n} x_i}{1-p}$$

$$\frac{n(1-p) + np}{p} = \sum_{i=1}^{n} x_i$$

$$\frac{n}{p} = \sum_{i=1}^{n} x_i \quad \Rightarrow \quad \hat{p} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

$\therefore$ El estimador de máxima verosimilitud de p es

$$\hat{p} = \frac{1}{\bar{X}}$$

S))

Si se desconocen dos o más parámetros de una distribución entonces se plantea la función de verosimilitud como una función de todos los parámetros desconocidos; y se optima aplicando logaritmos y resolviendo el sistema de ecuaciones que se obtiene de las derivadas parciales de la función logaritmo de la función de verosimilitud.

S))

Ejemplo 3.10

Considere que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una distribución normal con media μ y variancia σ^2 desconocidas. Construir los estimadores de máxima verosimilitud para dichos parámetros y decir si son insesgados o no.

Resolución

Para la distribución normal

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad , \quad -\infty < x < \infty$$

Por lo que la función de verosimilitud es

$$L(x_1, x_2, \ldots, x_n \;;\; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$L(x_1, x_2, \ldots, x_n \;;\; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Tomando logaritmos

$$\ln L = n \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \ln e$$

$$\ln L = n \ln \frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad \ldots (a)$$

Derivando parcialmente e igualando a cero

$$\frac{\partial}{\partial\mu}[\ln L] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \qquad \ldots (b)$$

$$\frac{\partial}{\partial\sigma}[\ln L] = n(\sqrt{2\pi}\,\sigma)\left(-\frac{1}{\sqrt{2\pi}\,\sigma^2}\right) + \frac{2}{2\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \qquad \ldots (c)$$

Y las ecuaciones (b) y (c) forman un sistema de dos ecuaciones con dos incógnitas.
De (b)

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \quad \mu = \frac{1}{n}\sum_{i=1} x_i$$

$$\mu = \bar{x} \qquad \ldots (d)$$

$$\therefore \text{ El estimador para } \mu \text{ es } \bar{X}, \qquad \hat{\mu} = \bar{X}$$

Sustituyendo (d) en (c)

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$-n + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\therefore$ El estimador para σ^2 es

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = S_n^2$$

Como se demostró anteriormente $\bar{X}$ es un estimador insesgado de μ y S_n^2 es un estimador sesgado de σ^2.

S))

ESTIMACIÓN DE PARÁMETROS POR INTERVALOS DE CONFIANZA

La estimación puntual es útil pues proporciona el valor más representativo para el parámetro que se desea estimar; sin embargo, la probabilidad de que el estimador tome el valor del parámetro es prácticamente cero, por lo que en algunos problemas es mucho más útil un intervalo para el cual la probabilidad de que el parámetro se encuentre en dicho intervalo sea alta.

La *estimación por intervalos* parte de construir intervalos aleatorios. Un *intervalo aleatorio* es un intervalo en donde al menos uno de sus límites es una v.a. En general, un intervalo aleatorio se construye a través de los estadísticos L_1 y L_2 tales que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$$

donde $1 - \alpha$ recibe el nombre de *nivel* o *coeficiente de confianza*, L_1 y L_2 se denominan *límites de confianza* inferior y superior, respectivamente y α se llama *nivel de significación* o *nivel de significancia* o simplemente *significancia*.

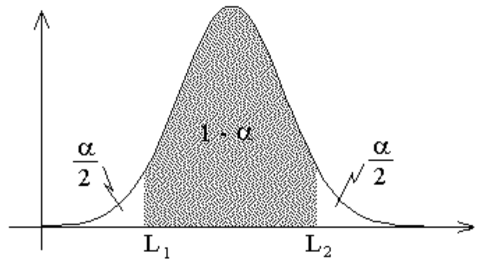


Fig. 3.3. Límites de confianza

Considérese que se tiene una variable aleatoria X contenida en un intervalo dado, por ejemplo:

$1 \le X \le 2$ (Intervalo determinístico)

el cual se podría escribir como:

$$1 \le X \le 2$$
$$\frac{1}{X} \le 1 \le \frac{2}{X}$$
$$X \ge 1 \ge \frac{X}{2}$$
$$2X \ge 2 \ge X$$
$$X \le 2 \le 2X \qquad \text{(Intervalo aleatorio)}$$

En donde a partir de una condición se ha generado un intervalo aleatorio, en el cual, la constante que ha quedado dentro del intervalo puede ser el valor de un parámetro objetivo θ .

Definición 3.7 Intervalo de confianza

Un intervalo de confianza para el parámetro poblacional θ al nivel de confianza $\mathbf{100}(1-\alpha)\%$, siendo α un valor en el intervalo $[0,1]$, se define como un intervalo de la forma $L_1 \le \theta \le L_2$ cuyos extremos son estadísticos y tiene la propiedad de que

$$P(L_1 \le \theta \le L_2) = 1 - \alpha$$

Para obtener el intervalo de confianza se utiliza el *método del pivote*. El método del pivote se basa en la determinación de una expresión pivote que es función de las mediciones de la muestra y del parámetro desconocido θ , en donde θ es la única cantidad desconocida y el pivote tiene una distribución de probabilidad que no depende del parámetro θ .

Por ejemplo, si el estadístico $\hat{\Theta}$ tiene una distribución normal, tal que $\hat{\Theta} \sim N(\mu, \sigma^2)$ entonces se puede calcular $P(z_1 \le \hat{\Theta} \le z_2) = 1 - \alpha$, y despejando θ se obtiene:

$$P(L_1 \le \theta \le L_2) = 1 - \alpha$$

y L_1 y L_2 son los estadísticos del intervalo de confianza.

INTERVALO DE CONFIANZA PARA LA MEDIA

En el capítulo anterior se estudió el estadístico

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

como estimador de la media poblacional μ, y si se considera una muestra grande $n \geq 30$, extraída de una población con σ^2 conocida, entonces del teorema central del límite $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n} \right)$ y en consecuencia $Z \sim N(0, 1)$, donde

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

por lo que

$$P\left(L_1 \leq \mu \leq L_2\right) = 1 - \alpha$$

se obtiene de

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} \right) = 1 - \alpha$$

De donde el intervalo de confianza de dos lados para la media con un nivel de confianza de $(1 - \alpha)\, 100\%$, cuando la muestra es grande es:

$$\boxed{\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}} \quad \ldots (3.1)$$

y los límites son:

$$L_1 = \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

$$L_2 = \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

el valor $z_{\frac{\alpha}{2}}$ se obtiene de tablas de distribución normal estándar de forma que $P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

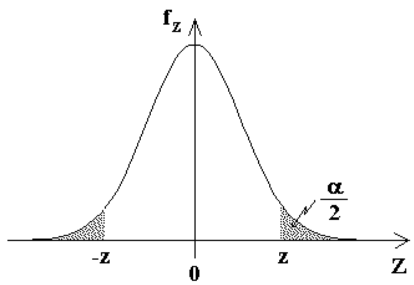


Fig. 3.4. Nivel de significancia

El denotar a z como $z_{\frac{\alpha}{2}}$, es una notación común en estadística, pero no está completamente generalizada.

Cuando la muestra es pequeña $(n < 30)$ y la población tiene una distribución normal con variancia conocida, entonces puede emplearse la expresión (3.1), como se ilustra en el siguiente ejemplo.

))))))))))))))))))))))))))))))))))))

Ejemplo 3.11

Construir un intervalo del 95% de confianza para la media de una población normal con variancia 4, y $n = 10$. Además, utilizando los datos: $4, 7, 5, 10, 23, 17, 9, 15, 7,$ y 10 obtener una estimación.

Resolución

Puesto que se desea un intervalo para la media, la variancia es conocida y la distribución de la población es normal, de la expresión (3.1) se tiene

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

donde $\bar{x} = \frac{1}{n} \sum_{1=1}^{n} x_i = \frac{1}{10}(4 + 7 + 5 + 10 + 23 + 17 + 9 + 15 + 7 + 10)$

$$\bar{x} = 10.7$$

$$\sigma_X = \sqrt{\sigma_X^2} = 2 \quad , \quad 1 - \alpha = 0.95$$

$n = 10$

y de tablas $P(Z \geq z_{\frac{\alpha}{2}}) = \dfrac{\alpha}{2} = 0.025 \qquad \Rightarrow \qquad z_{\frac{\alpha}{2}} = 1.96$

El intervalo aleatorio es

$$\bar{X} - 1.96\left(\frac{2}{\sqrt{10}}\right) \leq \mu_X \leq \bar{X} + 1.96\left(\frac{2}{\sqrt{10}}\right)$$

y sustituyendo $\bar{X}$ por $\bar{x}$ se obtiene la estimación

$$10.7 - 1.96\left(\frac{2}{\sqrt{10}}\right) \leq \mu_X \leq 10.7 + 1.96\left(\frac{2}{\sqrt{10}}\right)$$

$$9.46 \leq \mu_X \leq 11.9 \qquad \text{Estimación}$$

))))))))))))))))))))))))))))))))))))

El último intervalo no es aleatorio, es decir, la estimación no es un intervalo de probabilidad 0.95, puesto que la probabilidad de que la media poblacional esté contenida en el intervalo es 1 ó 0, lo que significa que la media está o no contenida en el intervalo.

La interpretación del intervalo aleatorio es, que en repetidas utilizaciones de la fórmula, $(1 - \alpha)$ 100% de las veces el intervalo generado (estimación) contendrá a la media.

En el ejemplo anterior, se construyó un intervalo del 95% de confianza, por lo que se interpreta considerando que en repetidas utilizaciones de la fórmula, el 95% de las veces el intervalo generado contendrá a la media de la población.

A pesar de que la estimación contiene o no a la media de la población, es común decir que se tiene una confianza del 95% de que sí lo contenga, pero debe recordarse que esto ya no es una probabilidad.

En muchos casos no se conoce la variancia de la población, la cual se puede aproximar puntualmente mediante el estimador insesgado S^2_{n-1} y sustituirlo en lugar de σ^2_X en la expresión (3.1).

))))))))))))))))))))))))))))))))))))

Ejemplo 3.12

En una escuela de ingeniería, se seleccionaron 50 alumnos y se determinó el promedio de horas que estudian a la semana, el cual fue de $\bar{x} = 3.5$ [h] con una desviación estándar de los datos igual a $s_n = \sqrt{3.92}$ [h] . Con estos datos, estimar las horas que estudiaron en promedio los alumnos con un coeficiente de confianza igual a 95%.

Resolución

Se desea un intervalo para la media de las horas de estudio μ_X.

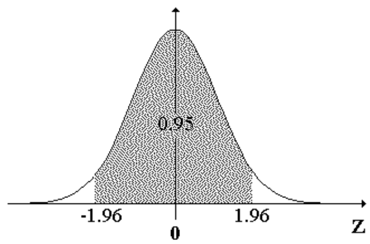
Puesto que la muestra es grande, y del TCL se tiene:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

como el valor de σ_X^2 se desconoce se puede aproximar mediante el estimador insesgado S_{n-1}^2 de donde la expresión (3.1) se reescribe como:

$$\boxed{\bar{X} - z_{\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}} \quad . . . (3.2)$$

al considerar un intervalo simétrico y centrado en μ_X y puesto que $1 - \alpha = 0.95$ entonces, de tablas $z_{\frac{\alpha}{2}} = 1.96$



y sustituyendo por los valores de la muestra

$$\bar{x} = 3.5$$

$$s_{n-1} = \sqrt{\frac{n\, s_n^2}{n - 1}} = \sqrt{\frac{(50)(3.92)}{49}} = 2$$

Y sustituyendo en el intervalo se tiene:

$$3.5 - \frac{2}{\sqrt{50}} 1.96 \leq \mu \leq 3.5 + \frac{2}{\sqrt{50}} 1.96$$

operando

$$2.9456 \leq \mu \leq 4.05$$

))))))))))))))))))))))))))))))))))))

Cuando el tamaño de la muestra es pequeño, $n < 30$, la muestra se extrae de una población con distribución normal y la variancia se desconoce, entonces debe utilizarse el estadístico

$$T = \frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

el cual tiene una distribución t con $n - 1$ grados de libertad, y el intervalo de confianza del $(1 - \alpha) 100\%$ es:

$$\boxed{\bar{X} - t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}}} \quad \ldots \quad (3.3)$$

donde $t_{\frac{\alpha}{2},(n-1)}$ se obtiene de tablas con distribución t de Student de forma que $P(T \ge t_{\frac{\alpha}{2},(n-1)}) = \frac{\alpha}{2}$

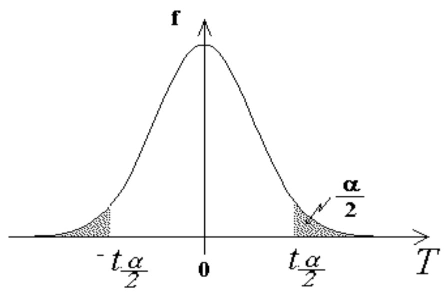


Fig. 3.5 Distribución t

))))))))))))))))))))))))))))))))))))

Ejemplo 3.13

Diez ejes de precisión fabricados en cierto proceso tienen un diámetro promedio de 0.908 cm, con una desviación estándar de 0.004 cm. Considerando que los datos provienen de una muestra aleatoria con distribución normal, construir un intervalo de confianza del 95% para el diámetro promedio real de los ejes fabricados.

Resolución

Puesto que la muestra es pequeña, los datos provienen de una población con distribución normal y la variancia es desconocida, entonces el intervalo de confianza para la media está dado por (3.3), de donde

$$\bar{X} - t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}} \le \mu_X \le \bar{X} + t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}}$$

y sustituyendo los valores obtenidos en la muestra

$$\bar{x} = 0.908$$

$$s_{n-1} = 0.004$$

$$n = 10$$

y de tablas $\quad t_{\frac{\alpha}{2},(n-1)} = t_{0.025,(9)} = 2.262$

∴

$$0.908 - 2.262 \frac{(0.004)}{\sqrt{10}} \le \mu \le 0.908 + 2.262 \frac{(0.004)}{\sqrt{10}}$$

operando

$$0.905 \le \mu \le 0.911$$

))))))))))))))))))))))))))))))))))))

Cuando se desea construir un intervalo de confianza para la media, existen entonces dos estadísticos que se pueden utilizar: Z y t, dependiendo de la información que se tiene. Por otro lado, también pueden construirse intervalos de un solo lado, como se observa en el siguiente ejemplo.

))))))))))))))))))))))))))))))))))))

Ejemplo 3.14

Un fabricante produce anillos de pistón para un motor de automóvil, se sabe que el diámetro de los anillos se distribuye aproximadamente en forma normal y con una desviación estándar $\sigma_X = 0.001 \ [\text{mm}]$. Una muestra aleatoria de 15 anillos tiene media de $71.036 \ [\text{mm}]$.

a) Construir un intervalo de confianza de dos lados del 95% con respecto al diámetro medio de los anillos de pistón.

b) Construir un límite de confianza inferior del 95% respecto al diámetro medio de los anillos de pistón.

Resolución

a) Puesto que la población tiene distribución normal, el intervalo de dos lados está dado por:

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

De tablas se obtiene que

$$P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} = 0.025$$

se satisface con $z_{\frac{\alpha}{2}} = 1.96$

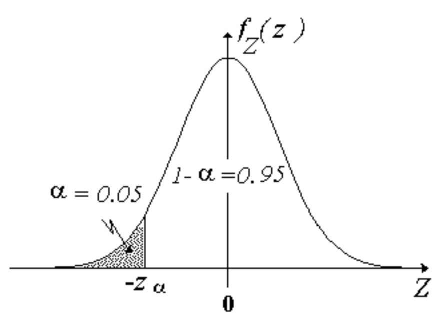
Sustituyendo $\bar{x} = 74.036$, $\sigma_X = 0.001$, $n = 15$

$$74.036 - 1.96\left(\frac{0.001}{\sqrt{15}}\right) \leq \mu_X \leq 74.036 + 1.96\left(\frac{0.001}{\sqrt{15}}\right)$$

$$74.03549 \leq \mu_X \leq 74.03650$$

b) Un límite de confianza inferior es de la forma $L_1 \leq \mu_X$ por lo que

$$\boxed{\bar{X} - z_{\alpha} \frac{\sigma_X}{\sqrt{n}} \leq \mu_X} \quad \ldots \quad (3.2a)$$



donde z_α se obtiene de $P(Z \leq -z_\alpha) = 0.05$

De tablas

$$-z_{\alpha} = -1.645$$

sustituyendo

$$74.036 - 1.645\left(\frac{0.01}{\sqrt{15}}\right) \leq \mu_X$$

$$74.03175 \leq \mu_X$$

))))))))))))))))))))))))))))))))))))

Como se observó en el ejemplo, es posible obtener intervalos unilaterales, tanto inferiores como superiores, procediendo de una manera muy similar que para los intervalos de dos lados.

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS

Al igual que el intervalo de confianza para la media, para el intervalo de confianza para la diferencia de medias pueden distinguirse dos casos: muestras grandes (variancia conocida o que se puede aproximar) y muestras pequeñas (distribución normal y variancia desconocida).

Definición 3.8

Cuando la muestra es grande el estadístico para la diferencia de medias está dado por

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Por lo que el intervalo de confianza de dos lados para la diferencia de medias con un nivel de confianza de $(1 - \alpha)\,100\%$ centrado en $\mu_1 - \mu_2$ es

$$\bar{X}_1 - \bar{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

donde $z_{\frac{\alpha}{2}}$ se obtiene de tablas con distribución normal estándar de forma que

$$P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

Si se desconocen las variancias de las poblaciones σ_1^2 y σ_2^2 y la muestra es grande, entonces se pueden sustituir por sus estimadores puntuales $S^2_{n_1-1}$ y $S^2_{n_2-1}$.

))))))))))))))))))))))))))))))))))))

Ejemplo 3.15

La resistencia del caucho a la abrasión aumenta si se agrega una carga de sílice y un agente de acoplamiento para enlazar químicamente a la carga con las cadenas de polímero del caucho. Cincuenta muestras de caucho con el agente de acoplamiento tipo I dieron una resistencia promedio de 92 y la variancia de las mediciones fue de 20. Cuarenta muestras de caucho con el agente de acoplamiento tipo II dieron un promedio de 98 y una variancia de 30 en sus mediciones. Estimar la diferencia verdadera entre las resistencias promedio a la abrasión en un intervalo de confianza del 95% .

Resolución

Puesto que las muestras son grandes y aproximando puntualmente las variancias, se tiene:

$$\bar{X}_1 - \bar{X}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} = 0.025$$

Sustituyendo:

$$92 - 98 - 1.96 \sqrt{\frac{20}{50} + \frac{30}{40}} \leq \mu_1 - \mu_2 \leq 92 - 98 + 1.96 \sqrt{\frac{20}{50} + \frac{30}{40}}$$

operando

$$-8.1019 \leq \mu_1 - \mu_2 \leq -3.8981$$

))))))))))))))))))))))))))))))))))))

El intervalo obtenido en el ejemplo anterior, $-8.1019 \leq \mu_1 - \mu_2 \leq -3.8981$, tiene la interpretación adicional de que, al no contener al cero, una de las medias es mayor que la otras. En este caso el intervalo indica que la media μ_2 es mayor que la media μ_1.

Cuando la variancia de la población es desconocida y la muestra es pequeña, se puede utilizar la distribución t para obtener el intervalo de confianza.

Definición 3.9

Cuando las muestras son pequeñas y provienen de poblaciones normales con variancias desconocidas pero iguales, entonces el estadístico para la diferencia entre dos medias está dado por

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde

$$S_p^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}$$

y

$$T \sim t_{(n_1 + n_2 - 2)}$$

Por lo que el intervalo de confianza de dos lados para la diferencia de medias con un nivel de confianza de $(1 - \alpha)100\%$ centrado en $\mu_1 - \mu_2$ es

$$\boxed{\bar{X}_1 - \bar{X}_2 - t_{\frac{\alpha}{2},(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\frac{\alpha}{2},(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad ...(3.4)$$

))))))))))))))))))))))))))))))))))))

Ejemplo 3.16

Para dos muestras extraídas de distribuciones normales se obtuvieron los siguientes resultados

$n_1 = 9$	$n_2 = 7$
$\bar{y}_1 = 43.71$	$\bar{y}_2 = 39.63$
$s_1 = 5.88$	$s_2 = 7.68$

Obtener un intervalo de confianza para la diferencia de medias con un coeficiente de confianza igual a 0.95.

Resolución

Puesto que la muestra proviene de una distribución normal y las muestras son pequeñas con variancias desconocidas

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{(8)(5.88)^2 + (6)(7.68)^2}{9 + 7 - 2} = 45.0305$$

$$s_p = 6.7108$$

valor de t con 14 grados de libertad para $1 - \alpha = 0.95$ es: $t_{0.025,(14)} = 2.145$

Puesto que el intervalo se obtiene mediante

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\frac{\alpha}{2},(n_1+n_2-2)}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

valuando

$$43.71 - 39.63 \pm 2.145\,(6.7108)\sqrt{\frac{1}{9} + \frac{1}{7}}$$

Finalmente

$$-3.17 \leq \mu_1 - \mu_2 \leq 11.33$$

))))))))))))))))))))))))))))))))))))

INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN

Definición 3.10

Si se toma una muestra de tamaño n de una población muy grande (o infinita), y X observaciones pertenecen a la clase de interés, entonces $\hat{P} = \frac{X}{n}$ es un estimador puntual de la proporción de la población que pertenece a la clase en cuestión, y la distribución de muestreo es

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

donde $Z \sim N(0,1)$

y n , p son los parámetros de la distribución Binomial.

Utilizando el estadístico $\dfrac{\hat{P} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ y aproximando la cantidad $p(1-p)$ mediante su estimador puntual $\hat{P}(1-\hat{P})$ se obtiene el intervalo de confianza de dos lados con un coeficiente $(1-\alpha)100\%$ para la proporción p es

$$\boxed{\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \quad .....(3.5)$$

El aproximar la suma de variables aleatorias de Bernoulli (distribución binomial), mediante la distribución normal, se realiza por el teorema central del límite, y debe recordarse del curso de Probabilidad que la aproximación es bastante buena si $np > 5$ cuando $p \le 0.5$, o bien $nq > 5$ cuando $p > 0.5$.

))))))))))))))))))))))))))))))))))))

Ejemplo 3.17

En una muestra al azar de 60 secciones de tubo en una planta química, 8 de ellos mostraron señales de corrosión seria. Construir un intervalo de confianza del 95 % para la proporción de los tramos de tubo con corrosión seria.

Resolución

Utilizando la fórmula (3.5), con $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ de tablas, y recordando que $\hat{P} = \dfrac{X}{n}$, se tiene:

$$\frac{8}{60} - 1.96\sqrt{\frac{\frac{8}{60}\left(1 - \frac{8}{60}\right)}{60}} \le p \le \frac{8}{60} + 1.96\sqrt{\frac{\frac{8}{60}\left(1 - \frac{8}{60}\right)}{60}}$$

Finalmente:

$$0.04731 \le p \le 0.21934$$

))))))))))))))))))))))))))))))))))))

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE PROPORCIONES

Definición 3.11

Si dos muestras independientes de tamaño n_X y n_Y se extraen de poblaciones infinitas con distribuciones de Bernoulli, X representa el número de observaciones de la primera muestra que corresponden a la clase de interés, y Y representa el número de observaciones de la segunda muestra que corresponden a la clase en cuestión, entonces la distribución de muestreo para la diferencia de proporciones está dada por

$$Z = \frac{(\hat{P}_X - \hat{P}_Y) - (p_X - p_Y)}{\sqrt{\dfrac{p_X(1-p_X)}{n_X} + \dfrac{p_Y(1-p_Y)}{n_Y}}}$$

donde $\quad Z \sim N(0,1)$

De la definición (3.11) se obtiene el intervalo de confianza de dos lados para la diferencia de proporciones, con un nivel de confianza de $(1-\alpha)100\%$, el cual es

$$(\hat{P}_1 - \hat{P}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \le p_1 - p_2 \le (\hat{P}_1 - \hat{P}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \quad \text{...(3.6)}$$

))))))))))))))))))))))))))))))))))))

Ejemplo 3.18

Dos grupos de 80 pacientes tomaron parte en un experimento en el cual un grupo recibió píldoras que contenían un antialérgico, mientras que al otro grupo se le administró un placebo, es decir, una píldora sin droga alguna. En el grupo que recibió el medicamento 23 exhibieron síntomas alérgicos, mientras que en el otro grupo 41 los exhibieron. Obtener un intervalo de confianza del 99 % para la diferencia entre las proporciones.

Resolución

Sustituyendo en la fórmula (3.6) con

$$\hat{p}_1 = \frac{x}{n_1} = \frac{23}{80} \qquad\qquad \hat{p}_2 = \frac{y}{n_2} = \frac{41}{80}$$

Y $z_{\frac{\alpha}{2}} = z_{0.005} = 2.575$ de tablas, se tiene

$$\left(\frac{23}{80}-\frac{41}{80}\right)-2.575\sqrt{\frac{\left(\frac{23}{80}\right)\left(1-\frac{23}{80}\right)}{80}+\frac{\left(\frac{41}{80}\right)\left(1-\frac{41}{80}\right)}{80}}$$

$$\leq p_1 - p_2 \leq$$

$$\left(\frac{23}{80}-\frac{41}{80}\right)+2.575\sqrt{\frac{\left(\frac{23}{80}\right)\left(1-\frac{23}{80}\right)}{80}+\frac{\left(\frac{41}{80}\right)\left(\frac{41}{80}\right)\left(1-\frac{41}{80}\right)}{80}}$$

operando

$$-0.419 \leq p_1 - p_2 \leq -0.031$$

))))))))))))))))))))))))))))))))))))

INTERVALO DE CONFIANZA PARA LA VARIANCIA

Definición 3.12

Si X es una v.a. con distribución normal con media μ_X y variancia σ_X^2 desconocidas, entonces el estadístico empleado es

$$X^2 = \frac{n-1}{\sigma_X^2} S_{n-1}^2$$

donde $X^2 \sim \chi^2_{(n-1)}$

Utilizando el estadístico $\frac{n-1}{\sigma_X^2} S_{n-1}^2$ se obtiene el intervalo de confianza de dos lados con un coeficiente de confianza de $(1-\alpha)100\%$ para σ_X^2, el cual es

$$\boxed{\frac{(n-1)S_{n-1}^2}{X^2_{\frac{\alpha}{2},(n-1)}} \leq \sigma_X^2 \leq \frac{(n-1)S_{n-1}^2}{X^2_{1-\frac{\alpha}{2},(n-1)}}} \qquad ...(3.7)$$

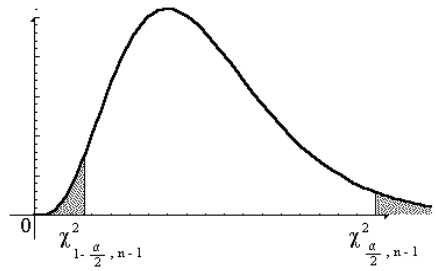


Fig. 3.6. Distribución Ji cuadrada

)))))))))))))))))))))))))))))))))))))

Ejemplo 3.19

Considérense los siguientes datos:

8.2	8.28	8.24
8.23	8.21	8.25
8.24	8.23	8.24
8.25	8.2	8.26
8.19	8.23	8.26

Obtener:

a) Un intervalo de confianza de dos lados del 95% para σ_X^2.

b) Un intervalo de confianza inferior del 95% para σ_X^2.

c) Un intervalo de confianza superior de 95% para σ_X^2.

Resolución

De los datos de la tabla se obtiene

$$\bar{x} = 8.23$$

$$s_{n-1} = 0.0253$$

a) Sustituyendo en

$$\frac{(n-1)S_{n-1}^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} \le \sigma_X^2 \le \frac{(n-1)S_{n-1}^2}{\chi_{1-\frac{\alpha}{2},(n-1)}}$$

De tablas

$$\chi^2_{\frac{\alpha}{2},(n-1)} = \chi^2_{0.025,(14)} = 26.12$$

$$\chi^2_{1-\frac{\alpha}{2},(n-1)} = \chi^2_{0.975,(14)} = 5.63$$

Por lo que

$$\frac{14(0.0253)^2}{26.12} \le \sigma_X^2 \le \frac{14(0.0253)^2}{5.63}$$

$$0.0003438 \le \sigma_X^2 \le 0.0015916$$

b) Para un intervalo inferior

$$\frac{(n-1)S_{n-1}^2}{X_{\alpha,(n-1)}^2} \le \sigma_X^2$$

De tablas $\chi^2_{0.05,(14)} = 23.68$

entonces

$$\frac{14(0.0253)^2}{23.68} \le \sigma^2$$

$$0.0003784 \le \sigma_X^2$$

c) Para un intervalo superior

$$\sigma_X^2 \le \frac{(n-1)S_{n-1}^2}{X_{1-\alpha,(n-1)}^2}$$

De tablas

$$\chi^2_{0.95,(14)} = 6.57$$

entonces

$$\sigma_X^2 \le \frac{14(0.0253)^2}{6.57}$$

$$\sigma_X^2 \le 0.0013639$$

))))))))))))))))))))))))))))))))))))

INTERVALO DE CONFIANZA PARA LA RAZÓN DE VARIANZAS

Definición 3.13

Si X y Y son vv.aa. independientes con distribuciones normales con medias μ_X y μ_Y desconocidas y varianzas σ_X^2 y σ_Y^2 desconocidas, respectivamente, entonces el estadístico empleado es

$$F = \frac{\dfrac{S_Y^2}{\sigma_Y^2}}{\dfrac{S_X^2}{\sigma_X^2}}$$

donde

$$F \sim F_{(n_Y - 1, n_X - 1)}$$

Utilizando el estadístico $\dfrac{S_Y^2 \big/ \sigma_Y^2}{S_X^2 \big/ \sigma_X^2}$ se obtiene el intervalo de confianza de dos lados con un coeficiente de confianza de $(1 - \alpha)100\%$ para la relación de las variancias $\dfrac{\sigma_X^2}{\sigma_Y^2}$, el cual es

$$\boxed{\frac{S_X^2}{S_Y^2} F_{1-\frac{\alpha}{2},(n_Y-1,n_X-1)} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{S_X^2}{S_Y^2} F_{\frac{\alpha}{2},(n_Y-1,n_X-1)}} \qquad \text{...(3.8)}$$

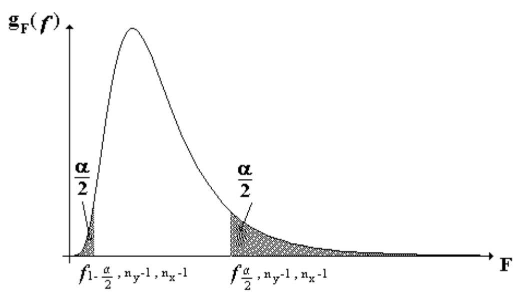


Fig. 3.7 Distribución F

O bien ,utilizando el recíproco de F, se puede construir el intervalo de confianza de la forma:

$$\boxed{\frac{S_X^2}{S_Y^2}\ \frac{1}{F_{\frac{\alpha}{2},\,(n_X-1,\,n_Y-1)}} \le \frac{\sigma_X^2}{\sigma_Y^2} \le \frac{S_X^2}{S_Y^2}\ \frac{1}{F_{1-\frac{\alpha}{2},\,(n_X-1,\,n_Y-1)}}}$$

))))))))))))))))))))))))))))))))))))

Ejemplo 3.20

Se extraen dos muestras independientes de poblaciones normales obteniéndose los siguientes datos

$$\begin{array}{ll} n_1 = 15 & n_2 = 10 \\ \bar{x}_1 = 300 & \bar{x}_2 = 325 \\ s_1^2 = 16 & s_2^2 = 49 \end{array}$$

Construir un intervalo de confianza de dos lados del 95 % con respecto a la relación de las variancias $\frac{\sigma_1^2}{\sigma_2^2}$.

Resolución

El intervalo está dado por

$$\frac{S_1^2}{S_2^2}\ f_{1-\frac{\alpha}{2},(n_2-1,n_1-1)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\ f_{\frac{\alpha}{2},(n_2-1,n_1-1)}$$

De tablas

$$f_{1-\frac{\alpha}{2},(n_2-1,n_1-1)} = f_{0.975,(9,14)} = \frac{1}{f_{0.025,(14,9)}} \approx \frac{1}{3.8} = 0.263$$

$$f_{\frac{\alpha}{2},(n_2-1,n_1-1)} = f_{0.025,(9,14)} = 3.21$$

sustituyendo

$$\frac{16}{49}(0.263) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{16}{49}(3.21)$$

$$0.085 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.048$$

))))))))))))))))))))))))))))))))))))

DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA

Una de las interrogantes más comunes es la determinación del tamaño de la muestra cuando se va a hacer un muestreo, para la estimación de un parámetro. Para determinar el tamaño de la muestra se utilizan los conceptos de la estimación por intervalos, en donde se acepta un "error" sobre la estimación puntual, por ejemplo para la media se tiene el intervalo

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

en donde el error es $E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, que se suma y se resta al estimador puntual para generar el intervalo confianza. Al fijar un valor para el error, y conociendo o aproximando la desviación estándar, se puede despejar el tamaño de la muestra.

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2$$

El redondeo se hace siempre hacia arriba, para asegurar el nivel de significancia.

Para una proporción el procedimiento es semejante, y se obtiene el despeje

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 p\,(1-p)$$

que requiere a p como dato. Puesto que p es desconocido puede aproximarse con una estimación $\hat{p}$, o en el caso más extremo, utilizar $p = 0.5$ que representa el máximo desconocimiento sobre p y por lo tanto genera el mayor tamaño de muestra. Un error E típico en una proporción es $E = 0.05$.

Cuando se tienen dos poblaciones y de cada población se desea extraer una muestra, se igualan los tamaños de muestra para realizar el despeje, obteniendose

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \left(\sigma_1^2 + \sigma_2^2 \right)$$

Para el caso de proporciones, el tamaño de las muestras se obtiene de:

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \left(p_1 q_1 + p_2 q_2 \right)$$

TÓPICOS ESPECIALES:

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS, CASOS ESPECIALES

Existen algunos casos especiales para los intervalos de confianza de diferencia de medias. El primero de ellos es cuando se tienen datos apareados, o en pares, es decir, las muestras aleatorias no son independientes y tienen el mismo tamaño. El segundo de ellos, que queda un poco más allá del objetivo del presente curso, se tiene cuando las muestras son pequeñas, independientes, con distribuciones aproximadamente normales con variancias desconocidas y diferentes.

DATOS EN PARES

Cuando se observan datos en pares y se espera que exista una fuerte correlación entre cada pareja de datos, se debe generar una nueva variable aleatoria para construir el intervalo de confianza.

Sea la variable aleatoria $D_i = X_{1i} - X_{2i}$, donde $i = 1, 2, \ldots, n$, entonces:

$$\mu_D = E(D) = \mu_1 - \mu_2,$$

y el intervalo se puede generar mediante:

$$\boxed{\bar{D} - t_{\frac{\alpha}{2},(n-1)} \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + t_{\frac{\alpha}{2},(n-1)} \frac{S_D}{\sqrt{n}}}$$

donde $\bar{D}$ y S_D son la media y la desviación estándar muestrales, que se calculan mediante:

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} (X_{1i} - X_{2i})$$

$$S_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left[(X_{1i} - X_{2i}) - \bar{D} \right]^2}$$

S))))))))))))))))))))))))))))))))))))Q

Ejemplo 3.21

Cierto gimnasio afirma que una nueva rutina de ejercicios reduce la talla de la cintura de una persona dos centímetros en promedio en un periodo de cinco días. Se indicó la talla de la cintura de seis hombres que participaron en este programa de ejercicios antes y después del periodo de cinco días y las cifras resultantes se registraron en la tabla:

	HOMBRES					
	1	2	3	4	5	6
Talla anterior de la cintura	90.4	95.5	98.7	115.9	104.0	85.6
Talla posterior de la cintura	91.7	93.9	97.4	112.8	101.3	84.0

Determinar si la afirmación de este gimnasio es cierta, calculando un intervalo de confianza del 95 % para la reducción promedio de la talla de la cintura. Asumir que la distribución de las diferencias de las tallas antes y después de la rutina es aproximadamente normal.

Resolución

Si X_1 es la talla anterior y X_2 la talla posterior, $D = X_1 - X_2$.

$n = 6$, $\bar{d} = 1.5$, $s_d = 1.543$, $t_{0.025,(5)} = 2.571$

utilizando un intervalo de confianza para $\mu_1 - \mu_2$ con observaciones apareadas, se tiene :

$$\bar{d} - t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \bar{d} + t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

$$1.5 - 2.571 \frac{(1.543)}{\sqrt{6}} \leq \mu_1 - \mu_2 \leq 1.5 + 2.571 \frac{(1.543)}{\sqrt{6}}$$

$$-0.12 \leq \mu_1 - \mu_2 \leq 3.12$$

La afirmación es válida.

S))))))))))))))))))))))))))))))))))))Q

VARIANCIAS DIFERENTES MUESTRAS PEQUEÑAS

Cuando el problema consiste en encontrar una estimación por intervalos para la diferencia de medias $\mu_1 - \mu_2$, las muestras son pequeñas, las poblaciones son aproximadamente normales y las variancias desconocidas no pueden considerarse iguales, entonces no existe un estadístico exacto para el problema; sin embargo, algunos autores han encontrado muy buenas aproximaciones utilizando el estadístico:

$$T^* = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

el cual tiene una distribución aproximadamente t, con ν grados de libertad, los cuales se aproximan mediante:

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1}\right] + \left[\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

o bien mediante

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1}\right] + \left[\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}\right]} - 2$$

puesto que ν difícilmente es entero se aproxima al entero más cercano.

El intervalo de confianza de dos lados queda entonces:

$$\boxed{\left(\overline{X}_1 - \overline{X}_2\right) - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\overline{X}_1 - \overline{X}_2\right) + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

USO DE LA FUNCIÓN DE VEROSIMILITUD PARA DETERMINAR LA SUFICIENCIA DE UN ESTIMADOR

La suficiencia de un estimador está relacionada con la función de verosimilitud a través del siguiente teorema

Teorema 3.1

Sea $\hat{\Theta}$ un estimador del parámetro θ, basado en la muestra aleatoria $X_1, X_2, \ldots, X_n$. Entonces $\hat{\Theta}$ es un estimador suficiente para θ si y sólo si la verosimilitud L se puede factorizar en dos funciones no negativas $g(\hat{\theta}, \theta)$ y $h(x_1, x_2, \ldots, x_n)$, i.e.

$$L(x_1, x_2, \ldots, x_n; \theta) = g(\hat{\theta}, \theta)\; h(x_1, x_2, \ldots, x_n)$$

S))))))))))))))))))))))))))))))))))))Q

Ejemplo 3.22

Sea $Y_1 \ , \ Y_2 \ , \ . \ . \ . \ . \ , \ Y_n$ una muestra aleatoria de una distribución Rayleigh con parámetro θ.

$$f_Y(y) \ = \ \begin{cases} \left(\dfrac{2y}{\theta} \right) e^{-\frac{y^2}{\theta}} & y > 0 \\ 0 & \textit{en otro caso} \end{cases}$$

Demostrar que $\sum_{i=1}^{n} Y_i^2$ es suficiente para θ.

Resolución

La función de verosimilitud es

$$L(\ y_1 \ , y_2 \ , \ldots , y_n \ ) \ = \ \frac{2^n y_1 y_2 \ldots y_n}{\theta^n} \ e^{-\sum_{i=1}^{n} \frac{y_i^2}{\theta}}$$

Si $g\left(\sum_{i=1}^{n} y_i^2 \ , \ \theta \right) \ = \ \dfrac{1}{\theta^n} \ e^{-\sum_{i=1}^{n} \frac{y_i^2}{\theta}}$

y $h(\ y_1 \ , y_2 \ , \ldots , y_n \ ) \ = \ 2^n \prod_{i=1}^{n} y_i$

entonces $\qquad L \ = \ g\left(\sum_{i=1}^{n} y_i^2 \ , \ \theta \right) h(\ y_1 \ , y_2 \ , \ldots , y_n \ )$

y puesto que la función de verosimilitud puede factorizarse en dos funciones no negativas $g(\hat{\theta} \ , \ \theta)$ y $h(x_1 \ , \ x_2 \ , \ . \ . \ . \ . \ , \ x_n \ )$

Y $\sum_{i=1}^{n} Y_i^2$ es suficiente para θ.

S))))))))))))))))))))))))))))))))))))Q

Evidentemente, existen varias factorizaciones de la función de verosimilitud L, con las cuales se puede probar la suficiencia. Cualquiera de ellas es igualmente válida.

BIBLIOGRAFÍA

Hines, William W. y Montgomery, Douglas C. - Probabilidad y Estadística para ingeniería, cuarta edición.- CECSA.- México, 2005.

Milton, Susan J. Y Arnold, Jesse C.- Probabilidad y Estadística para con aplicaciones para ingeniería y ciencias computacionales, cuarta edición.- McGraw-Hill.- México, 2004.

Devore, Jay L.- Probabilidad y Estadística para ingeniería y ciencias, séptima edición.- Cengage Learning.- México, 2008.

Mendenhall, William III. et al.- Introducción a la Probabilidad y Estadística.- Décimo cuarta edición.- Cengage Learning.- México 2015.

Wackerly Dennis D.- Mendenhall, William, *et al.*- Estadística Matemática con Aplicaciones, sexta edición.- Editorial Thomson.- México, 2002.

Walpole, Ronald E., *et al.*- Probability and Statistics for Engineers and Scientists.- Pearson.- USA, 2007.

Montgomery, Douglas C. y Runger, George C.-Probabilidad y Estadística aplicadas a la Ingeniería, segunda edición.- Limusa-Wiley.- México, 2002.

Scheaffer, Richard L. y McClave, James T.- Probabilidad y Estadística para Ingeniería.- Grupo Editorial Iberoamérica.- México, 1993.

Canavos, George C.- Probabilidad y Estadística Aplicaciones y Métodos.- McGraw-Hill.- México, 1988.

Meyer, Paul L.- Probabilidad y Aplicaciones Estadísticas.- Addison Wesley Iberoamericana.- México, 1992.

Spiegel, Murray R. et al.- Probabilidad y Estadística, cuarta edición.- Mc Graw-Hill.-México 2013.

Borras García, Hugo E., *et al.*- Apuntes de Probabilidad y Estadística.-Facultad de Ingeniería.- México, 1985.

Rosenkrantz, Walter A.- Introduction to Probability and Statistics for Scientists and Engineers.- McGraw-Hill.- EE.UU., 1997.

Ziemer, Rodger E.- Elements of Engineering Probability & Statistics.- Prentice Hall.- USA 1997.

ESTADÍSTICA

Profesores:
A. Leonardo Bañuelos Saucedo
Nayelli Manzanarez Gómez

NOTAS

TEMA 4

PRUEBAS DE HIPÓTESIS ESTADÍSTICAS

Tema IV
PRUEBAS DE HIPÓTESIS ESTADÍSTICAS

Definición 4.1
Una hipótesis es una afirmación acerca de la distribución de probabilidad de una variable aleatoria.

Las pruebas de hipótesis son parte de la inferencia estadística, y a menudo involucran a más de un parámetro de la distribución. Supóngase, por ejemplo, que se desea estimar el promedio de la estatura de los alumnos de la Facultad de Ingeniería, y se pretende saber si el promedio es 1.67 o no lo es. Lo anterior se expresaría:

$$\begin{aligned} H_0 &: \mu = 1.67\ [m] \\ H_1 &: \mu \neq 1.67\ [m] \end{aligned} \qquad \ldots(4.1)$$

Donde H_0 recibe el nombre de *hipótesis nula*, mientras que H_1 se denomina *hipótesis alternativa*.

En la expresión 7.1 se plantea una hipótesis alternativa *de dos lados*; sin embargo, es posible plantear hipótesis alternativas *de un lado*, generando propuestas como:

$$\begin{aligned} H_0 &: \mu = 1.67\ [m] \\ H_1 &: \mu > 1.67\ [m] \end{aligned} \qquad \ldots(4.2)$$

Para probar una hipótesis es necesario seleccionar una muestra aleatoria, y mediante un estadístico de prueba adecuado determinar si se acepta la hipótesis H_0 o se rechaza, aceptándose entonces la alternativa H_1. Con la finalidad de aceptar o rechazar una hipótesis, deben generarse regiones de aceptación y rechazo, por ejemplo para la hipótesis sobre la media poblacional planteada por (4.1) se tiene:

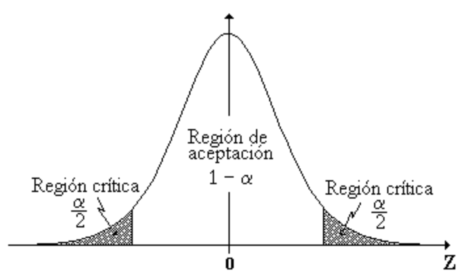


Fig. 4.1 Región de "aceptación"

Definición 4.2
Una prueba de hipótesis estadística para alguna característica desconocida de una población, es cualquier regla que permite rechazar o no rechazar una hipótesis nula, con base en una muestra aleatoria de la población.

ERRORES DE TIPO I Y TIPO II

La decisión que se toma de aceptar o rechazar una hipótesis según los datos observados en una muestra y empleando un estadístico de prueba adecuado, está sujeta a error. En particular se pueden cometer dos tipos de errores. Cuando la hipótesis nula se rechaza siendo que es verdadera se comete un *error del tipo I*, mientras que si se acepta la hipótesis nula cuando es falsa entonces se comete un *error del tipo II*.

Tabla 4.1 Tipos de error en las pruebas de hipótesis

		Estado de la naturaleza	
	Si la hipótesis es / Y la conclusión es	H_0 es verdadera	H_0 es falsa
Decisión	No se rechaza H_0	Ningún error	Error tipo II
	Se rechaza H_0	Error tipo I	Ningún error

Las probabilidades de cometer errores del tipo I y II se denotan mediante α y β respectivamente, es decir

$$\begin{aligned} &P\,(\,rechazar\;\;H_0 \mid H_0\;\;es\;\;verdadera\,) \;=\; P\,(\,error\;\;tipo\;\;I\,) \;=\; \alpha \\ &P\,(\,aceptar\;\;\;H_0 \mid H_0\;\;es\;\;falsa\,) \;=\; P\,(\,error\;\;tipo\;\;II\,) \;=\; \beta \end{aligned} \qquad ...(4.3)$$

Además α recibe el nombre de nivel o tamaño de significación de la prueba.

Tabla 4.2 Probabilidades de error en las pruebas de hipótesis

Si la hipótesis es / Y la conclusión es	H_0 es verdadera	H_0 es falsa
No se rechaza H_0	$1 - \alpha$	β
Se rechaza H_0	α	$1 - \beta$

Definición 4.3
La potencia de una prueba de hipótesis estadística es la probabilidad de rechazar una hipótesis falsa. Es decir:

$$\begin{aligned} Potencia\;\;de\;\;una\;\;prueba \;&=\; P\,(\,rechazar\;\;H_0 \mid H_0\;\;falsa\,) \\ &=\; 1 \;-\; \beta \end{aligned}$$

Como se comentó anteriormente, los resultados de una prueba de hipótesis están sujetos a error, por lo que no se dice que se aprueba la hipótesis nula, es más recomendable decir no se rechaza H_0. El no rechazar H_0 significa que no se tienen suficientes elementos para rechazarla, lo que no necesariamente significa que hay una alta probabilidad de que sea verdadera.

PRUEBAS DE HIPÓTESIS

Además de las pruebas de hipótesis unilaterales y bilaterales como fueron las ecuaciones (4.2) y (4.1), las pruebas se clasifican en simples y compuestas. Las *hipótesis simples* son aquellas que especifican el valor del parámetro al que se refieren, por ejemplo: $H_0 : p = \frac{1}{2}$.

Las *hipótesis compuestas* son aquellas que no especifican el valor del parámetro, por ejemplo: $H_0 : p > \frac{1}{2}$, $H_0 : p \neq \frac{1}{2}$.

Ejemplo 4.1

El tiempo transcurrido X entre dos señales consecutivas de un contador Geiger de partículas radioactivas, es una v.a. con distribución exponencial con parámetro λ.

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda x} & x > 0 \\ 0 & en\ otro\ caso \end{cases}$$

A fin de probar la hipótesis H_0 de que para un material en particular $\lambda = 2$, contra la alternativa H_1, de que $\lambda = 1$, se realiza una sola observación de X y se decide no rechazar H_0 si y sólo si el valor observado de X ocurre en el intervalo $[\,0, 1\,]$. Calcular los tamaños de los errores tipo I y tipo II.

Resolución

Para el error tipo I, se tiene

$$P(error\ tipo\ I) = \alpha = P(rechazar\ H_0 \mid H_0\ es\ verdadera)$$

$$\alpha = P(X > 1 \mid \lambda = 2) = \int_1^{\infty} 2e^{-2x}\, dx = e^{-2}$$

$$\alpha = e^{-2} \approx 0.135$$

$$P(error\ tipo\ II) = \beta = P(No\ rechazar\ H_0 \mid H_0\ es\ falsa)$$

$$\beta = P(0 \le X \le 1 \mid \lambda = 1) = \int_0^1 e^{-\lambda}\, dx = 1 - e^{-1}$$

$$\beta = 1 - e^{-1} \approx 0.632$$

Debe observarse que si al hipótesis nula H_0 es compuesta, entonces no se puede determinar el valor de α, es decir, si $H_0 : \lambda > 2$ entonces:

$$\alpha = P(X > 1 \mid \lambda > 2) = \int_1^{\infty} \lambda\, e^{-\lambda x}\, dx$$

y el valor de α depende de λ. Es por ello, que la hipótesis nula debe ser una hipótesis simple. De manera similar, cuando la hipótesis alternativa es compuesta tampoco se puede obtener el valor de β.

PRUEBAS DE HIPÓTESIS PARA LA MEDIA

Cuando se desea realizar una hipótesis con respecto a la media de una variable aleatoria X, debe suponerse con distribución normal, ya sea porque X se distribuye normalmente o por el cumplimiento del teorema del límite central. Si se considera que la media μ se desconoce pero se conoce la variancia σ^2, entonces la hipótesis bilateral puede formularse como:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 \qquad \ldots (4.4)$$

donde μ_0 es una constante específica, y el estadístico de prueba es

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad \ldots (4.5)$$

donde $Z \sim N(0, 1)$.

Ejemplo 4.2

Considérese una población con distribución normal con parámetros μ desconocido y $\sigma^2 = 4$. Con base en una muestra de 30 observaciones en la cual $\bar{x} = 10$ y $s^2 = 3.5$, determinar si es correcto suponer que $\mu = 12$ con un nivel de significancia de 0.01.

Resolución

La prueba de hipótesis estadística es:

$$H_0 : \mu = 12$$
$$H_1 : \mu \neq 12$$

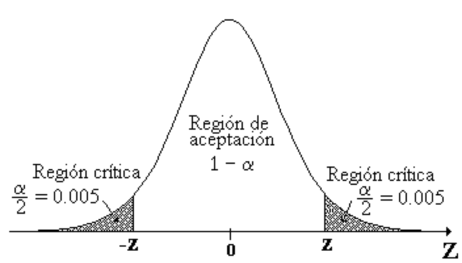
El estadístico de prueba es

$$\bar{X} \sim N\left(\mu, \frac{4}{30}\right)$$

estandarizando

$$Z_0 = \frac{\bar{X} - \mu_0}{\sqrt{\frac{4}{30}}} \sim N(0, 1)$$

Las regiones críticas y de aceptación son



$$P(Z > z) = \frac{\alpha}{2}$$

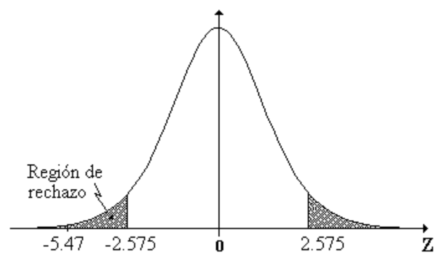
$$P(Z > z) = 0.005$$

De tablas $z = 2.575$

Y evaluando el estadístico de prueba para la muestra dada y suponiendo cierta H_0 se tiene

$$z_0 = \frac{\bar{x} - \mu_0}{\sqrt{\frac{4}{30}}} = \frac{10 - 12}{\sqrt{\frac{4}{30}}} = -5.47$$

De donde se observa que el estadístico de prueba se encuentra fuera de la región de aceptación, es decir $-5.47 < -2.575$ $(z_o < -z)$



Se concluye que, con base es esta muestra, no parece adecuado suponer $\mu = 12$, por lo que H_0 se rechaza.

Ejemplo 4.3

En un estudio del rendimiento de un proceso químico se ha observado que la variancia es $\sigma^2 = 5$, y en los últimos días de operación se han tenido los siguientes rendimientos:

$$89.95 \ , \ 91.3 \ , \ 88.75 \ , \ 90.8 \ , \ 91.6$$

¿Existe razón para creer que el rendimiento es menor a 90?

Resolución

La prueba de hipótesis estadística es: $H_0 : \mu = 90$

$H_1 : \mu < 90$

Suponiendo distribución normal en los datos, el estadístico de prueba es:

$$Z_0 = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

valuando
$$z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{90.48 - 90}{\frac{\sqrt{5}}{\sqrt{5}}} = 0.48$$

Y con $\alpha = 0.05$, de tablas, $z_{0.05} = -1.645$ y puesto que $z_{0.05} < z_0$ entonces:

no se rechaza H_0.

Con base en la información obtenida en la muestra no hay evidencia para rechazar la hipótesis nula, de que la media es igual a 90, con una significancia del 5%.

Cuando en la práctica no se conoce el valor de la variancia poblacional σ^2, puede sustituirse su valor por S_{n-1}^2 si la muestra es grande $(n \geq 30)$ sin tener un efecto perjudicial de consideración.

Si la variancia se desconoce y la muestra es pequeña $(n < 30)$ entonces el estadístico de prueba es

$$T_0 = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

siempre que la población tenga distribución normal.

PRUEBA DE HIPÓTESIS SOBRE LA IGUALDAD DE DOS MEDIAS

Si se desea probar que las medias de dos poblaciones (con distribuciones normales) son iguales, entonces el estadístico de prueba es

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \ldots (4.6)$$

La prueba con alternativa de dos lados es:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned} \qquad \ldots (4.7)$$

Cuando las variancias poblacionales se desconocen, se pueden utilizar las variancias muestrales para las poblaciones, siempre que las muestras sean grandes; si las muestras son pequeñas pero provienen de distribuciones normales con medias y variancias desconocidas, entonces se tienen dos casos.

Si se desea probar sobre la diferencia de medias, entonces el estadístico se modifica restando la diferencia de medias

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \ldots (4.6b)$$

Muestras pequeñas de poblaciones normales y variancias desconocidas pero iguales

El estadístico de prueba es:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{. . . (4.8)}$$

donde $$S_P^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} \qquad \text{. . . (4.9)}$$

y $T_0 \sim t_{(n_1 + n_2 - 2)}$

Muestras pequeñas de poblaciones normales y variancias desconocidas y diferentes

Cuando las variancias son diferentes, entonces no existe un estadístico t exacto para realizar la prueba sobre la igualdad de medias; sin embargo, una buena aproximación la proporciona el estadístico

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad \text{. . . (4.10)}$$

el cual tiene distribución aproximadamente t , i.e., $T_0^* \sim t_{(\nu)}$; donde el número de grados de libertad está dado por

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n + 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 + 1}} - 2 \qquad \text{. . . (4.11)}$$

y debe utilizarse el entero más cercano.

Ejemplo 4.3

Mediciones respecto del esfuerzo cortante obtenidas a partir de pruebas de compresión independientes para dos tipos de suelos dieron los resultados siguientes (mediciones en toneladas por metro cuadrado).

Suelo tipo I	Suelo tipo II
$n_1 = 30$	$n_2 = 35$
$\bar{y}_1 = 1.65$	$\bar{y}_2 = 1.43$
$s_1 = 0.26$	$s_2 = 0.22$

¿Difieren los dos suelos con respecto al esfuerzo cortante promedio, a un nivel de significación de $1\,\%$?

Resolución

$H_0 \;:\; \mu_1 - \mu = 0$

$H_1 \;:\; \mu_1 - \mu_2 \neq 0$

El estadístico de prueba es $$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$$

valuando $$z_0 = \frac{1.65 - 1.43}{\sqrt{\dfrac{(0.26)^2}{30} + \dfrac{(0.22)^2}{35}}} = 3.65$$

de tablas se obtiene

$z = 2.575$ con $\alpha = 0.01$

$\therefore$ se rechaza H_0.

Conclusión: Con base en la información contenida en la muestra, se rechaza la hipótesis nula, las medias son iguales; en favor de la hipótesis alterna, las medias son diferentes, con una significación del 1%.

PRUEBAS DE HIPÓTESIS PARA LA VARIANCIA

Si se desea probar la variancia de una población con distribución normal, entonces el estadístico de prueba es

$$\mathrm{X}_0^2 = \frac{(n-1)S^2}{\sigma_2^2} \qquad \ldots (4.12)$$

donde S^2 es la variancia muestral y $\mathrm{X}_0^2 \sim \chi^2_{(n-1)}$.

La prueba de hipótesis de dos lados es:

$$H_0 \;:\; \sigma^2 = \sigma_0^2$$

$$H_1 \;:\; \sigma^2 \neq \sigma_0^2$$

y la hipótesis nula se rechazaría si $$\chi_0^2 > \chi^2_{\frac{\alpha}{2},\,n-1} \qquad \ldots (4.13)$$

o bien si $\chi_0^2 < \chi^2_{1-\frac{\alpha}{2}, n-1}$ . . . (4.14)

Ejemplo 4.4

La dispersión o variancia, de tiempos de acarreo en un proyecto de construcción son de gran importancia para el sobrestante, ya que los tiempos muy variables de acarreo originan problemas en la programación de los trabajos. El encargado de los transportes dice que el intervalo de tiempo de acarreo no debe ser mayor que 40 minutos (este intervalo es la diferencia entre el tiempo mayor y el menor). Si se supone que estos tiempos de acarreo están distribuidos en forma aproximadamente normal, el sobrestante cree que la afirmación acerca de los límites quiere decir que la desviación estándar σ debe ser aproximadamente 10 minutos. Se midieron en realidad 15 tiempos de acarreo y se obtuvo un promedio de 142 minutos y una desviación estándar de 12 minutos. ¿Podría refutarse la afirmación de $\sigma = 10$ en el nivel de significancia del 5%?

Resolución

Se desea probar:

H_0: $\sigma^2 = 100$

H_1: $\sigma^2 > 100$

El estadístico de prueba es: $\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{(14)(12)^2}{100} = 20.16$

La región de rechazo es: $\chi^2 > \chi^2_{0.05,14} = 23.6848$

Puesto que $\chi_0^2 < \chi^2_{0.05} \Rightarrow 20.16 < 23.64$

No se rechaza H_0. Con base en la información de la muestra no hay suficiente evidencia para concluir que la desviación estándar es superior a 10 minutos, con $\alpha = 0.05$.

PRUEBAS DE HIPÓTESIS PARA LA IGUALDAD DE VARIANCIAS

Para probar la igualdad de dos variancias de poblaciones normales con parámetros μ_1, σ_1^2, μ_2 y σ_2^2 desconocidas, se utiliza el estadístico

$$F_0 = \frac{S_1^2}{S_2^2} \qquad \text{. . . (4.15)}$$

donde $F \sim F_{n_1-1, n_2-1}$

y la prueba de dos lados quedaría como:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad \text{. . . (4.16)}$$

la hipótesis H_0 sería rechazada si:

$$f_0 > f_{\frac{\alpha}{2}, n_1 - 1, n_2 - 1} \qquad \text{. . . (4.17a)}$$

o bien si

$$f_0 < f_{1 - \frac{\alpha}{2}, n_1 - 1, n_2 - 1} \qquad \text{. . . (4.17b)}$$

Para probar la hipótesis alternativa de un solo lado, quedando la prueba

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2 \qquad \text{. . . (4.18)}$$

Para rechazar H_0 debe cumplirse

$$f_0 > f_{\alpha, n_1 - 1, n_2 - 1} \qquad \text{. . . (4.19)}$$

Un concepto muy utilizado en las pruebas de hipótesis es el *nivel de significación alcanzado.* El nivel de significación alcanzado en una prueba, p, es un estadístico que representa el mínimo valor de α para el cual se rechaza la hipótesis nula. Es decir:

Si $\alpha \geq p$ se rechaza H_0.

Ejemplo 4.5

Dos máquinas producen piezas metálicas. Interesa la variancia del peso de estas piezas. Se han colectado los siguientes datos.

	Máquina 1	Máquina 2
n	25	30
$\bar{x}$	0.984	0.907
s^2	13.46	9.65

a) Probar la hipótesis de que las variancias de las dos máquinas son iguales. Emplear $\alpha = 0.05$.

b) Probar la hipótesis de que las dos máquinas producen piezas que tienen el mismo peso medio. Utilizar $\alpha = 0.05$.

Resolución

a) $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Es estadístico de prueba es $F_0 = \dfrac{S_1^2}{S_2^2}$

de donde $f_0 = \dfrac{13.46}{9.65} = 1.395$

De tablas: $f_{\frac{\alpha}{2}, 24, 29} = 2.15$

Puesto que $f_{\frac{\alpha}{2}, 24, 29} = 2.15 > 1.39 = f_0$

H_0 no se rechaza.

Conclusión: Con base en la información contenida en las muestras, no puede rechazarse la hipótesis nula, de que las varianzas son iguales, con una significación del 5%.

b) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

El estadístico de prueba es $T_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

donde $S_p^2 = \dfrac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}$

valuando: $t_0 = \dfrac{0.984 - 0.907}{\sqrt{11.37}\sqrt{\dfrac{1}{25} + \dfrac{1}{30}}} = 0.084$

De tablas, $t_{\frac{\alpha}{2}, n_1 + n_2 - 2} = t_{0.025\,,\,53} \approx 2.007$

Puesto que $t_{0.025\,,\,53} \approx 2.007 > 0.084 = t_0$

H_0 no se rechaza.

Conclusión: Con base en la información contenida en las muestras, no puede rechazarse la hipótesis nula, de que las medias son iguales, con una significación del 5%.

PRUEBA DE HIPÓTESIS SOBRE UNA PROPORCIÓN

La proporción es un caso particular de la media, por lo que no debe sorprender que el estadístico de prueba sea muy similar.

$$Z = \frac{X - n\, p_0}{\sqrt{n p_0 (1 - p_0)}}$$

o bien, al dividir el numerador y el denominador por n se tiene

$$Z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 (1 - p_0)}{n}}}$$

y las hipótesis se pueden plantear como:

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

para una prueba de dos lados.

Ejemplo 4.6

Considérese que cierta empresa dedicada a realizar estudios estadísticos observó que el 54% de 2207 personas entrevistadas pensaba que los trámites de titulación son demasiado complicados. ¿Se puede concluir con un nivel de significación de 5%, que la mayor parte de las personas de esta población piensan que los trámites de titulación son demasiado complicados?

Resolución

La prueba de hipótesis es:

$$H_0 \ : \ p = 0.5$$
$$H_1 \ : \ p > 0.5$$

El estadístico de prueba es $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$

valuando el estadístico de prueba $z_0 = \dfrac{0.54 - 0.5}{\sqrt{\dfrac{(0.5)(0.5)}{2207}}} = 3.76$

La región de rechazo es $Z > z_\alpha$

Puesto que $\alpha = 0.05$, $z_{0.05} = 1.645$

Conclusión: H_0 se rechaza. Con base en los datos de la muestra, existe evidencia para concluir que la mayoría de la población estudiada piensa que el sistema de titulación es muy complicado, con una significación del 5%.

PRUEBA DE HIPÓTESIS PARA UNA IGUALDAD DE PROPORCIONES

Nuevamente, el caso de una prueba de hipótesis para la igualdad de dos proporciones es equivalente a la de igualdad de medias, puesto que se utiliza el TCL para determinar la distribución del estadístico de prueba.

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1 - \hat{P})\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]}}$$

donde la estimación del parámetro común p está dada por:

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

Este estadístico considera la hipótesis nula $H_0 : p_1 - p_2 = 0$

o bien, aproximando con el estadístico utilizado en los intervalos de confianza

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\hat{P}_1 \hat{Q}_1}{n_1} + \frac{\hat{P}_2 \hat{Q}_2}{n_2}}}$$

En ambos casos para muestras grandes, donde pueda aplicarse el TCL.

Las hipótesis se pueden plantear como:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

para una prueba de dos lados.

Ejemplo 4.7

Un ingeniero industrial está tratando de determinar si un proceso nuevo reduce el número de imperfecciones en el acabado de un artículo. En 50 muestras del artículo con el proceso actual, 43 contenían ciertas imperfecciones. Con el proceso nuevo, en otras 50 muestras sólo 22 mostraron imperfecciones. ¿El proceso nuevo reduce significativamente la proporción de artículo que tienen imperfecciones? Usar $\alpha = 0.025$.

Resolución

Usando los subíndices
1: Proceso actual.
2: Proceso nuevo.

Lo que se desea probar es si $p_2 < p_1$, o lo que es lo mismo $p_1 - p_2 > 0$, por lo que las hipótesis se plantean:

$$H_0 : p_1 - p_2 = 0$$
$$H_1 : p_1 - p_2 > 0$$

$$\hat{p}_1 = \frac{43}{50} = 0.86 \qquad , \qquad \hat{p}_2 = \frac{22}{50} = 0.44$$

Estadístico

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

$$z_0 = \frac{0.86 - 0.44}{\sqrt{\frac{(0.86)(0.14)}{50} + \frac{(0.44)(0.56)}{50}}} = 4.9$$

Región de Rechazo: $Z > z_{0.025} = 1.90$

Conclusión: H_0 se rechaza. Con base en la información contenida en las muestras, se rechaza la hipótesis nula, que indica que las proporciones son iguales, en favor de la hipótesis alterna, que indica que la proporción para el proceso nuevo es menos que para el proceso actual, con una significación del 2.5%

PRUEBA DE BONDAD DE AJUSTE JI CUADRADA

Hasta este momento, se han estudiado pruebas de hipótesis estadísticas sobre parámetros poblacionales, en situaciones donde se conoce (o se supone) la distribución de las variables aleatorias. Otro tipo de pruebas se da cuando la distribución de la variable aleatoria bajo estudio se desconoce, y por lo tanto se desea "probar" si sigue una distribución teórica en particular. A este tipo de pruebas se les llama pruebas de bondad de ajuste.

En particular, para la prueba de bondad de ajuste ji cuadrada, considérese una muestra aleatoria de tamaño n de la distribución de una variable aleatoria X dividida en k clases (intervalos exhaustivos y mutuamente excluyentes), y sea O_i $i = 1, 2, \ldots, k$, el número de observaciones de la i-ésima clase. Si la hipótesis nula es

$$H_0 : \textit{La distribución de probabilidad es } f_X(x;\theta)$$

donde $f(x;\theta)$ es una distribución que se encuentra completamente especificada, incluyendo todos sus parámetros, entonces la hipótesis nula es simple.

Con el objeto de deducir un estadístico adecuado para H_0 considérese el caso en el que sólo se tienen dos clases, $k = 2$, entonces O_1 representa el número de observaciones en la clase 1 y O_2 el número de observaciones de la clase 2 con $O_2 = n - O_1$. Para las dos categorías excluyentes las probabilidades son p_1 y $p_2 = 1 - p_1$, entonces bajo la hipótesis nula la probabilidad de la muestra agrupada es igual a la función de probabilidad binomial con parámetros n y p_1, es decir, la variable O_1 tiene una distribución binomial. Estandarizando la variable aleatoria se tiene

$$Y = \frac{O_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

y si n es suficientemente grande entonces Y tiene una distribución aproximadamente normal estándar, por lo que al elevar al cuadrado se obtiene una variable aleatoria ji cuadrada con un grado de libertad.

$$\frac{(O_1 - np_1)^2}{np_1(1-p_1)} = \frac{(O_1 - np_1)^2(p_1+p_2)}{np_1p_2}$$

$$= \frac{p_2(O_1 - np_1)^2 + p_1(O_1 - np_1)^2}{np_1 p_2}$$

$$= \frac{(O_1 - np_1)^2}{np_1} + \frac{(n - O_2 - n(1 - p_2))^2}{np_2}$$

$$= \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2}$$

$$= \sum_{i=1}^{2} \frac{(O_i - np_i)^2}{np_i}$$

Por lo que el estadístico $\sum_{i=1}^{2} \frac{(O_i - np_i)^2}{np_i}$ tiene aproximadamente una distribución ji cuadrada con un grado de libertad, siempre que n sea lo suficientemente grande. De forma análogamente, para $k \geq 2$ clases, el estadístico

$$\sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$$

tiene aproximadamente una distribución ji cuadrada con $k - 1$ grados de libertad.

En resumen, la prueba de bondad de ajuste ji cuadrada consiste en comparar la frecuencia observada de la variable aleatoria en cada uno de los intervalos de clase de una tabla de distribución de frecuencia (O_i) y el valor esperado de la distribución hipotética $(E_i = n\ p_i)$. El estadístico de prueba es

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

donde $X_0^2 \sim \chi^2_{(k-p-1)}$.

El estadístico X_0^2 tiene distribución ji cuadrada con $k - p - 1$ grados de libertad, donde k es el número de intervalos de clase y p es el número de parámetros de la distribución hipotética. Por ejemplo $p = 0$ para una distribución discreta uniforme; $p = 1$ para una distribución de Poisson y $p = 2$ para una distribución normal.

La hipótesis nula, de que la distribución se ajusta a la considerada se rechaza si $\chi_0^2 > \chi^2_{\alpha, k-p-1}$.

Para realizar esta prueba, no se requiere que el ancho de clase sea constante, lo que se requiere es que la frecuencia esperada en cada intervalo no sea cero; sin embargo, el valor mínimo no se ha establecido de forma única, la mayoría de los autores utilizan los números 3, 4 ó 5 como mínimos, y es claro que con 5 se puede asegurar la aproximación a la distribución normal para construir el estadístico χ^2 , por lo que se recomienda que $n p_i \geq 5$, o bien, nunca menor que 4.

La prueba de bondad de ajuste puede utilizarse también cuando la variable es continua; sin embargo, debe hacerse énfasis en que la prueba de bondad de ajuste ji cuadrada es de naturaleza discreta, en el sentido de que compara frecuencias de observación y esperadas para un número finito de categorías. Para muestras muy grandes, la potencia de la prueba tiende a 1, lo que significa que es casi seguro rechazar la hipótesis nula.

Ejemplo 4.8

En un proceso de fabricación de tela, se cuenta con el número de defectos por metro cuadrado en 50 muestras , cada una de un metro cuadrado, se observaron los siguientes resultados

Número de defectos	Frecuencia de observación
0	0
1	3
2	5
3	10
4	14
5	8
6 o más	10

Probar la hipótesis de que los datos provienen de una distribución de Poisson. Utilizar $\alpha = 0.05$.

Resolución

Si se considera que Y , el número de defectos por metro cuadrado tiene una distribución de Poisson, entonces

$$Y \sim Poisson\ (\lambda)$$

El estimador puntual de λ es: $\hat{\lambda} = \overline{Y}$, de donde: $\hat{\lambda} = \overline{y} = \dfrac{\sum_{i=1}^{7} y_i f_i}{50} = \dfrac{199}{50} = 3.98$

La hipótesis estadística es:

H_0 : El número de defectos tiene distribución de Poisson con parámetro $\lambda = 3.98$.

H_1 : El número de defectos no tiene distribución de Poisson con parámetro $\lambda = 3.98$.

Para determinar el estadístico de prueba se obtiene la siguiente tabla:

y_i	$O_i = f_i$	$p_i = \dfrac{e^{-\hat{\lambda}}\hat{\lambda}^{y_i}}{y_i!}$	$E_i = np_i$
0	0	0.01869	0.934
1	3	0.07437	3.718
2	5	0.14799	7.400
3	10	0.19634	9.817
4	14	0.19536	9.768
5	8	0.15555	7.775
6 o más	10	0.21175	10.588

Puesto que para el primer intervalo, $Y_1 = 0$ se tiene que $E_1 = 0.934 < 3$, se agrupan los primeros dos intervalos, obteniéndose ahora la siguiente tabla

y_i	$O_i = f_i$	$P_i = \dfrac{e^{-\hat{\lambda}}\hat{\lambda}^{y_i}}{y_i!}$	$E_i = np_i$
1 o menos	3	0.09306	4.653
2	5	0.14799	7.400
3	10	0.19634	9.817
4	14	0.19536	9.768
5	8	0.15555	7.775
6 o más	10	0.21175	10.588

El estadístico de prueba es:

$$\chi_0^2 = \sum_{i=1}^{6} \frac{\left(O_i - E_i\right)^2}{E_i}$$

$$\chi_0^2 = \frac{(3 - 4.653)^2}{4.653} + \frac{(5 - 7.400)^2}{7.400} + \dots + \frac{(10 - 10.588)^2}{10.588}$$

$$\chi_0^2 = 4.89$$

Por otro lado, con $k = 6$ intervalos, $p = 1$ parámetros (λ), se tiene que:

$$\chi^2_{\alpha, k-p-1} = \chi^2_{0.05, 4} = 9.487 > \chi_0^2$$

No se rechaza la hipótesis nula, de que la distribución es Poisson con parámetro $\lambda = 3.98$, con una significación del 5%.

Tópicos especiales: Prueba de bondad de ajuste Kolgomorov-Smirnov

La prueba de bondad de ajuste χ^2 es muy útil; sin embargo, cuando la v.a. es continua, para realizar el agrupamiento se requiere de una gran cantidad de datos, con lo que el agrupamiento se vuelve más complicado, puesto que se deben buscar clases que no contengan menos 4 valores esperados. Cuando la v.a. bajo prueba es continua, el estadístico Kolgomorov-Smirnov resulta más adecuado.

Considérese la hipótesis nula, en la cual se especifica de manera completa la función de distribución de la variable aleatoria X,

$$H_0 : F_X(x) = F_{0_X}(x)$$

utilizando los estadísticos de orden $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ de una muestra aleatoria de tamaño n y definiendo la función de distribución acumulativa como

$$S_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \dfrac{k}{n} & x_{(k)} \le x < x_{(k+1)} \\ 1 & x \ge x_{(n)} \end{cases}$$

es decir, $S_n(x)$ es la proporción de los valores de la muestra que son iguales o menores a x.

El estadístico de Kolgomorov-Smirnov se define como

$$D_{0_n} = \max_{\forall x} \left| S_n(x) - F_{0_X}(x) \right|$$

donde $F_{0x}(x)$ se puede valuar puesto que es la distribución bajo prueba, y D_n es un estadístico independiente de la distribución F_{0x}. Los valores críticos del estadístico de $k - S$ (Kolgomorov-Smirnov) se muestran en el apéndice A, y la hipótesis nula se rechaza si $D_{0_n} > D_n$.

Ejemplo 4.9

Cierta empresa productora de champiñones ha registrado la demanda diaria de champiñón fresco en toneladas, obteniéndose los siguientes valores

38	67	28	49	47	59	51	57	52	56
35	76	58	48	63	34	68	53	26	36
32	61	33	48	42	72	66	59	43	44

Utilizar la prueba de bondad Kolmogorov-Smirnov para probar que la demanda diaria de champiñones tiene una distribución normal con media $\mu = 50$ y desviación estándar $\sigma = 13$. Usar $\alpha = 0.05$.

Resolución

H_0 : Los datos tienen distribución normal con media 50 y desviación estándar 13.

H_1 : Los datos no tienen distribución normal con media 50 y desviación estándar 13.

Ordenando los datos $x_{(i)}$ y calculando $S_n(x)$, $F_{0x}(x)$ y $\left|S_n(x) - F_{0x}(x)\right|$ se tiene

i	$x_{(i)}$	$S_n(x) = \frac{i}{n}$	$F_{0_X}(x) = P\left(X \le x_{(i)}\right)$	$\left|S_n(x) = F_{0_X}(x)\right|$
1	25	$\frac{1}{30}$	0.0272	0.006
2	28	$\frac{2}{30}$	0.0453	0.021
3	32	$\frac{3}{30}$	0.0831	0.017
4	33	$\frac{4}{30}$	0.0955	0.038
5	34	$\frac{5}{30}$	0.1092	0.057
6	35	$\frac{6}{30}$	0.1243	0.076
7	36	$\frac{7}{30}$	0.1408	0.093
8	38	$\frac{8}{30}$	0.178	0.089
9	42	$\frac{9}{30}$	0.2692	0.031
10	43	$\frac{10}{30}$	0.2951	0.038
11	44	$\frac{11}{30}$	0.3222	0.044
12	47	$\frac{12}{30}$	0.4087	0.009
13	48	$\frac{13}{30}$	0.4389	0.006
14	48	$\frac{14}{30}$	0.4389	0.028
15	49	$\frac{15}{30}$	0.4693	0.031
16	51	$\frac{16}{30}$	0.5307	0.003
17	52	$\frac{17}{30}$	0.5611	0.006
18	53	$\frac{18}{30}$	0.5913	0.009
19	56	$\frac{19}{30}$	0.6778	0.044

i	$x_{(i)}$	$S_n(x) = \frac{i}{n}$	$F_{0_X}(x) = P\left(X \le x_{(i)}\right)$	$\left\|S_n(x) = F_{0_X}(x)\right\|$
20	57	$\frac{20}{30}$	0.7049	0.038
21	58	$\frac{21}{30}$	0.7308	0.031
22	59	$\frac{22}{30}$	0.7556	0.022
23	59	$\frac{23}{30}$	0.7556	0.011
24	61	$\frac{24}{30}$	0.8013	0.001
25	63	$\frac{25}{30}$	0.8413	0.008
26	66	$\frac{26}{30}$	0.8908	0.024
27	67	$\frac{27}{30}$	0.9045	0.005
28	68	$\frac{28}{30}$	0.9169	0.016
29	72	$\frac{29}{30}$	0.9547	0.012
30	76	1	0.9773	0.023
			Máximo	0.093

De donde $D_{0n} = 0.093$

Y de tablas $D_{n,\,\alpha} = D_{30,0.05} = 0.23$

Y puesto que $D_{0_n} = 0.093 < 0.23 = D_{n,\alpha}$

H_0 no se rechaza.

Conclusión: A partir de la información contenida en la muestra, no puede rechazarse la hipótesis nula, de que los datos provienen de una población con distribución normal con media $\mu = 50$ y desviación estándar $\sigma = 13$, con una significación del 5%.

Apéndice A

Estadistico D_n de Kolmogorov-Smirnov

n	2	0.15	0.1	5	1
1	0.9	0.925	0.95	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	0.594	0.525	0.564	0.624	0.733
5	0.446	0.474	0.51	0.565	0.669
6	0.41	436	0.47	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	0.36	0.388	0.432	0.514
10	0.322	0.342	0.368	0.41	0.49
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.45
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	0.25	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.233	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.2	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	27

BIBLIOGRAFÍA

Hines, William W. y Montgomery, Douglas C. - Probabilidad y Estadística para ingeniería, cuarta edición.- CECSA.- México, 2005.

Milton, Susan J. Y Arnold, Jesse C.- Probabilidad y Estadística para con aplicaciones para ingeniería y ciencias computacionales, cuarta edición.- McGraw-Hill.- México, 2004.

Devore, Jay L.- Probabilidad y Estadística para ingeniería y ciencias, séptima edición.- Cengage Learning.- México, 2008.

Mendenhall, William III. et al.- Introducción a la Probabilidad y Estadística.- Décimo cuarta edición.- Cengage Learning.- México 2015.

Wackerly Dennis D.- Mendenhall, William, *et al*.- Estadística Matemática con Aplicaciones, sexta edición.- Editorial Thomson.- México, 2002.

Walpole, Ronald E., *et al*.- Probability and Statistics for Engineers and Scientists.- Pearson.- USA, 2007.

Montgomery, Douglas C. y Runger, George C.-Probabilidad y Estadística aplicadas a la Ingeniería, segunda edición.- Limusa-Wiley.- México, 2002.

Scheaffer, Richard L. y McClave, James T.- Probabilidad y Estadística para Ingeniería.- Grupo Editorial Iberoamérica.- México, 1993.

Canavos, George C.- Probabilidad y Estadística Aplicaciones y Métodos.- McGraw-Hill.- México, 1988.

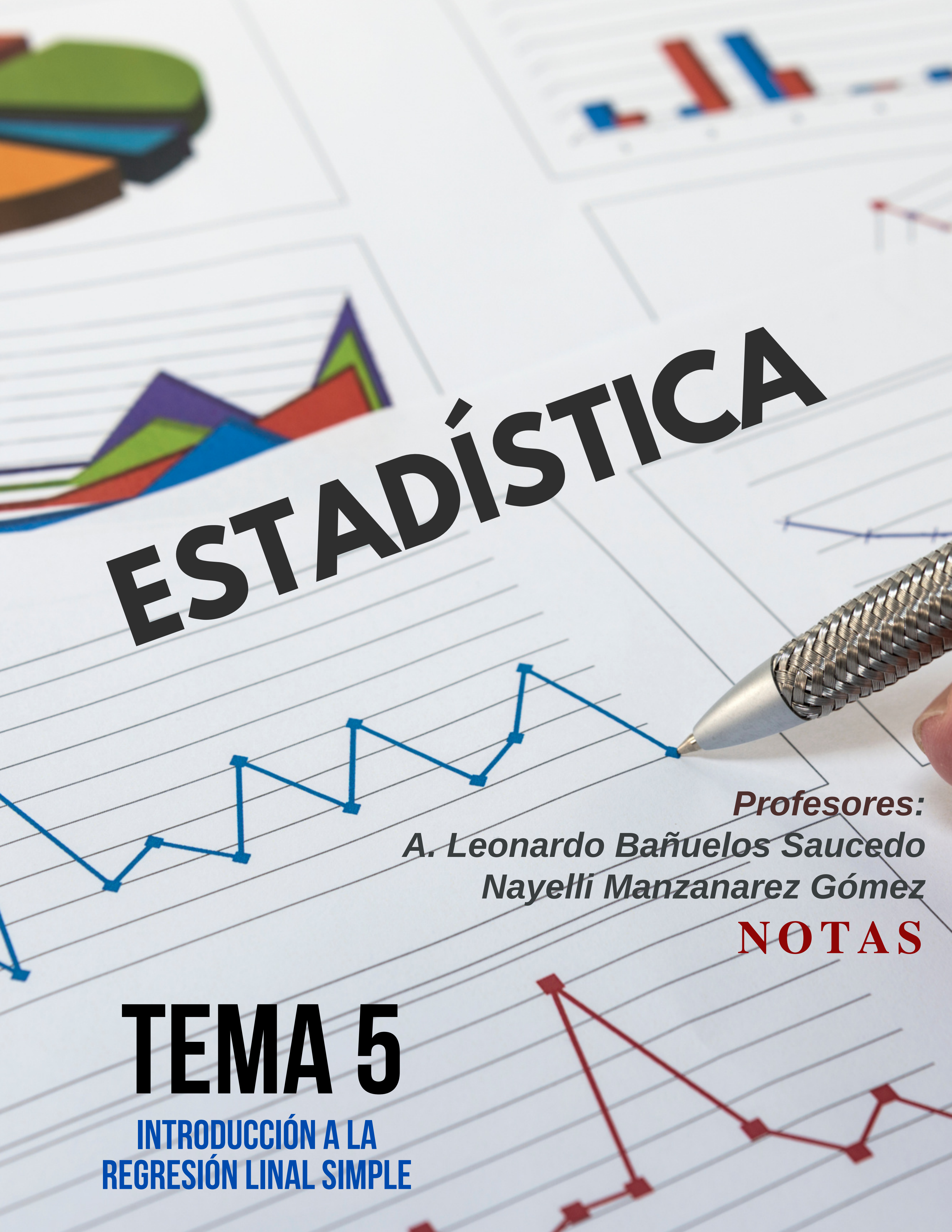
Meyer, Paul L.- Probabilidad y Aplicaciones Estadísticas.- Addison Wesley Iberoamericana.- México, 1992.

Spiegel, Murray R. et al.- Probabilidad y Estadística, cuarta edición.- Mc Graw-Hill.-México 2013.

Borras García, Hugo E., *et al*.- Apuntes de Probabilidad y Estadística.-Facultad de Ingeniería.- México, 1985.

Rosenkrantz, Walter A.- Introduction to Probability and Statistics for Scientists and Engineers.- McGraw-Hill.- EE.UU., 1997.

Ziemer, Rodger E.- Elements of Engineering Probability & Statistics.- Prentice Hall.- USA 1997.



ESTADÍSTICA

Profesores:
A. Leonardo Bañuelos Saucedo
Nayelli Manzanarez Gómez

NOTAS

TEMA 5

INTRODUCCIÓN A LA REGRESIÓN LINAL SIMPLE

TEMA V
INTRODUCCIÓN A LA REGRESIÓN LINEAL SIMPLE

INTRODUCCIÓN

La mayoría de las ramas de las matemáticas se dedica a estudiar variables que están relacionadas de manera determinística, esto es, que una vez que se sabe el valor de x, el valor de y puede conocerse por completo; sin embargo, existen muchas variables x y y que no están relacionadas determinísticamente, por ejemplo:

1) La estatura y el peso de una persona
2) Consumo de un artículo y su precio
3) Coeficiente de inteligencia y rendimiento de una persona

Para estudiar estos casos se utiliza la estadística multivariable, que se encarga de estudiar las relaciones entre dos o más variables aleatorias. Una técnica muy utilizada por la estadística multivariable es el análisis de regresión.

DISTRIBUCIÓN MULTINOMIAL

Como antecedente de la forma en la que se pueden relacionar las variables aleatorias, considérese una generalización de la distribución binomial a más variables. Esta generalización surge del hecho de considerar que cada ensayo pueda tener más de dos resultados posibles. Es decir, los resultados se pueden clasificar en: bueno, malo, regular; alto, medio, bajo; A, B, C, D, etc.

Sean $X_1, X_2, \ldots, X_k$; k variables aleatorias conjuntas que representan el número de éxitos del primer tipo x_1, el numero de éxitos del segundo tipo x_2, etc. Con $\sum_{i=1}^{k} x_i = n$ que se obtienen en n ensayos independientes, cada uno de los cuales permite k resultados mutuamente excluyentes, cuyas probabilidades son $p_1, p_2, \ldots, p_k$, con $\sum_{i=1}^{k} p_i = 1$, entonces las variables aleatorias tienen distribución multinomial con parámetros n y $p_1, p_2, \cdots, p_k$

$$f_{X_1 X_2 \ldots X_k}(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\; x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

para $x_i = 0, 1, \ldots, n$; $1 \le i \le k$ y $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$.

Definición 5.1

Si un ensayo determinado pude resultar en cualquiera de los k resultados $E_1, E_2, \ldots, E_k$ con probabilidades $p_1, p_2, \ldots, p_k$, entonces la distribución conjunta de las variables aleatorias X_1, $X_2, \ldots, X_k$, que representan el número de ocurrencias para $E_1, E_2, \ldots, E_k$ en n intentos independientes es:

$$f_{X_1 X_2 \ldots X_k}(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\ x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

con $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$.

Ejemplo 5.1

Supóngase que de un proceso de producción se seleccionan, de manera aleatoria, 25 artículos. Este proceso de producción por lo general produce un 90% de artículos listos para venderse y un 7% reprocesables. ¿Cuál es la probabilidad de que 22 de los 25 artículos estén listos para venderse y que dos sean reprocesables?

Resolución

Del enunciado:

p_1: Probabilidad de que el artículo este listo para venderse

p_2: Probabilidad de que sea reprocesable

p_3: Probabilidad de que no este listo y no sea reprocesable

$p_1 = 0.9$; $p_2 = 0.07$; $p_3 = 0.03$

Si X_1 es la variable aleatoria que representa el número de artículos listos para venderse, X_2 el número de artículos reprocesables y X_3 el resto de los artículos, entonces las variables aleatorias tienen distribución multinomial, por lo que:

$$P(X_1 = 2.2, X_2 = 2, X_3 = 1) = \frac{25!}{22!\ 2!\ 1!}(0.9)^{22}(0.07)^2(0.03)^1$$
$$= 0.09988$$

REGRESIÓN LINEAL

La regresión proporciona la posible relación entre las variables mediante una ecuación, con el objetivo de predecir una de ellas (variable dependiente o variable de salida) en función de la otra u otras variables (variable(s) independiente(s) o variable(s) de entrada). Existen dos tipos de regresión en general:

1) Regresión simple

2) Regresión múltiple

La regresión simple, se utiliza cuando se relacionan dos variables mientras que la múltiple se utiliza para más de dos variables. En este curso, se estudiará la regresión simple, en la cual, el tipo de curva puede ser lineal, polinomial, exponencial y algunos otros modelos. En este tema nos concentraremos en el estudio de la regresión lineal simple, haciendo lo que se llama análisis estadístico bivariado, denominado así por el manejo de dos conjuntos de datos . El caso más común de análisis estadístico bivariado es el ajuste por mínimos cuadrados.

AJUSTE POR MÍNIMOS CUADRADOS

Partiendo de que se desea obtener un modelo lineal para la variable independiente y en función de la variable dependiente x, se escribe

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde ε es un error aleatorio que se obtiene debido al modelo.

Sin considerar el error el modelo se puede escribir como

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde la pendiente y la ordenada al origen tiene un acento circunflejo para indicar que se trata de aproximaciones de los verdaderos parámetros.

Considerando el valor real y el aproximado para cada punto, se puede obtener la suma de los errores cuadrados, esto es:

$$SEC = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

obteniendo el mínimo de SEC en función de $\hat{\beta}_0$ y $\hat{\beta}_1$ se tiene:

$$\frac{\partial\, SEC}{\partial\, \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial\, SEC}{\partial\, \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

de donde

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

o bien:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_i \right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}}$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

si se simplifica la notación, mediante:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

entonces:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores (aproximaciones) insesgados de β_0 y β_1, que son los parámetros que se desea obtener.

Cabe aclarar que la ecuación de regresión que se obtenga es válida solo para parejas de valores comprendidos en el rango donde se ha experimentado.

Diagrama de Dispersión

Una vez que se ha determinado la ecuación de regresión, es útil la representación gráfica de los puntos de datos en el plano $x\,y$ en lo que se denomina *diagrama de dispersión*. Cuando la regresión aplicada es lineal, los puntos deben mostrar esa tendencia, aunque no debe esperarse que los puntos se ubiquen exactamente en una recta.

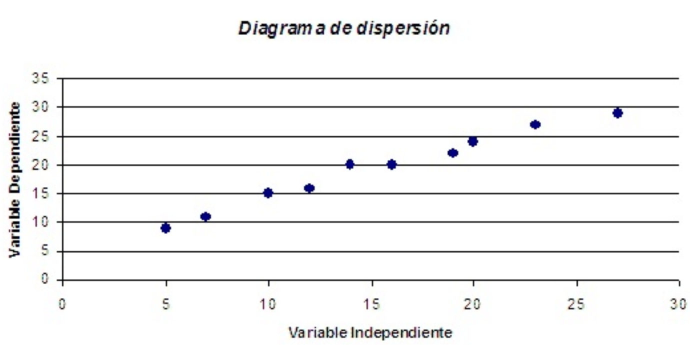


Fig. 5.1. Diagrama de dispersión

Covariancia

Definición 5.2
La covariancia de dos conjuntos de datos, es una medida de la dispersión promedio de los datos con respecto a sus medias. Se denota por Cov, y se define mediante:

$$Cov = \frac{SS_{xy}}{n}$$

donde

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

es la suma sobre las x (equis) y las y (yes)

o bien,

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

Coeficiente de determinación

Definición 5.3
El coeficiente de determinación lineal r^2 de la muestra es:

$$r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}$$

y representa la proporción de variación de y observada que se explica mediante el modelo de regresión.

De un conjunto de datos muestrales apareados (x,y) se puede obtener un ajuste por mínimos cuadrados, pero ¿qué tan bueno es el ajuste? ¿Qué tanto sirve para explicar el comportamiento de y el saber el valor de x? Si el valor de y es independiente de la x, entonces el valor más representativo de y sería $\bar{y}$, y para cada valor real y_i se obtendría un error con respecto de $\bar{y}$

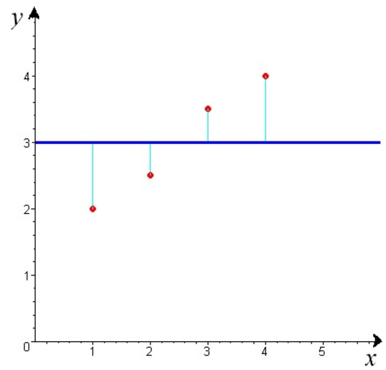


Fig. 5.2. Distancias $y_i - \bar{y}$

La suma de estos errores al cuadrado está dado por $SS_{yy} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$

mientras que al realizar el ajuste a una recta por mínimos cuadrados, el error se obtiene con la recta de ajuste $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

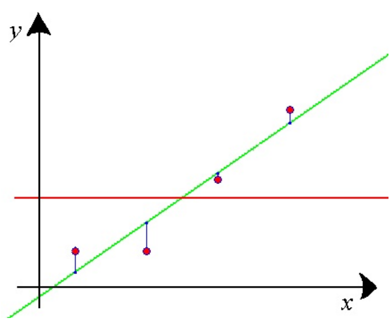


Fig. 5.3. Distancias $y_i - \hat{y}$

y la suma de errores al cuadrado es

$$\begin{aligned} \text{SEC} &= \sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2 \\ &= \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2 \end{aligned}$$

Es claro que $\dfrac{\text{SEC}}{SS_{yy}}$ es una proporción menor o igual que uno, e indica la variación que el modelo no aclara o no explica. Si $\dfrac{\text{SEC}}{SS_{yy}} = 0$, entonces $\text{SEC} = 0$ y los puntos experimentales u observados están contenidos todos sobre la recta de ajuste, por lo que no existe variación no explicada. Si $\frac{\text{SEC}}{SS_{yy}} = 1$ entonces $\text{SEC} = SS_{yy}$ y la recta obtenida es un horizontal que coincide con $\bar{y}$. Por lo que el modelo no explica nada adicional al promedio. De lo anterior se define el coeficiente de determinación.

Definición 5.4
El coeficiente de determinación muestral r^2 se determina mediante

$$r^2 = 1 - \frac{\text{SEC}}{SS_{yy}}$$

y representa la proporción de variación de y observada que se explica mediante el modelo de regresión.

Cuando r^2 es muy cercano a 1, el modelo explica en un mayor porcentaje el comportamiento de la variable independiente, pero si r^2 es cercana a 0, entonces el modelo proporcione muy poca explicación.

Para calcular las sumas de cuadrados, del error y sobre y, se pueden utilizar las siguientes fórmulas operativas:

$$\text{SEC} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \beta_0 \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i y_i$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i^2\right)}{n}$$

Coeficiente de correlación

El coeficiente de correlación proporciona el grado de asociación lineal de las variables x y y, en otras palabras. El coeficiente de correlación que se estudia en este curso es de tipo simple, es decir, considera solo dos variables asociadas en forma lineal.

Definición 5.5
El coeficiente de correlación r de la muestra es:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

y proporciona el grado de asociación lineal de las variables x y y,

donde

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Por la relación que guardan las variables, existen tres tipos de correlación:

Correlación directa o positiva: Se obtiene cuando al aumentar (disminuir) el valor de la variable independiente, aumenta (disminuye) también el valor de la variable dependiente. Si la correlación toma el valor de 1 se tiene correlación positiva perfecta.

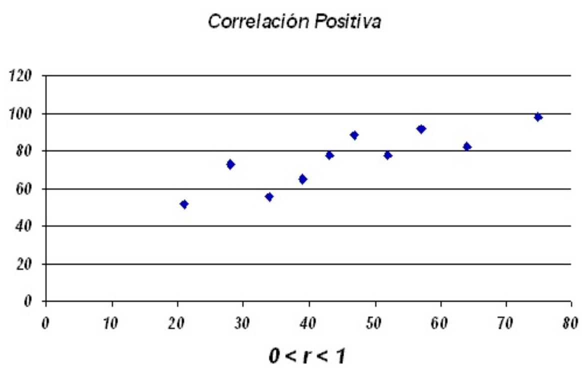


Fig 5.4. Correlación Positiva

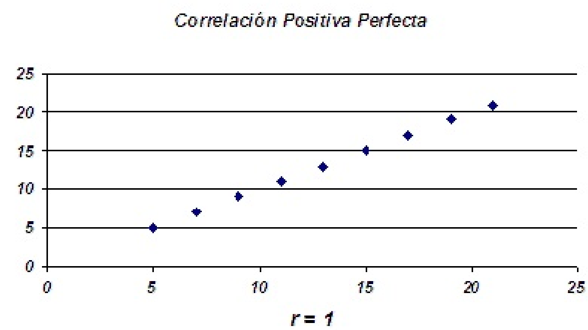


Fig. 5.5. Correlación Positiva Perfecta

Correlación inversa o negativa: Se obtiene cuando al aumentar (disminuir) el valor de la variable independiente, disminuye (aumenta) el valor de la variable dependiente. Si la correlación toma el valor de -1 se tiene correlación negativa perfecta.

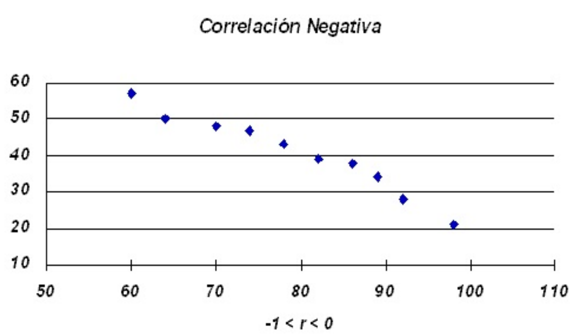


Fig. 5.5. Correlación Negativa

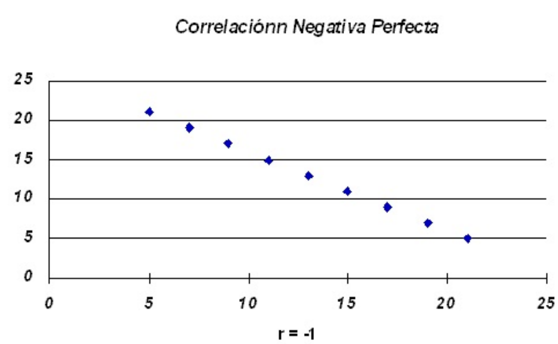


Fig. 5.6. Correlación Negativa Perfecta

<u>Correlación nula</u>: Se da cuando no existe relación lineal entre las variables.

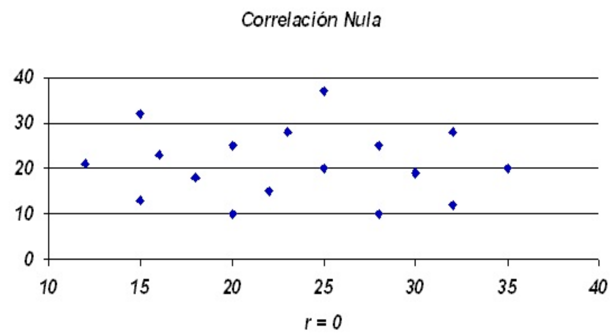


Fig. 5.7. Correlación Nula

Ejemplo 5.2

Emplear el método de mínimos cuadrados para ajustar los siguientes puntos a una recta.

x	1	2	3	4	5	6
y	1	2	2	3	5	5

a) ¿Cuáles son la estimaciones de β_0 y β_1 de mínimos cuadrados?

b) Obtener el coeficiente de correlación e interpretarlo.

Resolución

a) $\sum_{i=1}^{n} x_i = 21 \quad \sum_{i=1}^{n} y_i = 18 \quad n = 6$

$$\sum_{i=1}^{n} x_i^2 = 91 \quad \sum_{i=1}^{n} y_i^2 = 68$$

$$\sum_{i=1}^{n} x_i y_i = 78$$

$$SS_{xx} = 91 - \frac{(21)^2}{6} = 17.5$$

$$SS_{yy} = 68 - \frac{(18)^2}{6} = 14$$

$$SS_{xy} = 78 - \frac{(21)(18)}{6} = 15$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6}{7} \approx 0.8571$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0$$

La ecuación de la recta de mínimos cuadrados es:

$y = 0.8571\ x$

b) El coeficiente de correlación es:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{15}{\sqrt{(17.5)\ (14)}} = 0.9583$$

Las variables x y y tienen una buena asociación lineal.

Ejemplo 5.3

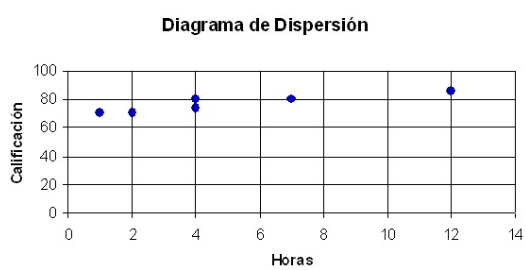
Los siguientes datos representan el número de horas de estudio (x) y la calificación obtenida (y) en un examen para una muestra de 6 estudiantes.

Estudiante	A	B	C	D	E	F
Horas	1	2	4	4	7	12
Calificación	71	71	74	80	80	86

a) Representar los datos en un diagrama de dispersión.
b) Ajustar a los datos un modelo lineal de regresión empleando el criterio de mínimos cuadrados.
c) Si estudia 5 horas, ¿cuál calificación esperaría?
d) Calcular la covariancia y el coeficiente de determinación. Interpretar los resultados de la relación de las variables.

Resolución

a)



b)

	x	y	x^2	xy	y^2
	1	71	1	71	5041
	2	71	4	142	5041
	4	74	16	296	5476
	4	80	16	320	6400
	7	80	49	560	6400
	12	86	144	1032	7396
Sumas	30	462	230	2421	35754

De donde:

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 230 - \frac{(30)^2}{6} = 80$$

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 2421 - \frac{(30)(462)}{6} = 111$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{111}{80} = 1.3875$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{462}{6} - 1.3875\left(\frac{30}{6}\right)$$

$$= 70.0625$$

$$\therefore \hat{y} = 70.0625 + 1.3875x$$

c) Utilizando la recta de regresión

$$\hat{y} = 70.0625 + 1.3875(5)$$

d) $\text{Cov} = \dfrac{SS_{xy}}{n} = \dfrac{111}{6} = 18.5$

$$SS_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= 35754 - \frac{(462)^2}{6} = 180$$

Por lo que:

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \frac{111^2}{(80)(180)} = 0.8556$$

La variable x explica el 85.56% del comportamiento de y.

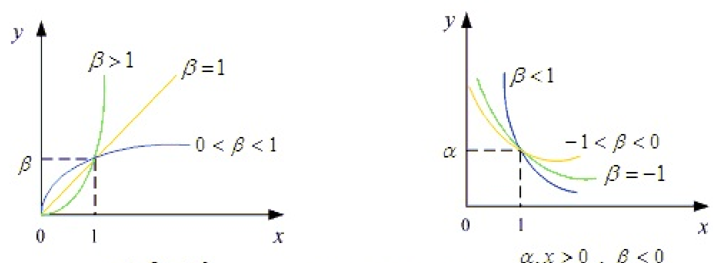
MODELOS LINEALIZABLES

En algunas ocasiones, se puede descubrir que la relación entre x y y no está dada por una recta, ya sea por diagramas o analizando el coeficiente de determinación; sin embargo, es posible que la relación no lineal existente entre x y y pueda ser linealizada, a estos modelos no lineales se les llama ***transformablemente lineales***. Las funciones no lineales, sus gráficas, las transformaciones y las formas lineales que resultan se resumen en la tabla 5.1

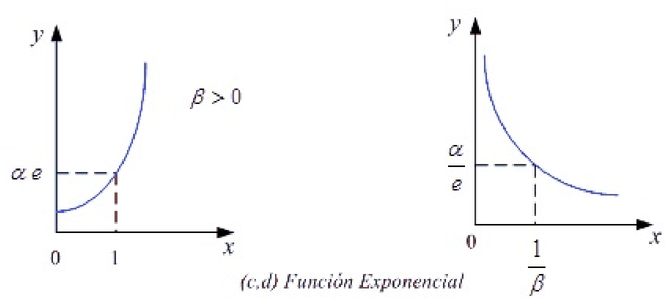
.

Tabla 5.1. Modelos transformablemente lineales

Figura	Función linealizable	Transformación	
Potencia (a, b)	$y = \alpha x^{\beta}$	$y' = \ln y, x' = \ln x$	$\alpha = e^{\beta_0}$ $\beta = \beta_1$
Exponencial (c, d)	$y = \alpha e^{\beta x}$	$y' = \ln y$	$\alpha = e^{\beta_0}$ $\beta = \beta_1$
Logarítmica (e, f)	$y = \alpha + \beta \ln x$	$x' = \ln x$	$\alpha = \beta_0$ $\beta = \beta_1$
Hiperbólica (g, h)	$y = \dfrac{x}{\alpha x - \beta}$	$y' = \dfrac{1}{y}, x' = \dfrac{1}{x}$	$\alpha = \beta_0$ $\beta = -\beta_1$
Recíproca (i, j)	$y = \alpha + \beta\left(\dfrac{1}{x}\right)$	$x' = \dfrac{1}{x}$	$\alpha = \beta_0$ $\beta = \beta_1$



(a,b) Función Potencia



(c,d) Función Exponencial

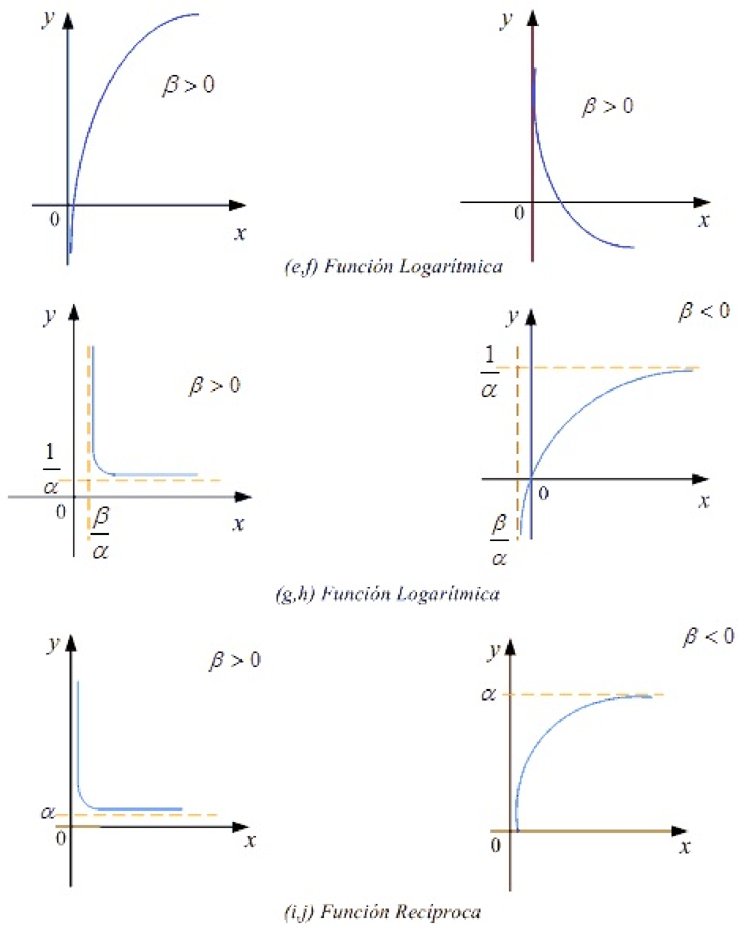


Fig. 5.2. Gráficas de modelos transformablemente lineales

Es decir, si la función de los datos es de tipo potencia, a los datos (x,y) originales, se les hará la transformación indicada y se hará la regresión lineal con esas variables transformadas, en este caso con (x',y'), para el caso de tener una función tipo exponencial la regresión lineal se hará con la x de los datos originales y la y transformada (y').

Cuando se emplean estas transformaciones se debe tener cuidado sobre la forma del modelo antes y después de la transformación, es decir, una vez que se tenga el modelo lineal, se debe regresar al modelo que linealizamos obteniendo sus parámetros α y β, para poder utilizarlo cuando se quiera conocer un valor de y dado uno de x, también deben tenerse en cuenta la medida de mejoría R^2.

Ejemplo 5.4

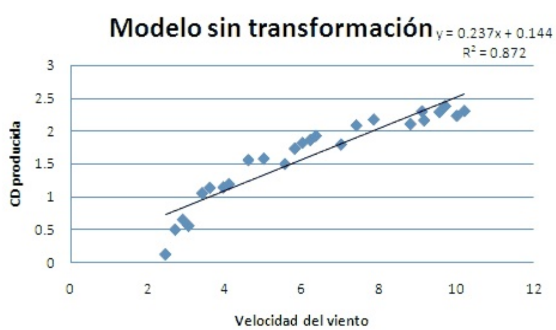
Un ingeniero investiga el uso de un molino de viento para generar electricidad. Ha reunido datos sobre la corriente directa (CD) producida por su molino y la velocidad correspondiente. Los datos se resumen en la siguiente tabla:

x (CD)	y (velocidad)	x (CD)	y (velocidad)
5	1.582	5.8	1.737
6	1.822	7.4	2.088
3.4	1.057	3.6	1.137
2.7	0.5	7.85	2.179
10	2.236	8.8	2.112
9.7	2.386	7	1.8
9.55	2.294	5.545	1.501
3.05	0.558	9.1	2.303
9.15	2.166	10.2	2.31
6.2	1.866	4.1	1.194
2.9	0.653	3.95	1.144
6.35	1.93	2.45	0.123
4.6	1.562		

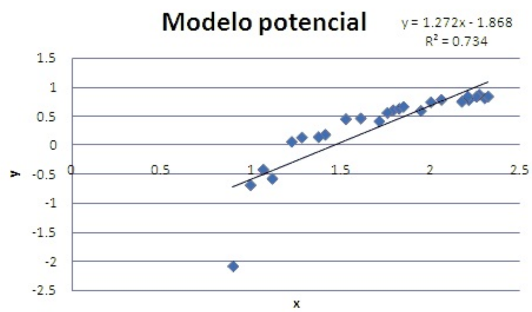
Determinar un modelo lineal adecuado para relacionar a x y a y.

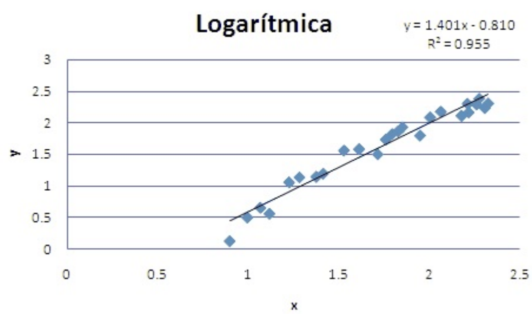
Resolución

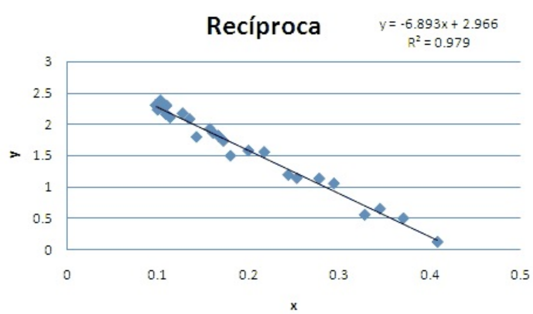
Se realiza la regresión, sin aplicar ninguna transformación:



Puede observarse el valor de r^2=0.872 , (R^2 en los programas estadísticos) y en el diagrama de dispersión podemos identificar que puede parecerse a la función potencial, logarítmica y recíproca, por lo tanto, se realizan las transformaciones correspondientes para cada una y se realizan las regresiones.







Al llevar a cabo las transformaciones, se puede observar que con la transformación a la función recíproca r^2 mejoró considerablemente de 0.872 a 0.979, por lo cual para modelar nuestro problema debemos obtener la función recíproca, obteniendo sus parámetros, de acuerdo a la tabla 5.1, éstos son:

$\alpha = \beta_0$

$\beta = \beta_1$

Partiendo del modelo lineal al cual llegamos mediante la transformación:

$\hat{y} = -6.893x + 2.996$

$\alpha = 2.996$

$\beta = -6.893$

Por lo tanto la función recíproca para nuestros datos queda de la forma:

$$y = \alpha + \beta\left(\frac{1}{x}\right)$$

Finalmente:

$$y = 2.996 - 6.893\left(\frac{1}{x}\right)$$

Si se quisiera estimar el valor de y para $x = 3.8$, entonces deberá obtenerse mediante

$$y = 2.996 - 6.893\left(\frac{1}{3.8}\right) = 1.1820$$

ESTIMACIÓN DE INTERVALOS PARA LOS COEFICIENTES DE REGRESIÓN Y BANDA DE CONFIANZA

El método de mínimos cuadrados nos proporcionó las estimaciones puntuales de β_0 y β_1; sin embargo, es posible obtener estimaciones de intervalo para estos parámetros. Considerando que los errores aleatorios ε_i son independientes y siguen una distribución normal con media cero y varianza σ^2, entonces el estimador insesgado de la varianza de los errores es:

$$\hat{\sigma}^2 = \frac{SEC}{n-2}$$

donde

$$SEC = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 SS_{xy}$$

Para obtener $\hat{\sigma}^2$ se divide entre $n-2$ porque se pierden dos grados de libertad al tener que estimar la ordenada al origen y la pendiente β_0 y β_1 antes de obtener la estimación $\hat{y}_i$.

El intervalo de confianza para la pendiente del modelo de regresión lineal está dado por:

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},(n-2)} \sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},(n-2)} \sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}}$$

y de forma similar el intervalo de confianza para la ordenada al origen está dado por:

$$\hat{\beta}_0 - t_{\frac{\alpha}{2},(n-2)} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},(n-2)} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}$$

También puede construirse un intervalo de confianza para la variable dependiente a partir de un valor específico de la variable independiente, es decir, construir un intervalo de confianza para

$$E(y \mid x_0)$$

a partir de la estimación

$$\widehat{E(y \mid x_0)} \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

obteniendose el intervalo

$$\hat{y}_0 - t_{\frac{\alpha}{2},(n-2)} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)} \leq E(y \mid x_0) \leq \hat{y}_0 + t_{\frac{\alpha}{2},(n-2)} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)}$$

y como puede observarse, la longitud del intervalo depende del valor de x_0, por lo que al obtener diversos intervalos y localizarlos junto a la recta re regresión se obtiene una banda de confianza. La longitud del intervalo es mínima cuando $x_0 = \bar{x}$ y se incrementa a medida que el valor x_0 se aleja de $\bar{x}$, lo cual nos indica que la recta de regresión no debe utilizarse para extrapolar, puesto que el error aumenta significativamente.

PRUEBAS DE HIPÓTESIS PARA LOS COEFICIENTES DE REGRESIÓN

La prueba de hipótesis más importante que se puede hacer sobre la regresión es la validación de la pendiente, esto es, probar que la pendiente debe estar incluida en el modelo y por lo tanto $\beta_1 \neq 0$.

Recordando quie $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados de β_0 y β_1, es decir, $E(\hat{\beta}_0) = \beta_0$ y $E(\hat{\beta}_1) = \beta_1$. Por otro lado,

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]$$

y

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_{xx}}$$

donde σ^2 es la variancia del error aleatorio $\boldsymbol{\varepsilon}$

La prueba de la utilidad del modelo se da en la tabla 5.2

Tabla 5.2. Prueba de la utilidad del modelo lineal

Pruebas unilaterales		Prueba bilateral
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$	$H_1: \beta_1 \neq 0$
Estadístico de prueba: $T_0 = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{SS_{xx}}}}$		
Región de rechazo:		
$t_0 < -t_{\alpha, n-2}$	$t_0 > t_{\alpha, n-2}$	$t_0 > t_{\frac{\alpha}{2}, n-2}$

donde $\hat{\sigma}^2$ es la varianza del error, es decir, $\hat{\sigma}^2 = \dfrac{SEC}{n-2}$.

El no rechazar la hipótesis nula $H_0: \beta_1 = 0$, equivale a comprobar que no existe relación lineal entre las variables x y y.

También puede realizarse una prueba de hipótesis sobre el valor de la ordenada al origen de la forma

$$H_0: \beta_0 = \beta_{0_0}$$
$$H_1: \beta_0 \neq \beta_{0_0}$$

donde el estadístico de prueba es

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0_0}}{\sqrt{\hat{\sigma}^2 \left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{SS_{xx}} \right]}}$$

y la hipótesis nula se rechaza si $|t_0| > t_{\frac{\alpha}{2}, (n-2)}$.

Ejemplo 5.5

Utilizando los datos del ejemplo 5.2, determinar si existe evidencia suficiente para decir que la pendiente de la recta difiere significativamente de cero. Utilizar $\alpha = 0.05$.

x	1	2	3	4	5	6
y	1	2	2	3	5	5

Resolución

Del ejercicio 5.2

$$\sum_{i=1}^{n} x_i = 21, \quad \sum_{i=1}^{n} y_i = 18, \quad n = 6, \quad \sum_{i=1}^{n} x_i^{\,2} = 91, \quad \sum_{i=1}^{n} y_i^{\,2} = 68, \quad \sum_{i=1}^{n} x_i y_i = 78$$

$$SS_{xx} = 91 - \frac{(21)^2}{6} = 17.5$$

$$SS_{yy} = 68 - \frac{(18)^2}{6} = 14$$

$$SS_{xy} = 78 - \frac{(21)(18)}{6} = 15$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{6}{7} \approx 0.8571$$

La prueba de hipótesis requerida es:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

El estadístico de prueba es: $$T_0 = \frac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{SS_{xx}}}}$$

asumiendo que los errores son independientes y tienen distribución normal.

Donde $$\hat{\sigma}^2 = \frac{SEC}{n-2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}$$

$$\hat{\sigma}^2 = \frac{14 - \dfrac{6}{7}(15)}{6-2} = 0.2857$$

Sustituyendo

$$t_0 = \frac{\dfrac{6}{7}}{\sqrt{\dfrac{0.2857}{17.5}}} = 6.71$$

De tablas $t_{\frac{0.05}{2},(4)} = 2.776$

La región de rechazo es $|\ t_0\ | > t_{\frac{0.05}{2},(4)}$

y puesto que:

$$t_0 = 6.71 > 2.776 = t_{\frac{0.05}{2},(4)}$$

Conclusión: Se rechaza H_0 con $\alpha = 0.05$. Existe suficiente evidencia para concluir que la pendiente es significativamente diferente de cero.

BIBLIOGRAFÍA

Hines, William W. y Montgomery, Douglas C. - Probabilidad y Estadística para ingeniería, cuarta edición.- CECSA.- México, 2005.

Milton, Susan J. Y Arnold, Jesse C.- Probabilidad y Estadística para con aplicaciones para ingeniería y ciencias computacionales, cuarta edición.- McGraw-Hill.- México, 2004.

Devore, Jay L.- Probabilidad y Estadística para ingeniería y ciencias, séptima edición.- Cengage Learning.- México, 2008.

Mendenhall, William III. et al.- Introducción a la Probabilidad y Estadística.- Décimo cuarta edición.- Cengage Learning.- México 2015.

Wackerly Dennis D.- Mendenhall, William, *et al*.- Estadística Matemática con Aplicaciones, sexta edición.- Editorial Thomson.- México, 2002.

Walpole, Ronald E., *et al.*- Probability and Statistics for Engineers and Scientists.- Pearson.- USA, 2007.

Montgomery, Douglas C. y Runger, George C.-Probabilidad y Estadística aplicadas a la Ingeniería, segunda edición.- Limusa-Wiley.- México, 2002.

Scheaffer, Richard L. y McClave, James T.- Probabilidad y Estadística para Ingeniería.- Grupo Editorial Iberoamérica.- México, 1993.

Canavos, George C.- Probabilidad y Estadística Aplicaciones y Métodos.- McGraw-Hill.- México, 1988.

Meyer, Paul L.- Probabilidad y Aplicaciones Estadísticas.- Addison Wesley Iberoamericana.- México, 1992.

Spiegel, Murray R. et al.- Probabilidad y Estadística, cuarta edición.- Mc Graw-Hill.-México 2013.

Borras García, Hugo E., *et al*.- Apuntes de Probabilidad y Estadística.-Facultad de Ingeniería.- México, 1985.

Rosenkrantz, Walter A.- Introduction to Probability and Statistics for Scientists and Engineers.- McGraw-Hill.- EE.UU., 1997.

Ziemer, Rodger E.- Elements of Engineering Probability & Statistics.- Prentice Hall.- USA 1997.