A los asistentes a los cursos del Centro de Educación Continua

La Facultad de Ingeniería, por conducto del Centro de Educación Continua, otorga constancia de asistencia a quienes cumplan con los requisitos establecidos para cada curso. Las personas que deseen que aparezca su título profesional precediendo a su nombre en el diploma, deberán entregar copia del mismo o de su cédula profesional a más tardar 8 días antes de la terminación del curso, en las oficinas del Centro, con la Srita. Delgadillo.

El control de asistencia se efectuará al terminar la primera hora de cada día de clase, mediante listas especiales en las que los interesados anotarán personalmente su asistencia. Las ausencias serán computadas por las autoridades del Centro.

Se recomienda a los asistentes participar activamente con sus ideas y experiencias, pues los cursos que ofrece el Centro están planeados para que los profesores expongan una tesis, pero sobre todo para que coordinen las opiniones de todos los interesados constituyendo verdaderos seminarios.

Al finalizar el curso se hará una evaluación del mismo a través de un cuestionario diseñado para emitir juicios anónimos por parte de los asistentes.

Las personas comisionadas por alguna institución deberán pasar a inscribirse en las oficinas del Centro en la misma forma que los demás asistentes.

Ing. Melesio Gutiérrez P.

DIRECTORIO DE ASISTENTES AL CURSO DE PROBABILIDAD Y ESTADISTICA FUNDAMENTOS Y APLICACIONES (DEL 29 DE AGOSTO AL 6 DE OCTUBRE DE 1972)

NOMBRE Y DIRECCION	EMPRESA Y DIRECCION
1. ING. ALEJANDRO BASURTO O. Av. Universidad No. 329 Edif.53-403 México, D. F.	PEMEX. Marina Nacional No. 329 México, D. F.
2. ING. EDUARDO BEAVEN SALAS Bolivar No. 927 México 13, D. F.	SECRETARIA DE MARINA Azueta No. 9 7o. Piso México, D. F.
3. ING. GERARDO CABADA LAGUNES Dr. Barragán No. 745-203 México, D. F.	SECRETARIA DE OBRAS PUBLICAS Av. Fernando y Niño Perdido México, D. F.
4. SR. J. AGUSTIN DE OVANDO PACHECO Bosque de Jazmines No. 145 Col. Bosques de las Lomas México, D. F.	PEMEX. Marina Nacional No. 329 México, D. F.
5. ING. AGUSTIN EGUIARTE GONZALEZ Emma No. 136 Col. Nativitas México 13, D. F.	SECRETARIA DE OBRAS PUBLICAS Fernando No. 268 Col. Alamos México 13, D. F.
6. ING. ARMANDO ESCANERO MUÑOZ Av. Universidad No.62-B México, D. F.	SECRETARIA DE OBRAS PUBLICAS Av. Fernando No. 268 Col. Alamos México 13, D. F.
7. SR. J. IGNACIO GUERRERO SANDERS Av. José F. Gutiérrez No. 347 México 16, D. F.	BANCO NACIONAL AGROPECUARIO, S. A. Baja California No. 169 México, D. F.
8. ING. FELIPE LOO GOMEZ Edif. Allende No. 716-D Unidad Nonoalco Tlatelolco México 3, D. F.	SECRETARIA DE OBRAS PUBLICAS DIRECCION GENERAL DE AEROPUERTOS Xola No. 1755 - 3er. Piso Col. Narvarte México 12, D. F.
9. ING. MARIANO MARTINEZ L. G. Cien Fuegos No. 899 México 14, D. F.	COMISION FEDERAL DE ELECTRICIDAD Ródano No. 14 1er. Piso México 5, D. F.

DIRECTORIO DE ASISTENTES AL CURSO DE PROBABILIDAD Y ESTADISTICA FUNDAMENTOS Y APLICACIONES (DEL 29 DE AGOSTO AL 6 DE OCTUBRE DE 1972)

NOMBRE Y DIRECCION	EMPRESA Y DIRECCION
10. SR. J. BENJAMIN MORALES OROZCO Coruña No. 206-11 Col. Viaducto Piedad México 13, D. F.	BODEGAS RURALES CONASUPO, S.A. DE C. V. Donceles No. 89-4o. Piso México, D. F.
11. ING. EMILIO RODRIGUEZ ARNAIZ Cerro del Nombre No. 56-5 Col. Romero de Terreros México 21, D. F.	CONTROL DATA DE MEXICO Mariano Escobedo No. 56-5 Col. Polanco México 5, D. F.
12. SR. JESUS SILVA AVILA Playa Copacabana No. 153 Col. Marte México 13, D. F.	ESTUDIANTE (ESIME) Instituto Politecnico Nacional México, D. F.
13. SR. J. LUIS VAZQUEZ SANTAMARIA Retorno Fuente de Petroleos No. 3 México, D. F.	PETROLEOS MEXICANOS Marina Nacional No. 329 México, D. F.

PROBABILIDAD Y ESTADISTICA. FUNDAMENTOS Y APLICACIONES

Fecha	Duración	Tema	
29 Agosto	2 1/2 Horas	Elementos de Probabilidad	Dr. Felipe Ochoa Rosso
		Experimento aleatorio	
		Eventos. Conjunto de eventos	
		Conceptos de probabilidad	
		Enfoque clásico	
		Enfoque Bayesiano	
		Algunos teoremas fundamentales con enfoque de conjuntos	
		Probabilidad condicional	
		El teorema de Bayes	
31 Agosto	2 1/2 Horas	Eventos independientes	Dr. Felipe Ochoa Rosso
		Eventos mutuamente exclusivos	
		Ley de adición	
		El concepto de probabilidad condicional	
		El teorema de Bayes	
		Concepto de variable aleatoria	
5 Septiembre	2 1/2 Horas	Modelos Probabilisticos	M. en C. Alfonso Cobarrubias Lugo
		Tratamiento de la información	
		Histogramas	
		Definición de términos	
		Distribución de frecuencia de una variable	
		Contínua	
		Discreta	
		Ejemplos	
		Parámetros	
		Medidas de tendencia central	
		Medidas de dispersión	
		Fórmulas para cálculo numérico	
		Ejemplos	

7 Septiembre 2 1/2 Horas Distribuciones de Probabilidad M. en C. Alfonso Cobarrubias Lugo

Definición de probabilidad
Probabilidad condicional, teorema de Bayes
Eventos Independientes
Variables Aleatorias de 1 dimensión

- Definición
- Variables aleatorias discretas
- Distribución Binominal
- Variables aleatorias contínuas
- Distribuciones acumuladas
- Distribuciones mixtas
- Variable aleatoria uniformemente distribuida

12 Septiembre 2 1/2 Horas Funciones de Variables Aleatorias M. en C. Alfonso Cobarrubias Lugo

Variable aleatoria de 2 o más dimensiones
- Distribuciones de probabilidad condicional y marginales
- Variables aleatorias independientes
- Distribución del producto y del cociente de variables
- Aleatorias independientes
- Variables aleatorias de n dimensiones

Caracterizaciones de variables aleatorias
- Valor esperado
- Esperanza de una función de variables aleatorias
- Propiedades del valor esperado
- Variancia y sus propiedades
- Correlación
- Esperanza condicional

14 Septiembre 2 1/2 Horas Distribuciones Teóricas de Probabilidad M. en C. Fernando Gómez Jardón

Distribución Binominal o de Bernoulli
- Probabilidad de éxito
- Probabilidad de fracaso
- Ejemplos

		Distribución de Poisson Aproximación de la distribución binominal a la de Poisson Aplicaciones Distribución Normal Fórmula y gráfica Estandarización interpretaciones prácticas y matemática Ejemplos Distribución Exponencial Aplicación en teoría de colas	
19 Septiembre	2 1/2 Horas	Estimaciones de los Parámetros del Modelos El método de los momentos Propiedades de los estimadores Distribuciones de los estimadores El método de máxima verosimilitud Ejemplos	M. en C. Fernando Gómez Jardón
21 Septiembre	2 1/2 Horas	Pruebas e Hipotesis Hipótesis estadísticas Errores Tipo I y II, nivel de significancia Pruebas sobre la media de la población Pruebas sobre la variancia de la población	M. en C. Fernando Gómez Jardón
26 Septiembre	2 1/2 Horas	Elementos de Estadística Bayesiana Modelo Decisional Valor esperado de las decisiones Teoría de la Utilidad. Ejemplos.	M. en C. Romárico Arroyo Marroquír
28 Septiembre	2 1/2 Horas	Ejemplo	M.en C. Romárico Arroyo Marroquín
3 Octubre	2 1/2 Horas	El método de mínimos cuadrados Inferencia basada en estimadores de mínimos cuadrados	M. en I.Salomán Camhaji Samra

5 Octubre	2 1/2 Horas	Regresión curvilínea. Ejemplos Regresión múltiple Correlación Ejemplos con el teletipo	M. en I. Salomón Camhaji Samra

ELEMENTOS DE PROBABILIDAD

Dr. F. Ochoa

IPESA Consultores

1. INTRODUCCION

"TOMA DE DECISIONES BAJO INCERTIDUMBRE".

El que toma decisiones tiene que seleccionar un determinado curso de acción entre varios disponibles, de manera que el resultado de su decisión sea óptimo conforme a una determinada medida de efectividad asociada al resultado de su decisión.

En la mayoría de estas situaciones, la decisión involucra juicios sobre variables para las cuales existe incertidumbre respecto a los valores que deberán de tomar.

Ante este marco, la Moderna Teoría Estadística de la Decisión nos proporciona una herramienta para buscar los mejores cursos de acción tomando en cuenta la incertidumbre del problema.

La base matemática de dicha herramienta del cálculo la constituye la llamada "Teoría de Probabilidad".

Esta es una rama del Análisis Matemático que permite manejar el concepto de incertidumbre.

Incertidumbre
- debida a variación natural
- falta de entendimiento de las causas y efectos de sistemas
- debida a conocimiento profesional incompleto (falta de datos)

Si un problema se va a resolver mediante un modelo abstracto que tome en cuenta esta incertidumbre, dicho modelo se dice *probabilístico* y está sujeto a análisis mediante la Teoría de Probabilidad.

El concepto central en probabilidad es la Variable aleatoria. (Valor específico no puede determinarse antes de un experimento).

El que toma decisiones no podrá predecir enteramente el futuro.

Considerará la posibilidad de que ocurran determinados *eventos* y calculará la mayor o menor plausibilidad de su ocurrencia.

2. EXPERIMENTO ALEATORIO

Son experimentos cuyos resultados se conocen con incertidumbre.
Ejemplo: No. de artículos de una cierta clase que serán demandados el mes próximo.
Pueden ser 1, 2, 3.......N (No. población potencial)

3. ESPACIO DEL EXPERIMENTO (SAMPLE SPACE).

Es el conjunto de todos los resultados de un experimento.

Los elementos del conjunto se llaman (SAMPLE POINTS) PUNTOS DE LA MUESTRA, cada uno asociado con un solo resultado distinguible.

4. EVENTO.

A es un conjunto de puntos del espacio del experimento.

{ Evento simple.- Un solo punto del espacio
{ Evento compuesto.- Dos o más puntos del espacio.

5. ALGEBRA DE LOS EVENTOS.

Relaciones entre eventos:

a. Si dos eventos no contienen puntos comunes se dicen *disjuntos* o mutuamente exclusivos. $A \cap B = \phi$

b. Eventos simples son mutuamente exclusivos.

c. Si se tienen dos eventos A y B, por $A \cap B$ se indica el nuevo evento que contiene los puntos comunes a A y B.
El evento se llama intersección.

d. La unión de dos eventos A y B, se denota $A \cup C$ y constituye el evento tal que es el conjunto de puntos que aparecen en A o en C.

e. Evento nulo Φ no contiene puntos del Universo. (Evento no puede ocurrir).

f. Universo.- Espacio de muestras (Conjunto todos eventos pueden ocurrir), I.

g. Conjuntos colectivamente Exhaustivos. Si y solo si $A_1 + A_2 + \ldots + A_N = I$

h. El complemento A^c de un evento A (la no ocurrencia de A) consiste en un evento que contiene todos los puntos del espacio no contenidos en A.

6. AXIOMAS.

Enunciados en los que se base el algebra de eventos.

(1) $A + B = B + A$

(2) $(A^c)^c = A$

(3) $A + (B + C) = (A + B) + C$

(4) $AB = (A^c + B^c)^c$ ó $(AB)^c = A^c + B^c$

(5) $A + (BC) = (A + B)(A + C)$

(6) $A A^{C} = \phi$

(7) $A I = A$

(14) $A + \phi = A$

7. RELACIONES ENTRE EVENTOS

$AB = BA$ (8)

$A^{C} B^{C} = (A + B)^{C}$ (9)

$A(B + C) = AB + AC$ (10)

$A + A^{1} = I$ (11)

$A + I = I$ (12)

$A \phi = \phi$ (13)

$A + \phi = A$ (14)

EJEMPLO.

Demostrar que $A + B + C = A + A^{C}B + (A + A^{C}B)^{C} C$

La unión de 3 eventos como la unión de 3 eventos mutuamente exclusivos.

Del axioma (7)

$$A + B = A + B\ I$$

$$= A + B\ (A + A^c) \qquad \text{de (11)}$$

$$= A + BA + BA^c \qquad \text{de (10)}$$

$$= AI + BA + BA^c \qquad \text{de (7)}$$

$$= A\ (I + B) + BA^c \qquad \text{de (8) y (10)}$$

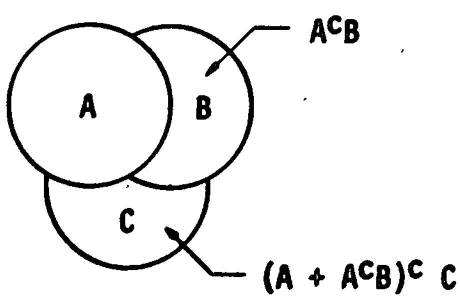
$$= AI = BA^c$$

$$= A + A^cB$$

Sea $A + B = E \rightarrow E + C = E + E^cC$

$$(A + B) + C = (A + A^c\ B) + (A + A^c\ B)^c\ C$$

$$A + B + C = A + A^c\ B + (A + A^c\ B)^c\ C$$



$\Theta \equiv D$

8. INTERPRETACION DE PROBABILIDADES.

. A cada punto del espacio de definición de un experimento le vamos a asignar una medida de probabilidad.

La teoría de probabilidad no nos dice cómo se obtienen, simo cómo usarlos de manera consistente.

ENFOQUE CLASICO

. Explicación Intuitiva de la Medida de la Probabilidad.

Si se pudiera realizar el experimento una cantidad considerable M de veces y se cuenta el número de veces N que un evento simple se observó, el número $\frac{N}{M} = p$, lo utilizaríamos como la medida de la probabilidad de que el evento simple ocurra.

Esto es adecuado cuando el fenómeno es repetitivo.

Esta interpretación de frecuencias relativas se debe a Von Mises.

ENFOQUE BAYESIANO

Cuándo la repetición del fenómeno es imposible o carece de significado? Implica una interpretación más liberal del concepto de probabilidad de un evento.

Las probabilidades asignadas a cada evento simple de un experimento pueden considerarse como un conjunto de "pesos" (factores de ponderación), que expresan la medida subjetiva de la plausibilidad subjetiva de que un evento simple ocurra.

"Medida subjetiva del grado de aceptación de una persona sobre un juicio o predicción".

9. AXIOMAS SOBRE PROBABILIDAD.

Notación:

$P[A] \rightarrow$ probabilidad de un evento aleatorio A.

Para el espacio de definición del experimento debe cumplirse:

Axioma No. 1 $0 \leq P[A] \leq 1$

Axioma No. 2 $P[I] = 1$

La probabilidad de un evento cierto es la unidad.

Axioma No. 3 La probabilidad de un evento que es la unión de dos eventos mutuamente exclusivos es la suma de las probabilidades de estos dos eventos:

$$P[A \cup B] = P[A] + P[B]$$

Puesto que I es la unión de todos los eventos simples, el axioma 3 implica que el axioma 2 puede escribirse:

$$\sum_{\forall i} P\ [\ E_i\ ] = 1$$

donde E_i son eventos simples asociados con puntos de prueba individuales.

10. RELACIONES ENTRE PROBABILIDADES DE EVENTOS.

De las relaciones entre eventos, y de los axiomas sobre probabilidad se derivan ciertas relaciones entre las probabilidades de los eventos.

a. Probabilidad de un Evento = Suma de las probabilidades asignadas a cada evento simple que forma parte del evento.

b. Probabilidad de un evento imposible

$$P\ [\ \phi\ ] = 0$$

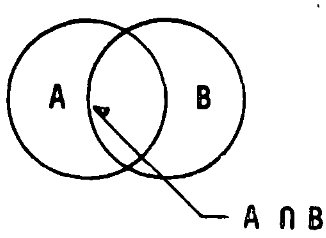
c. Probabilidad del evento unión de dos eventos.- (Mutuamente exclusivos) o <u>no</u>

$$P\ [\ A \cup B\ ] = P\ [\ A\ ] + P\ [\ B\ ] - P\ [\ A \cap B\ ]$$

$A = B^c A + A \cap B$ (mutuamente exclusivos)

$B = A^c B + A \cap B$

$A \cup B = B^c A + A^c B + A \cap B$ (mutuamente exclusivos)



$$P\,[A] = P\,[B^C A] + P\,[A \cap B]$$

$$P\,[B] = P\,[A^C B] + P\,[A \cap B]$$

$$P\,[A \cup B] = P\,[B^C A] + P\,[A^C B] + P\,[A \cap B]$$

$$P\,[A \cup B] = P\,[A] + P\,[B] - P\,[A \cap B]$$

Θ Σ D

11. PROBABILIDAD CONDICIONAL.

La "probabilidad condicional" de un evento A, dado que el evento B ha ocurrido, $P\,[A / B]$, se define como el cociente de la probabilidad de la intersección de A y B entre la probabilidad de B:

$$P\,[A / B] = \frac{P\,[A \cap B]}{P\,[B]}$$

Si $P\,[B] = 0$; $P\,[A / B]$ indefinida.

En las aplicaciones, generalmente se busca $P\,[A \cap B]$.

$$P\,[A \cap B] = P\,[A / B] \,.\, P\,[B] = P\,[B / A]\; P\,[A]$$

12. INDEPENDENCIA

Dos eventos A y B se dicen independientes si y solo si

$$P\ [A / B] = P\ [A]$$

Por tanto: para eventos independientes probabilísticamente:

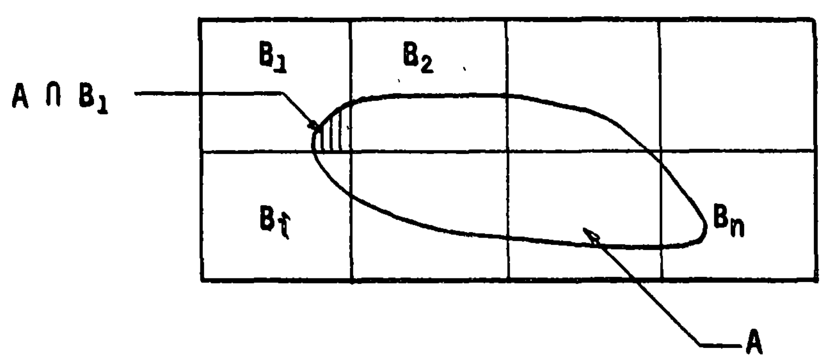
$$P\ [A \cap B] = P\ [A] \cdot P\ [B]$$

13. TEOREMA DE BAYES

Dado un conjunto de eventos mutuamente exclusivos y colectivamente exhaustivos B_1 , B_2 , ..., B_n

$$P\ [A] = P\ [A \cap B_1] + P\ [A \cap B_2] + ... + P\ [A \cap B_n]$$

$$= \sum_{i=1}^{n} P\ [A \cap B_i] = \sum_{i=1}^{n} P\ [A / B_i] .\ P\ [B_i]$$



Al examinar la probabilidad condicional de B_j dado A se tiene:

$$P\left[B_j / A\right] = \frac{P\left[B_j \cap A\right]}{P\left[A\right]} = \frac{P\left[A \cap B_j\right]}{P\left[A\right]}$$

$$\boxed{P\left[B_j / A\right] = \frac{P\left[A / B_j\right] P\left[B_j\right]}{\sum_{i=1}^{n} P\left[A/B_i\right] P\left[B_i\right]}}$$

Incorpora nueva información con probabilidad anterior subjetiva, para obtener nuevos valores de las probabilidades.

SESION 3. TRATAMIENTO DE LA INFORMACION

M EN C. ALFONSO COBARRUBIAS LUGO

3.1. HISTOGRAMAS

3.1.1. Definición de Términos.

Conceptos tales como dato, objeto, individuo, variable, colección y conjunto tienen el mismo significado en el lenguaje común y corriente y en estadística, sin embargo los términos observación y población tienen un significado diferente.

En estadística una observación es un registro, usualmente numérico, de información. Existen varios tipos de registros, estos pueden ser, entre otros: - resultados, medidas, rangos, categorías, etc.

Un resultado es una estimación numérica de objetos, individuos, etc., en una cierta escala. Ejemplos: La calificación obtenida en una prueba, el número de goles anotados por un equipo.

Una medida es, como comunmente se interpreta, un valor numérico que indica el tamaño o alcance de un objeto, v.gr: la altura de un hombre, el peso de un objeto, la distancia entre dos puntos, etc.

El término "rango" se usa para expresar la posición de un objeto con respecto a una calidad especificada para todos los otros objetos bajo consideración, v.gr: la calidad de un producto puede ser muy bueno, regular y malo; etc.

El término categoría se usa para concepto tales como: color de ojos, tipo de metal usado en un cierto producto, tipo de árbol, etc.

Como ejemplos de observaciones que no corresponden a ninguno de los tipos - antes mencionados, se tiene: Número de productos defectuosos en un proceso de fabricación, número de alumnos inscritos en una clase, etc.

Una observación siempre se lleva a cabo en una o más características de un - objeto, v.gr: en un día se pueden tener observaciones para medir la temperatura máxima, en una placa pueden ser para medir su ancho, etc. Así pues llamamos variable a la temperatura máxima observada ya que puede variar día con día y siempre se considerará que el valor de la variable es un número, por lo que los términos observación y valor de la variable son sinónimos siempre y cuando la información registrada sea numérica.

Las variables pueden ser contínuas o discretas. Una variable contínua es - aquella que puede tener un valor real comprendido entre dos números distintos y una variable discreta es aquella que sólo puede tomar valores aislados.

Ejemplos de variables discretas: el número de gentes en un salón, el número de accidentes en un año, el número de países en el continente, el número de ases en una mano de poker, etc.

Ejemplos de variables contínuas: La velocidad de un coche, la velocidad de un avión, la intensidad de un ruido o luz, la vida útil de un foco, etc.

Nota Aclaratoria: Por ejemplo la vida útil de un foco puede considerarse como variable discreta. Es verdad que un conjunto de datos solo tomará valores aislados, por lo que se les puede considerar discretos, pero la vida útil de un foco sólo puede ser medida hasta la precisión del instrumento, digamos décimas de segundo, esto introduce un error de redondeo por lo que el valor medido solo se aproxima a un número que podría caer entre dos números reales distintos.

En estadística no nos interesa una observación sino como se relaciona con - una colección, nos interesa una colección que tenga un atributo común observable, a la colección de tales observaciones se le llama población. A las - poblaciones de valores de variables sencillas se les llama poblaciones univariadas, una población de dos variables se les llama bi-variada, a una de - tres variables tri-variada, en términos generales a la que tenga más de una variable se le llama multivariada.

Ejemplos:

Univariada:	Peso de un hombre
Bi-variada:	Peso y estatura de un hombre
Tri-variada:	Peso, estatura y edad de un hombre.
Multi-variada:	Peso, estatura, edad, presión anterial y colesterol de un hombre.

Una población puede consistir de un número finito o infinito de observaciones. Si es la población es finita y contiene N observaciones, se dice que N es el tamaño de la población.

Si N es suficientemente pequeño podemos comprender la naturaleza de la población mediante la inspección del conjunto de observaciones .

Si N es muy grande o la población es infinita hay necesidad de reducirla - para comprender mejor su naturaleza, esto puede hacerse de dos maneras:

a) Tabular o gráficamente mediante la clasificación de valores de la variable en rangos.

b) Aritméticamente encontrando el valor numérico de ciertas características, llamadas parámetros, que se determinan de las observaciones y que definen a la población, es decir que presentan los aspectos más importantes de la población.

El principal interés de la estadística se refiere a estos parámetros. Como las poblaciones son diferentes, no pueden ser descritas de la misma manera; sin embargo tienen ciertas características o parámetros en común que deben ser estudiados en primer término.

En la mayoría de los casos el investigador no tiene observaciones de todos los objetos de una población, sino que usualmente cuenta con una muestra la cual es una sub-colección de la población de valores. Cuando esto se presenta, lo más que puede hacer es estimar la naturaleza de la población. Esto se logra determinando la naturaleza de la muestra y usando las propiedades de la muestra para estimar las propiedades que caracterizan a la población. El proceso mediante el cual llegamos a conclusiones respecto a una población a partir de una muestra tomada de la población, se conoce como Inferencia Estadística.

3.1.2. Distribuciones de Frecuencia en una Variables.

En este punto trataremos diversos métodos para describir la distribución de observaciones en una población o en una muestra.

3.1.2.1. Variables Continuas.

Si la población o la muestra es pequeña, probablemente la mejor manera de describirla sea listando los valores de la variable. Por ejemplo, el coeficiente de permeabilidad de tres muestras de arena listados en orden ascendente pueden describir la población en forma satisfactoria.

Nota Aclaratoria. Si el objetivo de la investigación es únicamente describir un conjunto de observaciones, entonces no importa si el conjunto de datos se considera como muestra o población. Sin embargo cuando el principal objetivo es utilizar la muestra para hacer una inferencia estadística respecto a la población, los métodos aritméticos para muestras difieren en algunos aspectos de los métodos para poblaciones y, en este caso, la descripción de la muestra se considera parte del procedimiento estadístico.

EJEMPLO: A continuación se presenta la distribución del porcentaje de individuos en cada estado de la Unión Americana de 65 años o más de edad. Fuente: Bureau of the Census, County and City Data Book, 1956

TABLA 1

Estado	%	Estado	%	Estado	%	Estado	%
1	6.5	13	10.4	25	9.8	37	8.9
2	5.9	14	10.2	26	6.9	38	5.4
3	7.8	15	8.0	27	10.8	39	8.5
4	8.5	16	6.6	28	8.1	40	7.1
5	8.7	17	10.2	29	4.9	41	6.7
6	8.8	18	7.0	30	8.5	42	6.2
7	8.3	19	10.0	31	5.5	43	10.5
8	8.6	20	7.2	32	7.8	44	6.5
9	6.4	21	9.0	33	8.9	45	8.9
10	7.4	22	7.0	34	8.7	46	6.9
11	8.7	23	10.3	35	8.7	47	9.0
12	9.2	24	8.6	36	8.4	48	6.3

Generalmente, lo primero que se hace para entender mejor la naturaleza de los datos es arreglarlos en orden numérico ascendente, o listar cada número con la frecuencia con que ocurre. La tabla 2 muestra los datos de la tabla anterior arreglados en orden ascendente.

TABLA 2 % Listados en orden ascendente

4.9	6.4	6.9	7.8	8.5	8.7	8.9	10.2
5.4	6.5	7.0	7.8	8.5	8.7	9.0	10.2
5.5	6.5	7.0	8.0	8.5	8.7	9.0	10.3
5.9	6.6	7.1	8.1	8.6	8.8	9.2	10.4
6.2	6.7	7.2	8.3	8.6	8.9	9.8	10.5
6.3	6.9	7.4	8.4	8.7	8.9	10.0	10.8

La tabla 2 muestra que el porcentaje más alto es de 10.8 y el más bajo de 4.9, con un rango de 10.8 - 4.9 = 5.9.

Todas las muestras pueden ser representadas gráficamente, pero la idea es que estas gráficas sean tan "suaves" como sea posible, esto se puede lograr agrupando los datos en "clases". Lo anterior lleva asociadas las siguientes 2 preguntas: ¿cuántas clases se deben utilizar y de que longitud deben ser las clases?, ninguna de las dos preguntas tienen una respuesta que sea válida en todos los casos, sin embargo pueden mencionar ciertas reglas generales como: La longitud de las clases debe ser igual al menos que exista una razón suficientemente poderosa para hacer lo contrario, cuando se tengan más de 200 observaciones, dos usualmente el número de clases que permite una agrupación adecuada varía de 10 a 20.

El valor numérico que divide a dos clases se llama "límite" o "frontera", la longitud de la clase es la diferencia entre su límite superior y límite inferior.

Al valor medio entre dos límites se le llama "marca" y usualmente se le utiliza para representar a la clase. La frecuencia o número de veces con que las observaciones caen dentro de una clase se llama frecuencia de la clase.

Una vez decidido el número de clases y sus límites se puede construir una tabla de frecuencia como la que se muestra a continuación:

Clases			Frecuencia		
Límites	Marca	Frecuencia (fi)	Relativa fi/N	Acumulada	Acumulada Relativa
4.55-5.55	5.05	3	0.062	3	0.062
5.55-6.55	6.05	6	0.125	9	0.188
6.55-7.55	7.05	9	0.188	18	0.375
7.55-8.55	8.05	9	0.188	27	0.562
8.55-9.55	9.05	13	0.271	40	0.833
9.55-10.55	10.05	7	0.146	47	0.979
10.55-11.55	11.05	1	0.021	48	1.000

Partiendo de una tabla de frecuencias se puede construir un Histograma de Frecuencias como el que se ilustra en la Fig. 1. Para su construcción se utiliza un sistema coordenado rectangular en el cual el eje de las abscisas se utiliza

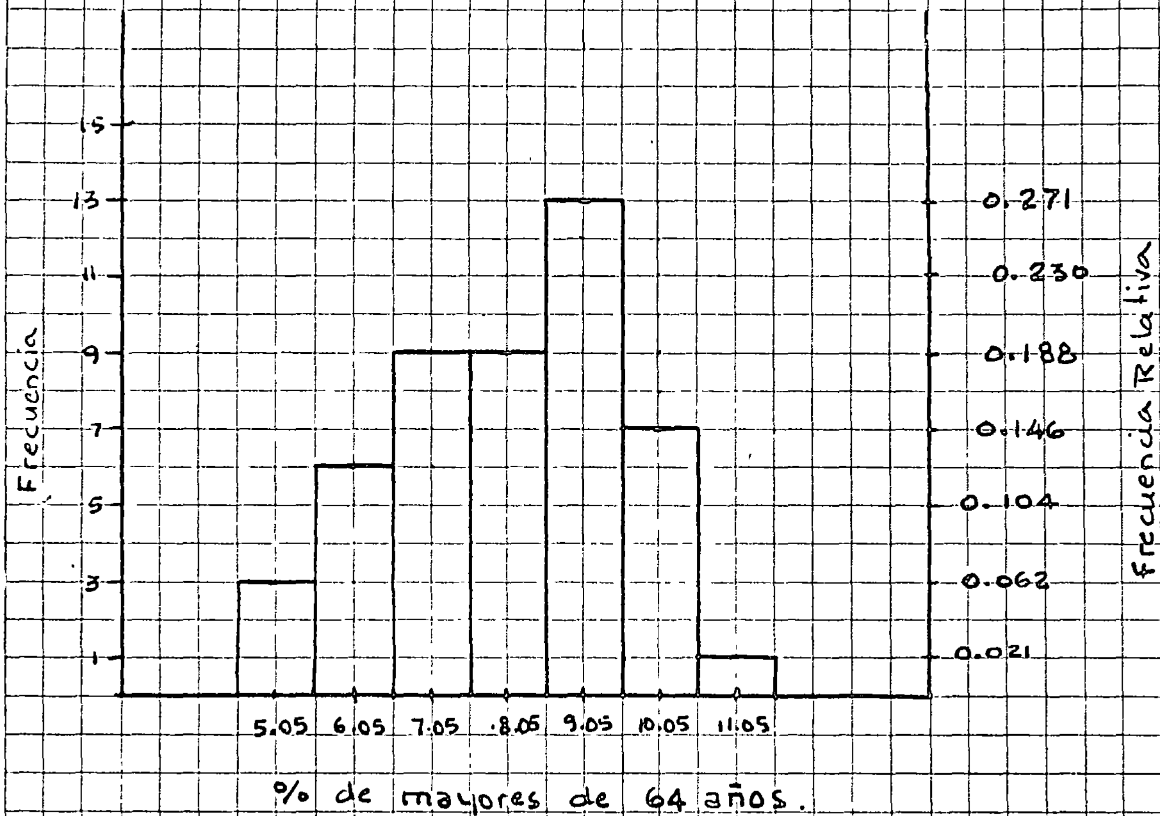


FIG 1 HISTOGRAMA DE FRECUENCIAS

para los límites y marcas de las clases y la ordenada para las frecuencias. Se construye sobre cada clase un rectángulo cuya área sea igual o proporcional a la frecuencia de la clase de tal forma que el área total del histograma o sea la suma de los rectángulos, es igual a la frecuencia total.

Los histogramas se utilizan para datos agrupados, este agrupamiento introduce aproximaciones tales como que todos los valores comprendidos dentro de un intervalo aparecen con la misma frecuencia.

El polígono de frecuencias, como el que se muestra en la Fig. 2 ilustra otra forma de representar diagramas de frecuencia. Un polígono de frecuencia se construye marcando en un sistema de ejes coordenados puntos (X_i, f_i) y uniéndolos con líneas. Para completar la figura normalmente se utiliza un punto de frecuencia nula al principio y al final de la gráfica.

Los polígonos de frecuencia pueden inducir graves errores si no se utilizan adecuadamente, estos errores pueden ser, por ejemplo: la lectura de la frecuencia correspondiente a un punto, etc. Estos polígonos se utilizan principalmente para indicar la forma de un conjunto de datos.

Existe otro tipo de diagrama de frecuencias llamado "Polígono de Frecuencias Acumuladas" como el que se muestra en la figura 3. La diferencia principal entre este tipo de polígono y el de frecuencias consiste en la localización del punto (X_j', f_i') correspondiente a (X_i, f_i). La abscisa X_i' es el límite superior o inferior de la clase dependiendo de si se trata de un diagrama de "más de" ó "menos de".

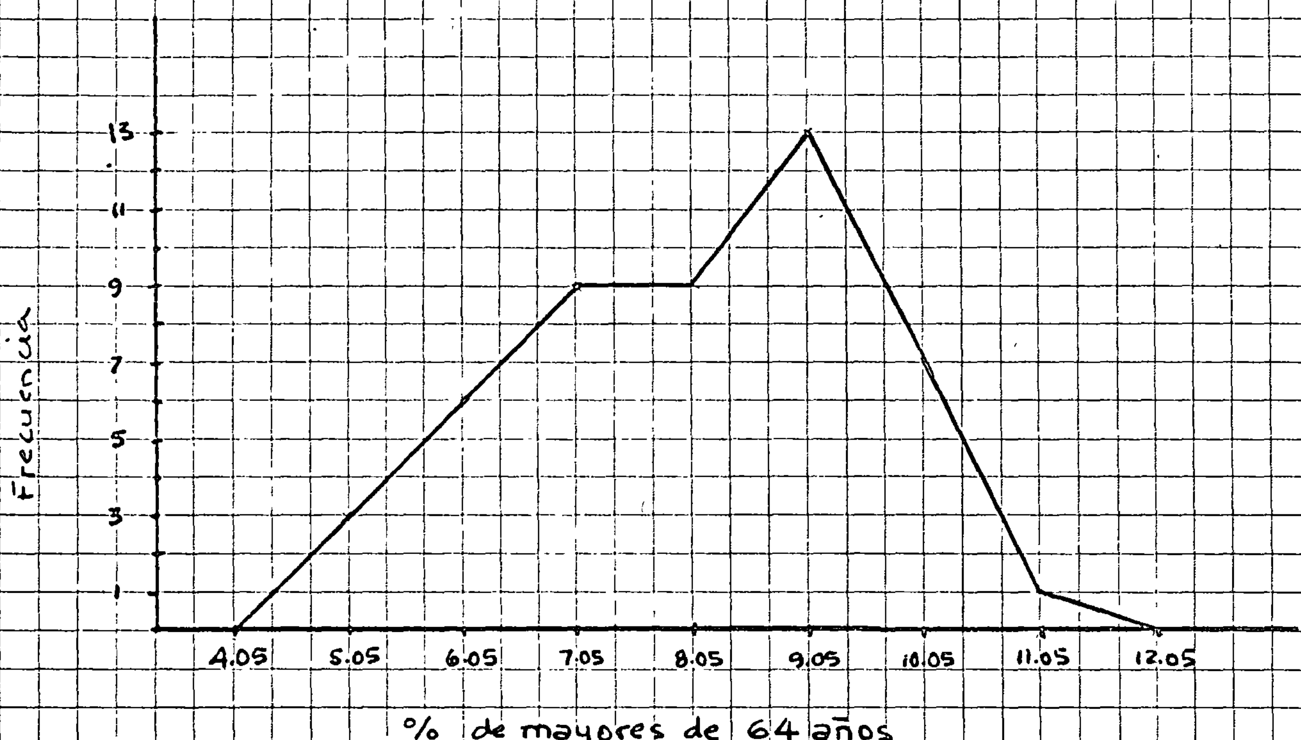


FIG 2. Polígono de Frecuencias

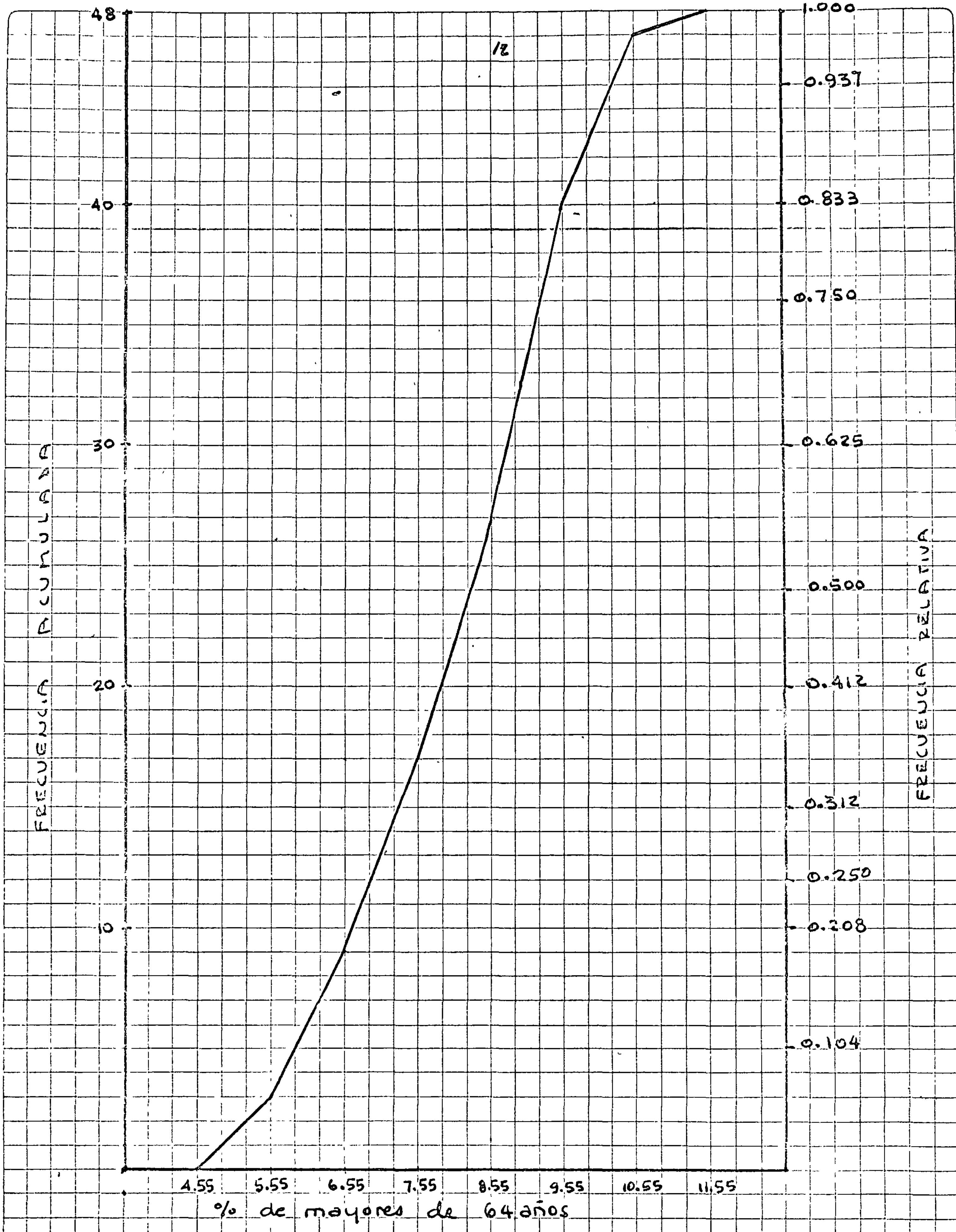


Fig 3 POLIGONO DE FRECUENCIAS ACUMULADAS

3.1.2.2. Variables Discretas.

Cuando se tienen observaciones resultantes de contar coches que pasan por una esquina, alumnos que asisten a una clase, etc. se tienen variables discretas. Para describir variables discretas y al mismo tiempo comparar los métodos con los utilizados para variables contínuas utilizaremos el siguiente ejemplo.

La tabla 2.5 muestra la distribución de frecuencia del f'_c de cilindros de concreto.

Tabla 2.5

f'_c	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada
200	2	0.002	2	0.002
201	0	0.000	2	0.002
202	4	0.005	6	0.007
203	21	0.026	27	0.033
204	47	0.058	74	0.092
205	104	0.129	178	0.220
206	123	0.152	301	0.373
207	148	0.183	449	0.556
208	146	0.181	597	0.739
209	91	0.113	686	0.849
210	79	0.098	765	0.947
211	27	0.033	792	0.980
212	8	0.010	800	0.990
213	7	0.009	807	0.999
214	1	0.001	808	1.000

La distribución de frecuencia de la tabla 2.5 se representa gráficamente en la Fig. 4 en la cual el número de cilindros corresponde a la abscisa y la frecuencia correspondiente es la ordenada. Esta figura muestra claramente que la variable solo toma valores aislados.

El polígono de frecuencias acumuladas se muestra en la Fig. 5. Hay autores que lo nombran diagrama de escalones donde el peralte es igual a la frecuencia.

3.2. PARAMETROS

3.2.1. Medidas de Tendencia Central.

Las observaciones tienen tendencia a agruparse alrededor de un cierto punto de la escala de referencia. Ahora consideraremos parámetros que miden este agrupamiento, estos parámetros son llamados de tendencia central o de localización.

El parámetro más importante se le llama "Media Aritmética" o simplemente media, está dado por :

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} \quad ; \text{ ó}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (1)$$

Para datos agrupados, se obtiene una aproximación de la media usando la si-

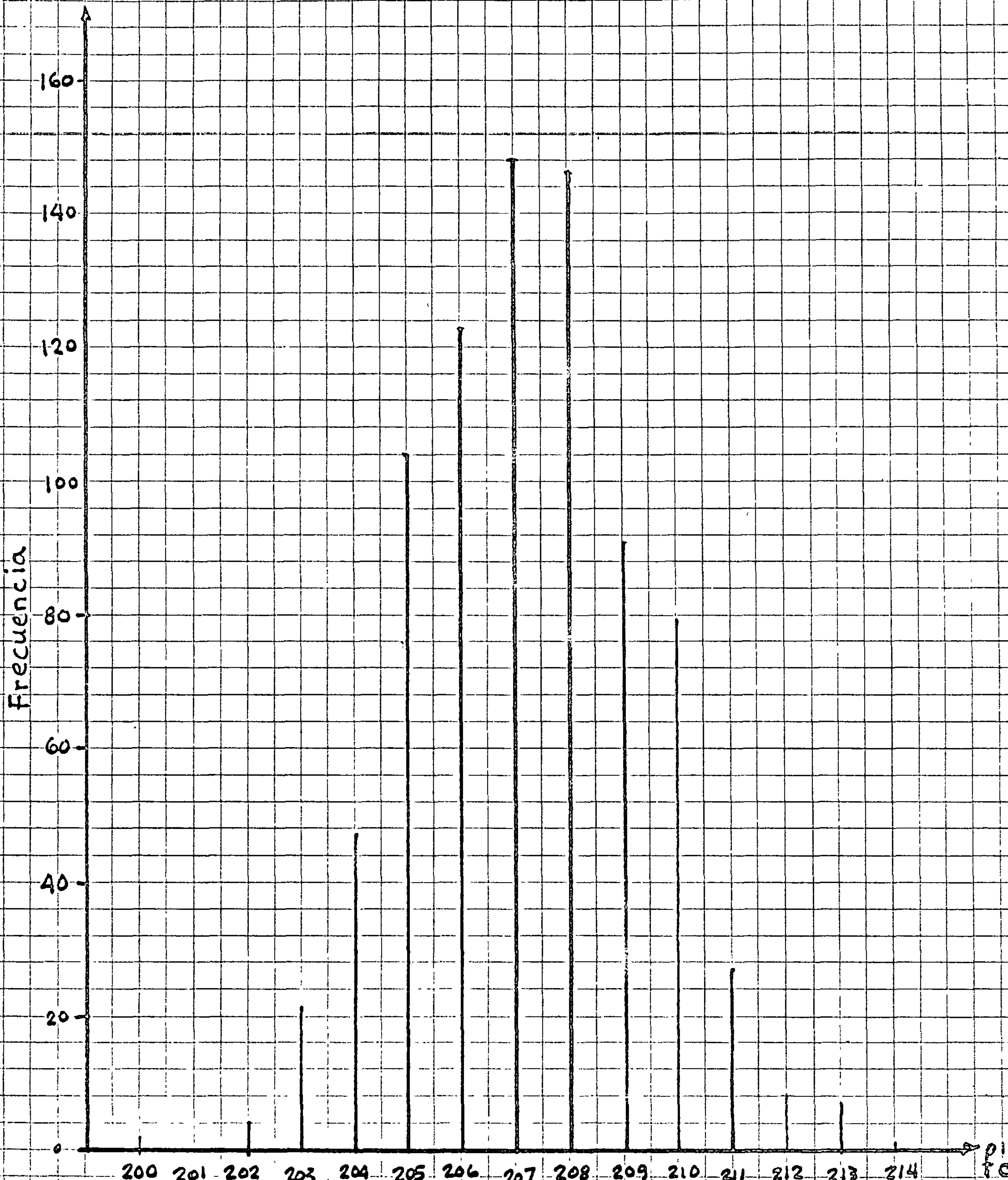


Fig 4 Distribución de Frecuencia del f'_c

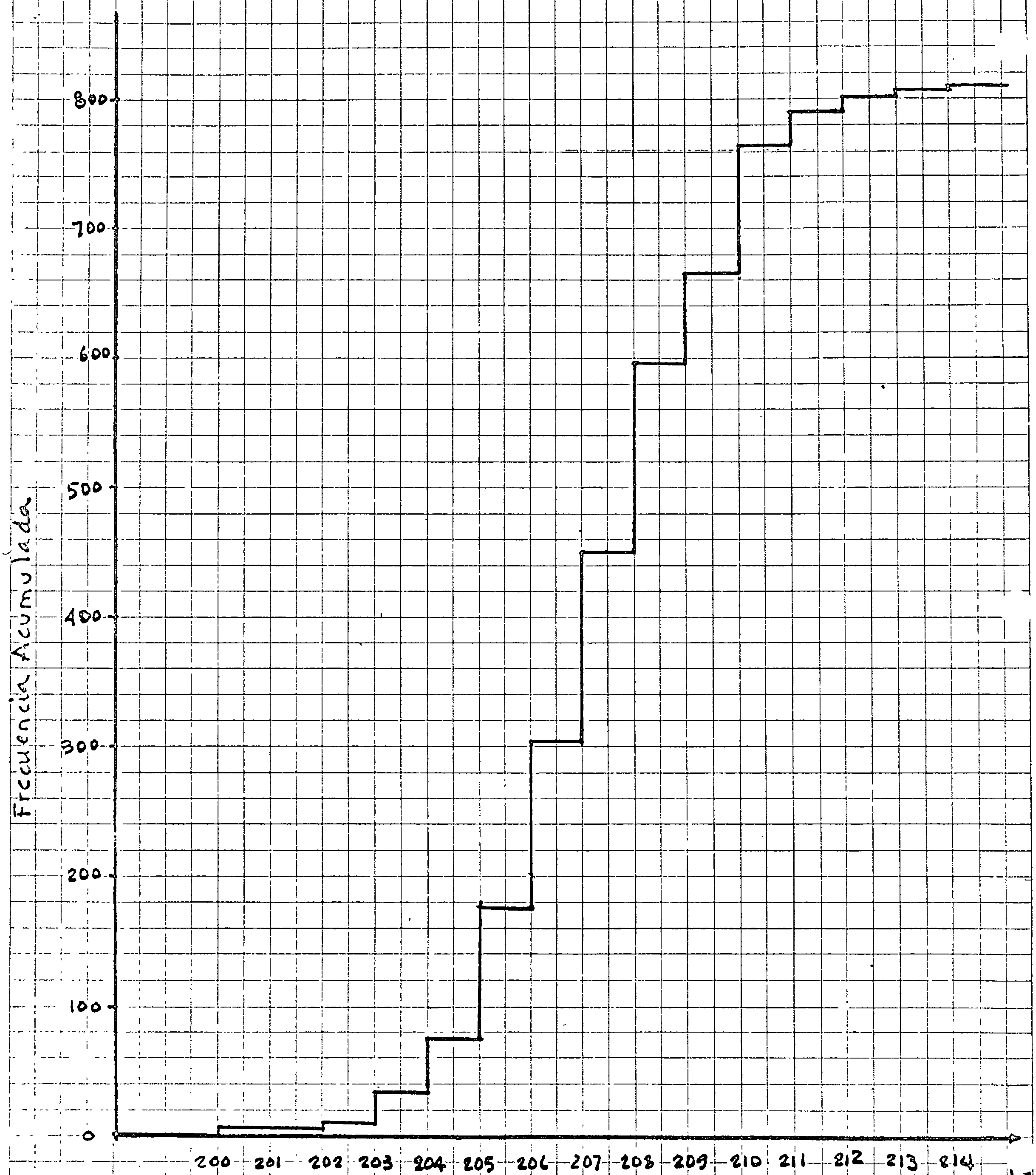


FIG 5. Polígono de Frecuencias Acumuladas.

guiente fórmula:

$$\mu_g = \frac{\sum_{i=1}^{k} f_i x_i}{N} \quad --- \quad (2)$$

dónde $N = \sum_{i=1}^{k} f_i$ donde k es el número de clases.

Ejemplo: Utilizando las observaciones del porcentaje de individuos en cada estado de 65 o más años de edad, tenemos

$$\mu = \frac{6.5 + 5.9 + \ldots + 6.3}{48}$$

$$\mu = 8.067\%$$

Aplicando (2)

$$\mu_g = \frac{3(5.05) + 6(6.05) + \ldots + 1(11.05)}{48}$$

$$\mu_g = 8.050\%$$

Los tres tipos de medida de tendencia central más usados son la mediana, la moda y las medias (aritmética, geométrica y armónica).

Mediana: Este parámetro se determina a partir de las observaciones ordenadas en forma ascendente y corresponde a la observación localizada en la mitad de la lista si el número de observaciones es impar. Si es número par se toma el promedio de las 2 observacio nes centrales.

Moda (M): Es el valor de X en una colección que ocurre con mayor frecuencia.

Media Geométrica (G) está dada por

$$G = \sqrt[N]{x_1 \, x_2 \ldots x_N} \qquad \text{ó} \qquad (3)$$

$$G = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

La Media Armónica (A) está dada por

$$A = \frac{N}{\sum_{i=1}^{N}\left(\frac{1}{x_i}\right)} \qquad (4)$$

3.2.2. Medidas de Dispersión.

A fin de determinar el grado de dispersión de los datos y como consecuencia obtener una mejor descripción de los mismos se tiene este tipo de parámetros, los más comúnmente usados son: Rango, Desviación Media, Varanza, Desviación Estandard y Coeficiente de Variación.

a) Rango: Está definido por :

$$R = x_N - x_1 \qquad (5)$$

b) Desviación Media: Está definida por

$$\delta_m = \frac{\sum_{i=1}^{N} |x_i - \mu_m|}{N} \qquad (6)$$

c) Varianza: Está definida por :

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad (7)$$

y para datos agrupados por:

$$\sigma_g^2 = \frac{\sum_{i=1}^{K} f_i (x_i - \mu)^2}{N} \qquad (8)$$

d) Desviación Estandar (d) se define como la raíz cuadrada de la varianza.

Ejemplo: para los datos de la tabla 1.

Aplicando (7)

$$\sigma^2 = \frac{(6.5 - 8.067)^2 + (5.9 - 8.067)^2 + \ldots + (6.3 - 8.067)^2}{48} = 2.13\,\%^2$$

$$\sigma = 1.46\%$$

Aplicando (8)

$$\sigma^2 = \frac{3(5.05 - 8.067) + 6(6.05 - 8.067) + \ldots + 1(11.05 - 8.067)^2}{48} = 2.29\,\%^2$$

e) El Coeficiente de Variación está dado por:

$$\nu = \frac{\sigma}{\mu} \quad ; \quad \nu = \frac{1.46}{8.067} = 0{,}181$$

Ejemplos. Demostrar que:

1.- La varianza de N valores dados X_1 X_2,... X_N está dada por

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}{N}$$

2.- Sumando una constante (positiva o negativa) a cada uno de los N valores, incrementa a la media la misma constante pero no afecta la varianza ni la desviación estándar.

3.- Multiplicando cada uno de los N valores por una constante, multiplica la media por esta constante y la varianza por la constante al cuadrado.

CURSO INTENSIVO DE

PROBABILIDAD Y ESTADISTICA

FUNDAMENTOS Y APLICACIONES

ING. FERNANDO GOMEZ JARDON

Sesión 7	14-IX-72	Modelos Probabilísticos Clásicos
Sesión 8	19-IX-72	Estimación de Parámetros de los Modelos.
Sesión 9	21-IX-72	Pruebas de Hipótesis.

DISTRIBUCION BINOMIAL O DE BERNOULLI

Experimentos tipo BERNOULLI son aquellos experimentos independientes en que sus resultados solo pueden ser dos: EXITO o FRACASO. Ambos mutuamente exclusivos e incompatibles

Ejem: el ganar o perder un concurso.
el estar defectuosa o nó alguna pieza fabricada.
el utilizar un dispositivo o nó, etc.

Probabilidad individual de ocurrencia (o de éxito) p es la probabilidad que en cada experimento de tipo Bernoulli se tenga un éxito. Como el éxito y el fracaso consideran exhaustivamente al espacio de eventos y si se designa con q a la probabilidad de fracaso, se debe cumplir

$$p+q=1$$

La función de distribución Binomial es un modelo probabilístico que permite calcular la probabilidad de que en n experimentos de tipo Bernoulli se obtengan x éxitos. Por lo tanto la variable aleatoria x puede escribirse como:

$$x = \xi_1 + \xi_2 + \dots + \xi_n$$

donde

ξ_i = variable aleatoria que puede tomar los valores

$$\begin{cases} 0 \text{ cuando se verifica el evento } F \text{ que representa fracaso} \\ 1 \text{ cuando se verifica el evento } E \text{ que representa éxito.} \end{cases}$$

sab iendo además que

$$P(E) = p$$
$$P(F) = q = (1-p)$$

Una posibilidad de obtener x éxitos en n experimentos sería:

$$P(\underbrace{E_1 \cap E_2 \cap \ldots \cap E_x}_{x \text{ éxitos}} \cap \underbrace{F_{x+1} \cap F_{x+2} \cap \ldots \cap F_n}_{n-x \text{ fracasos}})$$

como los experimentos son independientes, la probabilidad de las interseccio - nes de eventos se puede calcular mediante el producto de las respectivas pro - babilidades independientes de cada uno de los eventos intersectados, o sea que lo anterior será igual a:

$$\underbrace{p \cdot p \cdot \ldots \cdot p}_{x \text{ factores}} \cdot \underbrace{q \cdot q \cdot \ldots \cdot q}_{n-x \text{ factores}} = p^x q^{n-x} = p^x (1-p)^{n-x}$$

Pero lo anterior representa sólo una de las combinaciones de ocurrencias de - éxitos y fracasos que satisfacen el que en n experimentos se tengan x éxi- tos, donde estos últimos pueden aparecer en cualquier orden. Por lo que habrá que multiplicar la expresión $p^x(1-p)^{n-x}$ (constante, ya que el orden de los fac tores no altera al producto) por el número de permutaciones que pueden realizar se con n elementos donde x son iguales y n-x también, es decir:

$$P_n^{x,n-x} = \frac{n!}{x!(n-x)!} = C_x^n = \binom{n}{x}$$

finalmente la probabilidad de obtener x éxitos en n experimentos de tipo - Bernoulli queda:

$$\boxed{p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}} \qquad x = 0, 1, 2, \ldots, n$$

que es la función de distribución de probabilidades Binomial o de Bernoulli B (n,p) con

$$\mu_X = np$$
$$\sigma_X^2 = npq$$

LA DISTRIBUCION GEOMETRICA.

Considerando también experimentos de tipo Bernoulli puede interesar calcular - la probabilidad de que el primer éxito ocurra en el n-ésimo experimento o sea $P_N(n)$, la cual se puede calcular tomando en cuenta lo visto para la distribu - ción Binomial, como sigue:

$$\boxed{p_N(n) = q^{n-1}p = (1-p)^{n-1}p} \qquad n = 1, 2, \ldots$$

que es la distribución de probabilidad Geométrica G (n). n es el número del - experimento donde ocurre el primer éxito.

$$\mu_N = \frac{1}{p} \qquad \text{(período medio de retorno)}$$

$$\sigma_N^2 = \frac{q}{p^2} = \frac{1-p}{p^2}$$

Ejemplo:

Si la probabilidad de que la cotización de un contratista sea baja en cada una de 3 propuestas independientes es 1/3, calcular la distribución de probabilidades del número de propuestas exitosas sería:

$$p_X(x) \begin{cases} 0 & \binom{3}{0}\left(\frac{1}{3}\right)^0 \left(1-\frac{1}{3}\right)^3 = 8/27 \\ 1 & \binom{3}{1}\left(\frac{1}{3}\right)^1 \left(1-\frac{1}{3}\right)^2 = 12/27 \\ 2 & \binom{3}{2}\left(\frac{1}{3}\right)^2 \left(1-\frac{1}{3}\right)^1 = 6/27 \\ 3 & \binom{3}{3}\left(\frac{1}{3}\right)^3 \left(1-\frac{1}{3}\right)^0 = 1/27 \end{cases}$$

a) la probabilidad de que no gane sería cercana a 1/3 $(8/27)$

b) la probabilidad de que gane cuando más 2 sería

$$P[X \leq 2] = p_X(0) + p_X(1) + p_X(2)$$
$$= 1 - p_X(3) = \frac{26}{27}$$

c) la prob abilidad de que gane al menos 2 sería

$$P[X \geq 2] = p_X(2) + p_X(3) = \frac{7}{27}$$

aproximadamente igual a la probab ilidad de que no gane ningún concurso.

Si la decisión de los concursos se hace una por mes en los siguientes tres meses. La probabilidad de que se gane el primer concurso:

el 1er. mes $P_N(1) = \left(1-\frac{1}{3}\right)^{1-1}\left(\frac{1}{3}\right) = \frac{1}{3} = 0.33$

el 2o. mes $P_N(2) = \left(1-\frac{1}{3}\right)^{2-1}\left(\frac{1}{3}\right) = \frac{2}{9} = 0.22$

el 3er. mes $P_N(3) = \left(1-\frac{1}{3}\right)^{3-1}\left(\frac{1}{3}\right) = \frac{4}{27} = 0.148$

aplicando la distribución geométrica

DISTRIBUCION DE POISSON

Cuando se tiene un número muy grande de experimentos de tipo Bernoulli donde no es posible identificarlos individualmente a cada uno de ellos, por ejemplo, cuando los sucesos pueden ocurrir en cualquier instante de un intervalo de tiempo o en cualquier punto a lo largo de una línea o de una superficie, surge el modelo de la distribución de POISSON también denominado de ocurrencias aleatorias.

A partir de la distribución de BERNOULLI $p_x(n)=\binom{n}{x} p^x(1-p)^{n-x}$ se puede calcular la probabilidad de que en un intervalo de tiempo $t = n$ segundos, (que equivalen cada uno de ellos a experimentos de tipo Bernoulli) ocurran x sucesos, conociendo que la probabilidad individual de ocurrencia en cada segundo es p y despreciando la probabilidad de que ocurran 2 ó más sucesos en un mismo segundo o experimento. Sin embargo, puede convenir considerar duraciones de tiempo cada vez más y más pequeñas, con lo cual el número de experimentos de tipo Bernoulli crecen $(n \to \infty)$, pero al mismo tiempo la probabilidad de ocurrencia disminuye $(p \to 0)$, debiéndose mantener constante el número esperado de eventos en todo el intervalo, igual a $np = \nu$.

Tomando en cuenta lo anterior en la distribución Bernoulli:

$$p_x(x) = \frac{n!}{(n-x)!\,x!}\, p^x (1-p)^{n-x} \qquad \text{sustituyendo } \frac{\nu}{n} \text{ por } p$$

$$= \frac{n(n-1)(n-2)\ldots[n-(x-1)]}{x!} \left(\frac{\nu}{n}\right)^x \left(1-\frac{\nu}{n}\right)^{n-x}$$

$$= \left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\right] \frac{\nu^x}{x!} \frac{\left(1-\frac{\nu}{n}\right)^n}{\left(1-\frac{\nu}{n}\right)^x}$$

si se toma el límite de la exp resión anterior cuando $n \to \infty$ manteniendo - constante ν y sabiendo que:

$$\lim_{n \to \infty} \left(1 - \frac{\nu}{n}\right)^n = e^{-\nu}$$

$e = 2.7182818$

resulta la distrib ución de Poisson. $PO(\nu)$

$$\boxed{p_x(x) = \frac{\nu^x}{x!} e^{-\nu}} \qquad x = 0, 1, 2, \ldots, \infty$$

Algunas propiedades:

ν es el valor esperando del número de ocurrencias en el intervalo de tiempo considerado.

x es el número de ocurrencias que pueden presentarse en el intervalo de tiempo estudiado.

$\mu_x = \nu$

$\sigma_x^2 = \nu$

Es una distribución de probabilidades "regenerativa" (Significa que la suma de dos o más variables aleatorias de Poisson con parámetros $\nu_1, \nu_2, \ldots, \nu_k$ da por resultado otra variable aleatoria también de Poisson con parámetro. $\nu = \nu_1 + \nu_2 + \ldots + \nu_k$).

Con objeto de facilitar la aplicación práctica de la fórmula de Poisson se sustituye ν por λt donde λ representa la tasa media (valor esperado) del número de ocurrencias para un intervalo unitario (un segundo, un minuto, etc) y t el número de intervalos unitarios considerados, de tal manera que el producto $\lambda t = \nu$ (constante) constituye el valor esperado de ocurrencias en el intervalo total de tiempo formado por t intervalos unitarios; por lo tanto:

$$p_X(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$$

Lo anterior también puede verse facilmente a partir de la propiedad regenerativa, donde cada uno de los intervalos agregados correspondería a una variable aleatoria sumada.

La deducción de la distribución de Poisson a partir de la Binomial sugiere la utilización de la primera cuando n es grande y p chica facilitando notablemente los cálculos que implica la Binomial, sobre todo cuando n es grande. Para efectos prácticos esta sustitución puede hacerse sin pérdida de precisión cuando se cumple que $n \geq 50$ y $np \leq 5$.

DISTRIBUCION DE POISSON (ejemplo)

Si el número promedio de llamadas que llegan por minuto a un conmutador telefó nico es de 2.5, calcular la probabilidad de que en 2 minutos se reciban más de 3 llamadas.

$$p_X(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} = \frac{(2.5 \times 2)^x}{x!} e^{-2.5\times 2} = \frac{5^x}{x!} e^{-5}$$

la probabilidad que se busca es: $1 - P[x \leq 3]$

$$P[x \leq 3] = F_X(3) = P_X(0) + P_X(1) + P_X(2) + P_X(3)$$

$$P_X(0) = \frac{5^0 e^{-5}}{0!} = e^{-5}$$

$$P_X(1) = \frac{5 e^{-5}}{1!} = 5e^{-5}$$

$$P_X(2) = \frac{5^2 e^{-5}}{2!} = \frac{25}{2} e^{-5} = 12.5e^{-5}$$

$$P_X(3) = \frac{5^3 e^{-5}}{3!} = \frac{125}{6} e^{-5} = 2.083e^{-5}$$

$$F_X(3) = 20.583\ e^{-5} = 20.583 \times 0.00668 = 0.137$$

$$1 - 0.137 = 0.863$$

DISTRIBUCION EXPONENCIAL.

Si se considera un proceso de ocurrencias aleatorias cuyo número en un intervalo dado de tiempo t tiene una probabilidad que puede obtenerse mediante fórmula de Poisson, e interesa calcular la probabilidad de los tiempos entre ocurrencias, cuyos valores se representen a través de la variable aleatoria T, puede efectuarse el siguiente razonamiento:

La probabilidad de que T sea mayor que algún valor del tiempo t debe ser igual a la probabilidad de que no ocurran exentos en el intervalo de tiempo t, o sea:

$$1 - F_T(t) = p_X(0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

$$F_T(t) = 1 - e^{-\lambda t}$$

$$f_T(t) = \frac{dF_T(t)}{dt} = \lambda e^{-\lambda t} \quad \text{para } t > 0 \quad \text{(continua)}$$

$$\boxed{f_T(t) = \lambda e^{-\lambda t}}$$

notación $EX(\lambda)$

que es la distribución de probabilidad exponencial con las siguientes propiedades:

$$\mu_T = \frac{1}{\lambda}$$

$$\sigma_T^2 = \frac{1}{\lambda^2}$$

λ es la tasa media de ocurrencia de los eventos por unidad de tiempo.

T es el tiempo entre ocurrencias.

DISTRIBUCION EXPONENCIAL (ejemplo)

Si se sabe que a una gasolinería llegan a razón de 3 automóviles cada 10 minutos, cual es la probabilidad de que el encargado de una bomba tenga que esperar más de 5 minutos sin despachar.

$$f_T(t) = \lambda e^{-\lambda t} \quad ; \quad F_T(t) = 1 - e^{-\lambda t}$$

$\lambda = \frac{3}{10}$ automóviles por minuto.

la probabilidad de que tenga que esperar 5 minutos exactamente es:

$$f_T\ (5) = 0.3\ e^{-1.5} = 0.0669$$

la probabilidad de que tenga que esperar más de 5 minutos es:

$$P[T > 5] = 1 - F_T(5)$$

$$= 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$$

$$= e^{-0.3 \times 5} = e^{-1.5} = 0.224$$

DISTRIBUCION NORMAL O DE GAUSS.

Este modelo probabilístico ocurre básicamente cuando la incertidumbre sobre una variable física es el resultado de los efectos combinados de muchas otras variables cuyo análisis por separado resultaría demasiado complejo o no se justifica.

También se le denomina a este modelo el MODELO de SUMAS, debido a que puede demostrarse que en general la distribución de probabilidades de una suma de variables aleatorias se aproxima a una distribución, sobre todo cuando el número de variables sumadas es grande y ninguna de ellas es demasiado significativa respecto a las demás. (Teorema del límite central).

Como ejemplos se pueden plantear los errores en la medida de una distancia obtenida mediante varios tramos de mismo tamaño; el tiempo en realizar una obra que consta de una serie de actividades, con sus respectivas duraciones, etc.

Aparte de lo anterior esta distribución se utiliza en ocasiones en que a pesar de que el hecho que trata de representarse no corresponde a una suma de variables aleatorias su figura acampanada se ajusta al comportamiento por representar.

La forma de su expresión matemática es una doble exponencial:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \qquad -\infty \leq x \leq \infty$$

(variable aleatoria continua)

$e = 2.7182818..$

$\pi = 3.1416$

notación $N(\mu_x, \sigma_x)$

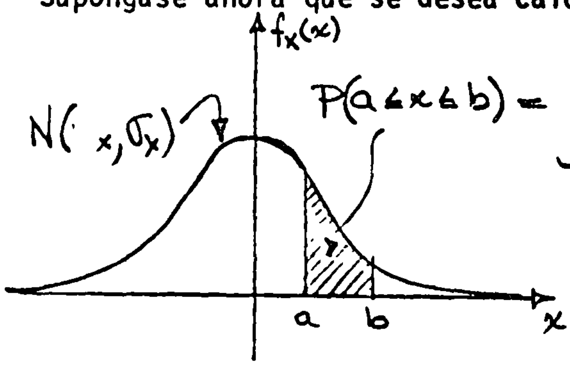
Propiedades:

1) Es simétrica respecto al eje $x = \mu_x$

2) El máximo está en $x = \mu_x$

3) Cuando x tiende a $\begin{cases} +\infty \\ ó \\ -\infty \end{cases}$ es asintótica con el eje x

Supóngase ahora que se desea calcular $P(a \leq x \leq b)$



$$P(a \leq x \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} dx$$

Como este tipo de cálculos es muy frecuente, es necesario buscar un artificio que en la práctica permita obtenerlos de manera sencilla y por lo tanto se procede de la siguiente forma, denominada LA ESTANDARIZACION de la variable aleatoria:

Sea el cambio de variable

$$z = \frac{x - \mu_x}{\sigma_x} \quad ; \quad dz = \frac{dx}{\sigma_x}$$

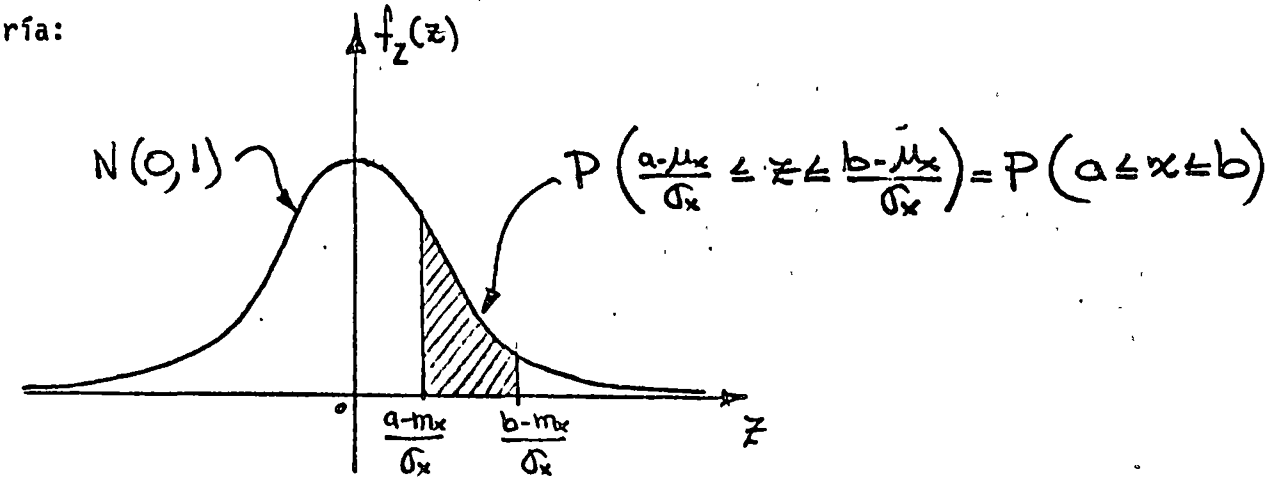
$$\text{si } x = a \ ; \quad z = \frac{a - \mu_x}{\sigma_x}$$

$$x = b \ ; \quad z = \frac{b - \mu_x}{\sigma_x}$$

con lo que

$$P(a \leq x \leq b) = \int_{\frac{a-\mu_x}{\sigma_x}}^{\frac{b-\mu_x}{\sigma_x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Si se observa la función que se esta integrando se puede afirmar que se trata de otra función de distrib ución normal denominada ESTANDARIZADA y que se caracteriza por tener $\mu_x = 0$ y $\sigma_x^2 = 1$ o sea N (0,1) cuya gráfica sería:



La importancia de lo anterior radica en el hecho de que CUALQUIER tipo de distribución normal N (μ_x, σ_x) puede transformarse a una N (0, 1) la cual se encuentra tabulada en casi todos los libros de Probabilidad y/o Estadística con lo que se facilita notablemente el manejo de esta distribución de probabilidades.

La distribución Normal o de Gauss también se emplea como una adecuada aproximación a otras distribuciones, aún del tipo discreto, para facilitar su cálculo.

En el caso de la distribución Binomial su aproximación es buena cuando n es grande y p está alejado de cero y uno (se recomienda en la práctica cuando np y nq son ≥ 5). Entonces la variable estandarizada $z = \frac{x - np}{npq}$ se comporta aproximadamente como una N (0, 1). Tratándose de la distribución de Poisson, esta se aproxima a la normal para valores de ν grandes, pudiendose utilizar en tal caso a la variable $z = \frac{x - \nu}{\sqrt{\nu}}$ distribuida N (0,1).

En la práctica para valores de $\nu > 10$.

Problema ilustrativo para la aproximación de la distribución normal a la binomial.

En un crucero donde la mitad de los vehículos que llegan de vuelta hacia la izquierda, encontrar la probabilidad de que entre 3 y 6 autos de 10 que llegan al crucero vayan a dar vuelta.

a) Usando la distribución Binomial.

b) Usando la aproximación a la Normal.

a) $B(np, \sqrt{npq}) \quad p = \frac{1}{2}$

$$P(3 \leq x \leq 6) = p_x(3) + p_x(4) + p_x(5) + p_x(6)$$

$$p_x(x) = \frac{n!}{(n-x)!\,x!} p^x (1-p)^{n-x}$$

$$p_x(3) = \frac{10!}{7!\,3!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \frac{120}{1024}$$

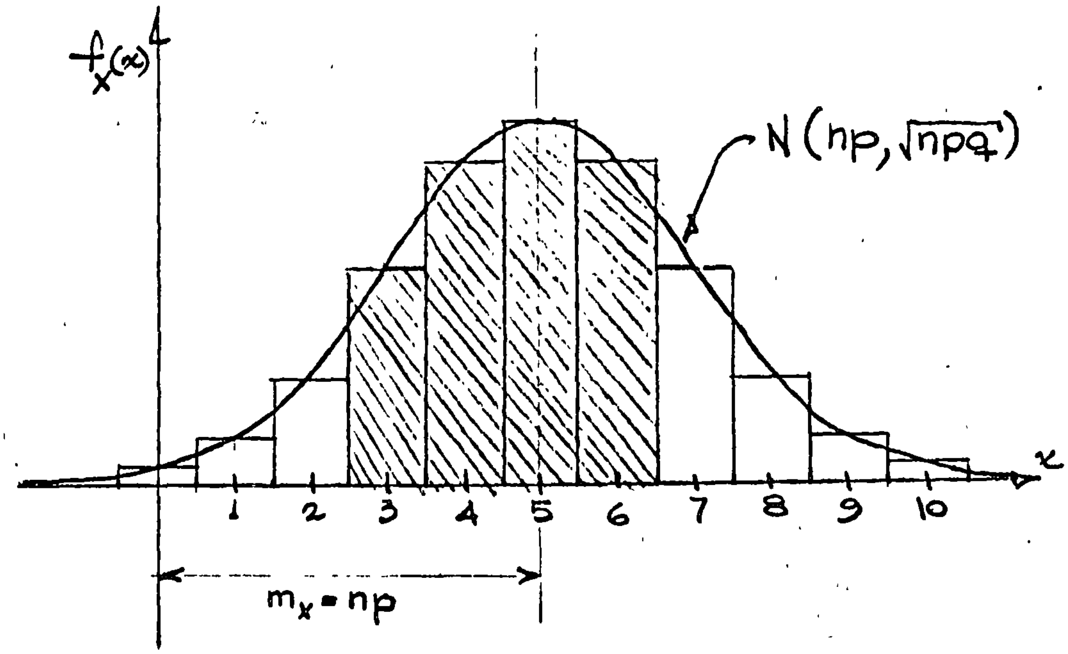
$$p_x(4) = \frac{10!}{6!\,4!}\left(\frac{1}{2}\right)^{10} = \frac{210}{1024}$$

$$p_x(5) = \frac{10!}{5!\,5!}\left(\frac{1}{2}\right)^{10} = \frac{252}{1024}$$

$$p_x(6) = \frac{10!}{4!\,6!}\left(\frac{1}{2}\right)^{10} = \frac{210}{1024}$$

$$P(3 \leq x \leq 6) = \frac{792}{1024} = 0.7734$$

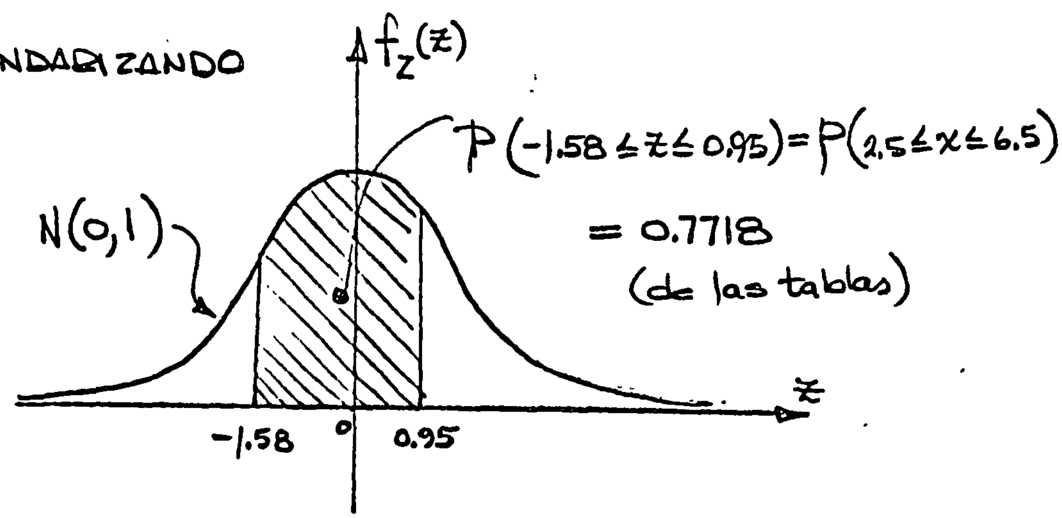
b)



$$z_{2.5} = \frac{2.5 - np}{\sqrt{npq}} = \frac{2.5 - 5}{\sqrt{2.5}} = -1.58$$

$$z_{6.5} = \frac{6.5 - np}{\sqrt{npq}} = \frac{6.5 - 5}{\sqrt{2.5}} = 0.95$$

ESTANDARIZANDO



ESTIMACION DE LOS PARAMETROS DEL MODELO

Para que los modelos probabilísticos dejen de ser una mera concepción teórica y puedan utilizarse en la práctica como modelos del comportamiento real de los elementos de la POBLACION que pretenden representar, es necesario, en primer lugar, estimar los parámetros contenidos en dichos modelos y en segundo lugar estudiar la naturaleza de estos últimos como variables aleatorias.

Teóricamente, los parámetros de los modelos (θ) se pueden obtener como una función de los elementos de la POBLACION, es decir $\theta = \Theta(x_1, x_2, \dots)$
Sin embargo, esta situación resulta dificil llevarla a cabo principalmente por el número de elementos de la población que en muchas ocasiones llega a ser infinito.

Para la distribución Normal o de Gauss los parámetros son:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i = E[x]$$

$$\sigma_x = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu_x)^2 = E[(x-\mu_x)^2]$$

En vista de la dificultad señalada, se recurre a estimar los parámetros a partir de las ESTADISTICAS muestrales o simplemente estadísticas $(\hat{\theta})$ que se obtienen en función de los n elementos de una MUESTRA (de tamaño n) que es un subconjunto finito de la POBLACION, o sea:

$$\hat{\theta}\ (\text{estadística estimadora de } \theta) = \hat{\theta}\ (x_1, x_2, \dots, x_n)$$

Como es facil suponer, las estadísticas para una misma población pueden tener diferentes valores dependiendo estre otras cosas de la mecanica del muestreo, por lo que, como se dijo anteriormente es necesario estudiarlas considerándolas como variables aleatorias, dando lugar a lo que se denomina DISTRIBUCIONES MUESTRALES.

Como ejemplos de estadísticas se tienen la media y la variancia de las muestras:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = E[x] \quad \text{(momento de orden uno respecto al origen)}$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = E\left[(x-\bar{x})^2\right] \quad \text{(momento de orden 2 respecto a la media)}$$

De acuerdo con lo anterior, es importante distinguir en el proceso de adaptación de los modelos probabilísticos a la realidad, 3 tipos diferentes de incertidumbre, a saber:

1) La que resulta de la naturaleza aleatoria del fenómeno por analizar y que es irreductible.

2) La que resulta de la estimación de los parámetros del modelo, llamada incertidumbre estadística, que puede reducirse en la medida en que se observe un mayor número de datos, y

3) La que resulta de la forma del modelo en si, en cuanto a que si el modelo realmente representa al fenómeno, la cual también puede disminuirse incrementando el número de observaciones.

METODO DE LOS MOMENTOS

Este método consiste en estimar los momentos de modelo probabilístico propuesto (μ_x, σ_x^2) a partir de los momentos de correspondientes la muestra $(\bar{x}, s_x^2)$ donde $\bar{x} = \hat{\mu}_x$ y $s_x^2 = \hat{\sigma}_x^2$. Normalmente si los parámetros no son equivalentes a algunos de los momentos, es posible contar con una expresión matemática que los relacione, como en el caso de la distribución exponencial donde $\mu_T = \frac{1}{\lambda}$ y por tanto $\hat{\lambda} = \frac{1}{\bar{t}}$, por ejemplo.

EJEMPLO 1 (OBTENCION DE LOS PARAMETROS DE UNA DISTRIBUCION NORMAL)

Se trata de adaptar el modelo de la distribución Normal a la resistencia de vigas de concreto reforzado.

DATOS		DATOS	
No. de Prueba	Resistencia	No. de Prueba	Resistencia
1	8560	14	8460
2	8700	15	8500
3	8420	16	8420
4	8580	17	8500
5	8460	18	8620
6	8440	19	8440
7	8460	20	8420

DATOS No. de Prueba	Resistencia	DATOS No. de Prueba	Resistencia
8	8440	21	8420
9	8550	22	8410
10	8430	23	8700
11	8620	24	8360
12	8580	25	8310
13	8650	26	8330
		27	8510

donde

$\bar{x} = 8490$

$S_x^2 = 10\,522$

$S_x = 103$

El modelo resultante será:

$$f_X(x) = \frac{1}{103\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-8490}{103}\right)^2}$$

EJEMPLO 2 (OBTENCION DEL PARAMETRO DE UNA DISTRIBUCION EXPONENCIAL)

Se supone un tráfico Poisoniano y se tienen los datos observados de los tiempos entre llegadas de vehículos clasificados en intervalos de clases de un -- minuto, según se muestra en la siguente tabla.

Tiempos entre llegadas	Frecuencia	Tiempos entre llegadas	Frecuencia
0-1	18	14-15	3
1-2	25	15-16	3
2-3	21	16-17	6
3-4	13	17-18	4
4-5	11	18-19	3
5-6	15	19-20	3
6-7	16	20-21	1

Tiempos entre llegadas	Frecuencia	Tiempos entre llegadas	Frecuencia
7-8	12	21-22	1
8-9	11	22-23	1
9-10	11	23-24	0
10-11	8	24-25	1
11-12	12	25-26	0
12-13	6	26-27	1
		28-29	1
		29-30	2
		30-31	1

De donde:

$\sum t_i f_i = 1640$

$\bar{t} = 7.66$ seg

Modelo $= f_T(t) = \lambda e^{-\lambda t} \quad t \geq 0$

$$\mu_T = \frac{1}{\lambda} \quad ; \quad \lambda = \frac{1}{\mu_T}$$

según el metodo de los momentos el estimador de μ_T es $\bar{t}$;

$$\hat{\lambda} = \frac{1}{\bar{t}} = \frac{1}{7.66} = 0.130/\text{seg}$$

Por lo tanto la probabilidad de que el tiempo entre llegadas fuera entre 8 y 10 segundos se podría calcular

$$F_T(10) - F_T(8) = 1 - e^{-0.130(10)} - \left[1 - e^{-0.130(8)}\right]$$

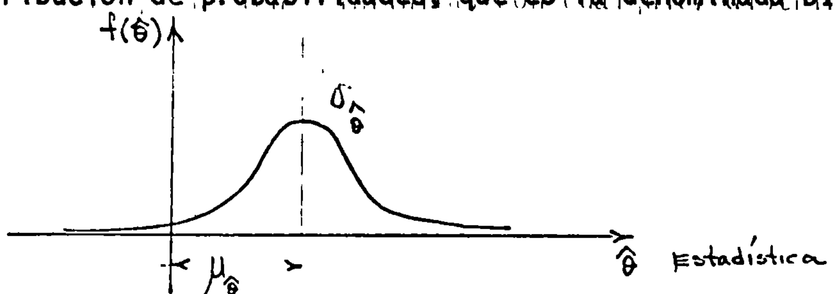
$$= e^{-1.04} - e^{-1.3} = 0.081$$

INTERVALOS DE CONFIANZA

Este método, a diferencia del de los momentos y del de máxima verosimilitud - que permiten una estimación puntual, es decir un valor único, conduce a la estimación de un intervalo donde con cierta probabilidad(confianza) se encuentra el parámetro por estimar. La aplicación de este método requiere del concepto de distribuciones muestrales que se vera a continuación.

DISTRIBUCIONES MUESTRALES

Si se considera el número de muestras del mismo tamaño con elementos independientes que pueden extraerse de una población y que con cada una de ellas se puede calcular una estadística, es fácil pensar que para los distintos valores de esta última, considerada como variable aleatoria, se puede obtener una distribución de probabilidades, que es la denominada DISTRIBUCION MUESTRAL.



Estimador insesgado se define como aquel que $E[\hat{\theta}] = \theta$ o sea que la media de su distribución muestral es igual al parámetro por estimar.

La distribución muestral de las medias corresponde a aquella distribución donde la variable aleatoria es la media de las muestras de un mismo tamaño que pueden extraerse de una población. Esta distribución tiene como propiedades más importantes a las siguientes:

1) Para poblaciones infinitas con cualquier distribución de probabilidad o muestreando con reemplazo se tiene una distribución con:

$$\mu_{\bar{x}} = \mu_x \quad \text{(media de la población)}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

n tamaño de la muestra
σ_x^2 varianza de la población

2) Para poblaciones finitas de tamaño N, con cualquier distribución de probabilidad muestreando sin reemplazo, la distribución tiene:

$$\mu_{\bar{x}} = \mu_x$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

n tamaño de la muestra
N tamaño de la población
$\sqrt{\frac{N-n}{N-1}}$ factor de corrección por tamaño de la población

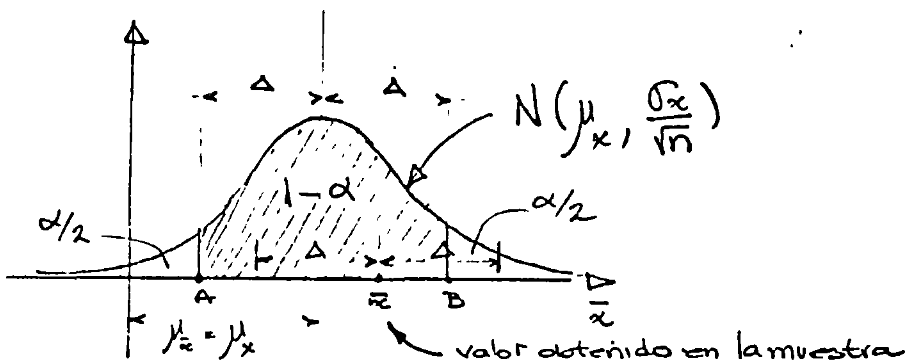
3) Para poblaciones con cualquier distribución de probabilidades y muestreando con muestras grandes (para efectos prácticos $n \geq 30$) la distribución muestral se aproxima a una distribución Normal.

$$N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

4) Para poblaciones distribuidas normalmente $N(\mu_x, \sigma_x)$ y para cualquier tamaño de muestra, la distribución muestral es una distribución normal.

$$N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

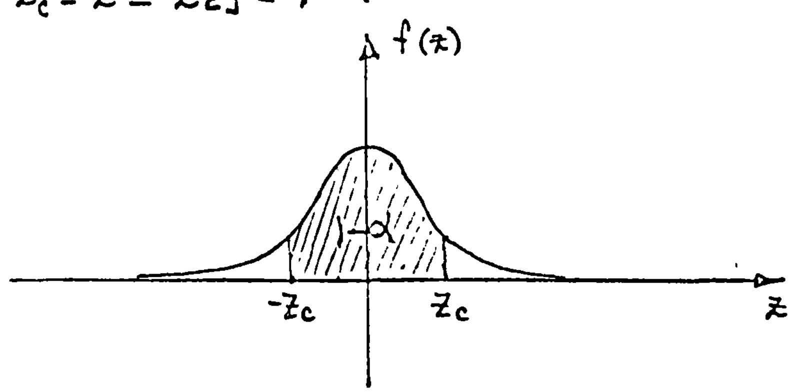
En el caso de que se muestree a una población infinita (o con reemplazo) utilizando muestras grandes y considerando que los valores de las medias corresponden a muestras más representativas en la medida en que aquellos se acerquen más a su valor esperado, y sí se tiene una confianza en el muestreo (de acuerdo con su mecánica de obtención) de que se obtendrá una de las muestras que este dentro del $(1-\alpha)\%$ (nivel de confianza) más representativas, el valor obtenido en el muestreo estara en el intervalo $A \leq \bar{x} \leq B$



Conforme a lo anterior es razonable pensar que con un $(1-\alpha)\%$ de confianza el valor del parámetro por estimar estará dentro del intervalo $\bar{x} \pm \Delta$, que será el intervalo de confianza del $(1-\alpha)\%$ para el parámetro por estimar que en este caso será μ_x

Finalmente el valor de Δ y por consiguiente de A y B en la figura anterior se pueden obtener sabiendo que la distribución muestral es una distribución normal $N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ que al estandarizarse en una distribución $N(0,1)$ los puntos A y B corresponden a $-z_c$ y a z_c respectivamente, de tal manera que

$$P[-z_c \leq z \leq z_c] = 1-\alpha$$



Por lo tanto aplicando la expresión del cambio de variable para la estandarización, se tendrá:

$$Z_c = \frac{B-\mu_x}{\sigma_x/\sqrt{n}} \quad ; \quad \Delta = B-\mu_x = Z_c \frac{\sigma_x}{\sqrt{n}}$$

Quedando finalmente el intervalo de confianza del $(1-\alpha)\,\%$ para la media de la población.

$$\boxed{\bar{x} \pm Z_c \frac{\sigma_x}{\sqrt{n}}}$$

donde

$$Z_c \mid P\left[-Z_c \leq Z \leq Z_c\right] = 1-\alpha$$

Por otro lado puede demostrarse que S_x (desviación estandar de la muestra) tiende a σ_x (desviación estandar de la población) cuando se trata de -- muestras grandes, y en este caso se pueden considerar iguales.

Por ejemplo, si la variable aleatoria x está distribuida $N(\mu_x, \sigma_x)$ en intervalo de confianza sería

$$\bar{x} \pm Z_c \frac{\sigma_x}{\sqrt{n}},$$

n tamaño de la muestra

$\bar{x} = \hat{\mu}_x$

$\sigma_x = S_x$

Si la v.a.x está distribuida Binomialmente con media $\mu_x = p$ y $\sigma_x = \sqrt{p(1-p)}$ (para cada experimento de tipo Bernoulli), el intervalo de confianza será

$$\bar{x} \pm Z_c \sqrt{\frac{p(1-p)}{n}} \qquad \bar{x} = \hat{\mu}_x = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$p = \hat{p}$$

Si la variable aleatoria t está distribuida según una exponencial con $\mu_T = \frac{1}{\lambda}$ y $\sigma_T = \frac{1}{\lambda}$

donde

$$\bar{t} \pm Z_c \frac{1}{\lambda\sqrt{n}} \qquad \bar{t} = \hat{\mu}_T = \frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} t_i$$
$$\lambda = \hat{\lambda}$$

Si la variable aleatoria x esta distribuida según una poisson con $\mu_x = \nu$ $\sigma_x^2 = \nu$

donde

$$\bar{x} \pm Z_c \sqrt{\frac{\nu}{n}} \qquad \bar{x} = \hat{\mu}_x = \hat{\nu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\nu = \hat{\nu}$$

EJEMPLO:

Si la proporción observada en una encuesta de 50 futuros inquilinos de una -- unidad habitacional, indica que 3 de cada 5 de ellos poseen automovil, ¿cuántos estacionamientos deberán construirse para una unidad habitacional con 423 de-- partamentos teniendo una confianza del 95%?

$$\bar{x} = \hat{p} \qquad \bar{x} \pm Z_{95\%} \sqrt{\frac{p(1-p)}{n}}$$

$$\hat{p} \pm \underbrace{1.96 \sqrt{\frac{0.6 \cdot 0.4}{50}}}_{0.136}$$

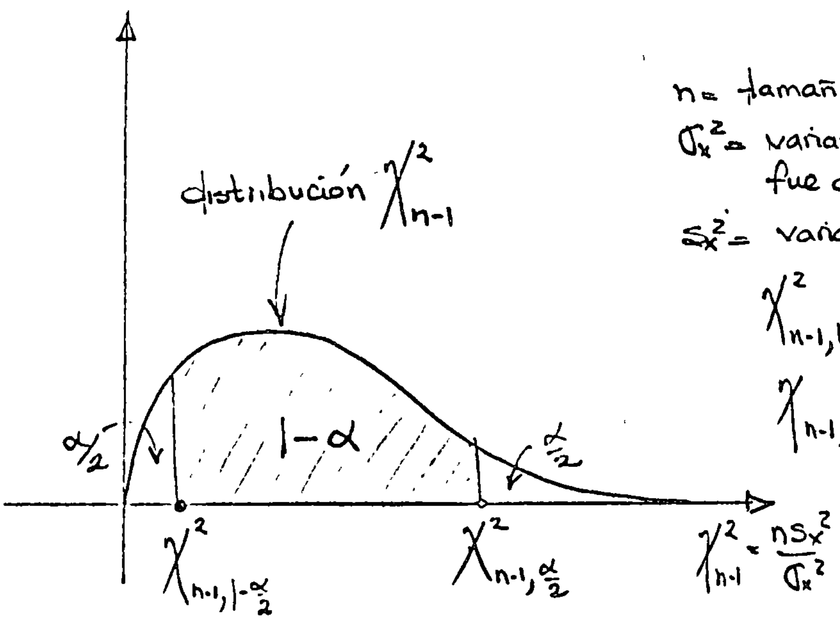
$$0.464 \leq p \leq 0.736$$

lo que se transformaria a No de estacionamientos

$$0.464 \times 423 = 196 \leq \text{nº de estacionamientos} \leq 0.736 \times 427 = 311$$

Distribuciones muestrales para la variancia de las muestras (cuando la variable x de la población esta normalmente distribuida)

En este caso se puede demostrar que la estadística s_x^2 modificada de la siguiente manera $\frac{n s_x^2}{\sigma_x^2} = \chi^2_{n-1}$ tiene una distribución de probabilidad llamada ji cuadrada con n-1 grados de libertad la cual se encuentra ampliamente tabulada como la Normal estandarizada. Conociendo esta propiedad se puede obtener el intervalo de confianza para la s_x^2 como se indica a continuación



donde.

n = tamaño de la muestra

σ_x^2 = variancia de la población de donde fue extraída

s_x^2 = variancia de la muestra

$\left.\begin{array}{l}\chi^2_{n-1,1-\frac{\alpha}{2}} \\ \chi^2_{n-1,\frac{\alpha}{2}}\end{array}\right\}$ se obtienen de las tablas.

pudiéndose escribir $P\left[\chi^2_{n-1,1-\frac{\alpha}{2}} \leq \frac{n s_x^2}{\sigma_x^2} \leq \chi^2_{n-1,\frac{\alpha}{2}}\right] = 1-\alpha$

haciendo transformaciones se llega a:

$$P\left[\frac{nS_x^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \leq \sigma_x^2 \leq \frac{nS_x^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

que es el intervalo de confianza del $(1-\alpha)\%$ para la variancia de la pobla-- ción.

Ejemplo: Encontrar el intervalo de confianza del 95% para la desviación - estandar de la población si en una muestra de tamaño 25 se obtuvo una desviación estandar $S_x = 10$

$n = 25$

$S_x = 10$

$n-1 = 24$

de las tablas $\chi^2_{24,0.977} = 39.4$

$\chi^2_{24,0.025} = 12.4$

$$\sqrt{\frac{25(10)^2}{39.4}} \leq \sigma_x \leq \sqrt{\frac{25(10)^2}{12.4}}$$

$$8 \leq \sigma_x \leq 4 .$$

Nota: Si se trata de muestras grandes el intervalo se cierra considerablemente por lo que en ese caso se puede aceptar que

$$\sigma_x^2 = S_x^2$$

ESTIMACIONES DE MAXIMA VEROSIMILITUD

La mayor ventaja de este tipo de estimadores es que sus propiedades han sido - ampliamente estudiadas independientemente de las distribuciones de probabilidad de las variables aleatorias que dan lugar a dichos estimadores.

Este método en esencia recomienda a las estadísticas que estiman a los parametros para los cuales la muestra extraída tiene mayor probabilidad de ser elegida.

En efecto: sea la muestra de elementos independientes $X_1, X_2, \ldots X_n$ para los cuales se puede definir una distribución conjunta de probabilidad (función de verosimilitud), en función de uno o varios parámetros (condicionada a dichos parámetros)

$$f_{x_1,x_2,\ldots,x_n}(x_1,x_2,\ldots,x_n|\theta_1,\theta_2,\ldots,\theta_k) = f(x_1|\theta_1,\ldots)\,f(x_2|\theta_1,\ldots)\ \ldots\ f(x_n|\theta_1,\ldots)$$

$$= \prod_{i=1}^{n} f_{x_i}(x_i|\theta_1,\theta_2,\ldots,\theta_k) = V(x_1,x_2,\ldots,x_n|\theta_1,\theta_2,\ldots,\theta_k)$$

(función de verosimilitud)

Después los valores de $\theta_1, \theta_2, \ldots \theta_k$ se obtienen haciendo maxima la función de verosimilitud (por ejemplo, igualando a cero las derivadas parciales)

De esta manera se puede demostrar que $\bar{x}$ y S_x^2 son estimadores de maxima verosimilitud

Ejemplo

ESTIMACION DE MAXIMA VEROSIMILITUD DEL PARAMETRO λ DE UNA DISTRIBUCION DE POISSON

$$p_x(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

sea una muestra de tamaño 8

$$V(x_1,x_2,\ldots,x_8|\lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \ldots \frac{\lambda^{x_8}}{x_n!} e^{-\lambda} = e^{-8\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1!\,x_2!\ldots x_n!}$$

Tomando logaritmos naturales de la función de verosimilitud

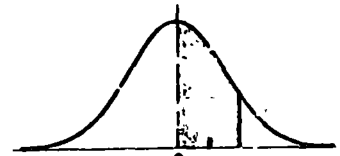
$$LV = -8\lambda + \sum_{i=1}^{n} x_i L\lambda - L[x_1!\,x_2!\ldots x_8!]$$

derivando la expresión anterior e igualandola a cero para obtener el máximo

$$\frac{dLV}{d\lambda} = -8 + \frac{\sum x_i}{\lambda} = 0 \quad \therefore \quad \boxed{\hat{\lambda} = \frac{\sum x_i}{8}}$$

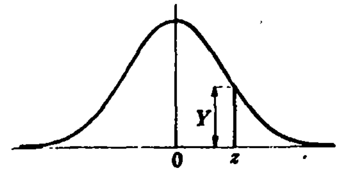
que es el estimador de maxima verosimilitud

AREAS BAJO LA CURVA NORMAL ESTANDARIZADA



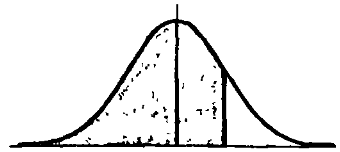
z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

ORDENADAS de la CURVA NORMAL ESTANDARIZADA



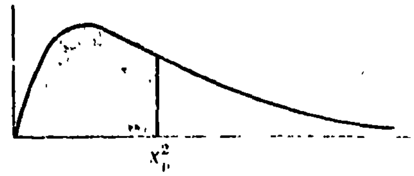
z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

PERCENTILES PARA LA DISTRIBUCION $\underline{t}$ DE STUDENT, CON ν GRADOS DE LIBERTAD



ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

PERCENTILES PARA JI CUADRADA, CON ν GRADOS DE LIBERTAD



ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	45.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.6
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

TEORIA DE LA DECISION

Enrique de la Mora A. Sept. 72

Análisis matematico
Toma de decisiones

- el estado del mundo es inciero
- se puede obtener información adicional sobre dicho estado mediante la experimentación

Objetivos → identificación de un camino de acción

↓

modelos

- consecuencias
- procesos de decisión
- proceso probabilístico: incertidumbre

procedimientos de la estadística + opiniones y sentimientos intuitivos de quien toma las decisiones

→ para considerar formalmente toda la información disponible dentro de un proceso racional y consistente de la toma de decisiones

Conceptos Básicos

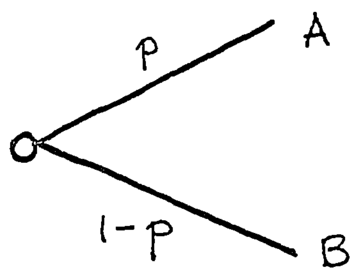
— Probabilidad

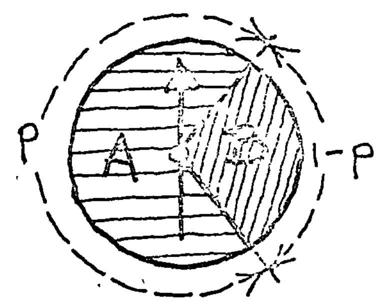
Sean E y F eventos que podrían resultar de una decisión

1) $P\{E\} \geq 0$ para cualquier E
2) $P\{E\} = 1$ si E ocurrirá con certeza
3) $P\{E \text{ ó } F\} = P\{E\} + P\{F\}$

   si E y F son mútuamente exclusivos

— Loterías





$(p, A; 1-p, B)$

Pago esperado de una lotería:

$$C = pA + (1-p)B$$

Conceptos Básicos

- Probabilidad

Eventos independientes: si E y F son eventos independientes, entonces:

$$P\{EF\} = P\{E\}P\{F\}$$

Probabilidad condicional:

$P\{E|F\}$ se lee: probabilidad de que suceda el evento E, dado que sucedió el evento F

Por ejemplo; si E es el evento de que en la estación de servicio den un diagnóstico de "motor bueno" y F es el evento de que el motor esté bien, entonces:

$P\{E|F\}$ es la probabilidad de que en la estación de servicio digan que el motor está bien dado que el motor está bien.

Se define: $$P\{E|F\} = \frac{P\{EF\}}{P\{F\}}$$

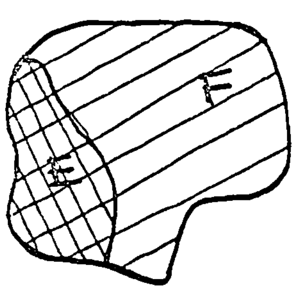


Diagrama de Venn

Problema Ilustrativo

— He decidido comprar un coche
en la agencia el precio és de 4200

— Alternativas
En el periódico anuncian uno igual con poco uso en 2700
Al visitarlo veo que está en perfectas condiciones pero tengo dudas del estado del motor: pienso que la probabilidad de que esté malo es de 0.3
Cambiar el motor cuesta 2400

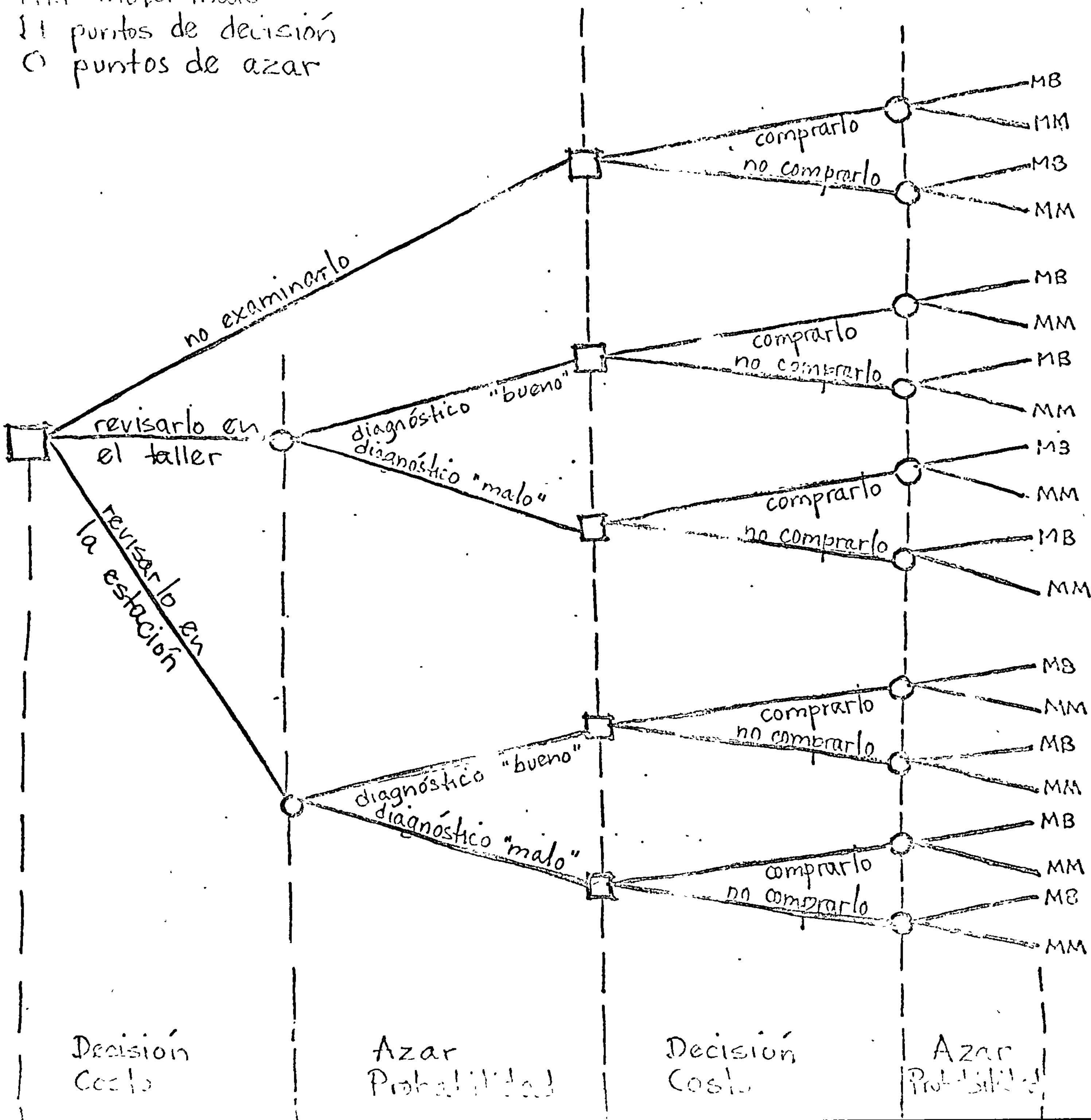
Para asegurarme del estado del motor puedo:
a) mandarlo revisar en un taller en el que me dirán con toda certeza el estado del motor con un costo de 200

b) mandarlo revisar en una estación de servicio en la que solo me costaría 120. Sin embargo, en dicha estación se equivocan el 10% de las veces.

— ¿ Que hacer ?

MB · motor bueno
MM motor malo
□ puntos de decisión
○ puntos de azar

maximizando ahorros respecto a la compra del coche nuevo



Continuación del problema (revisar coche en la estación)

Nomenclatura: MB = motor bueno
MM = motor malo
b = diagnóstico "motor bueno"
m = diagnóstico "motor malo"

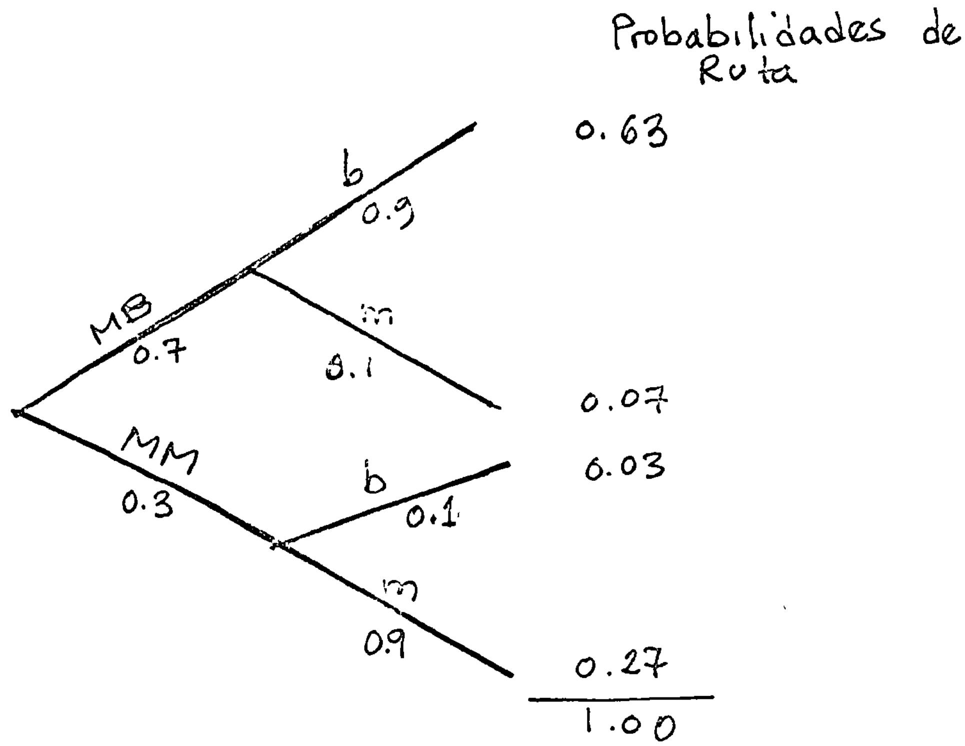
Se dispone de la siguiente información:

a) mi propio juicio de que: $P\{MB\} = 0.7$
$P\{MM\} = 0.3$

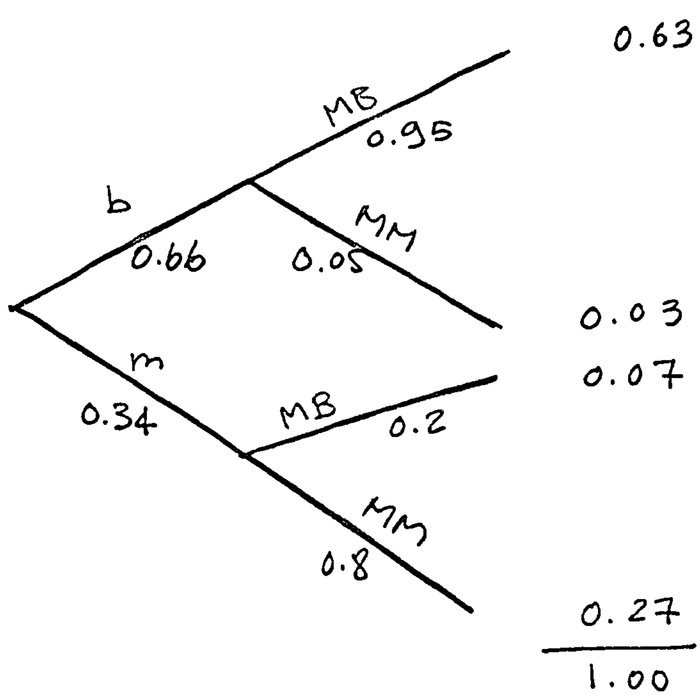
b) conozco el hecho de que en la estación de servicio se equivocan el 10% de las veces:

$$P\{b|MB\} = 0.9$$
$$P\{m|MM\} = 0.9$$
$$P\{b|MM\} = 0.1$$
$$P\{m|MB\} = 0.1$$

se conoce



se desea conocer

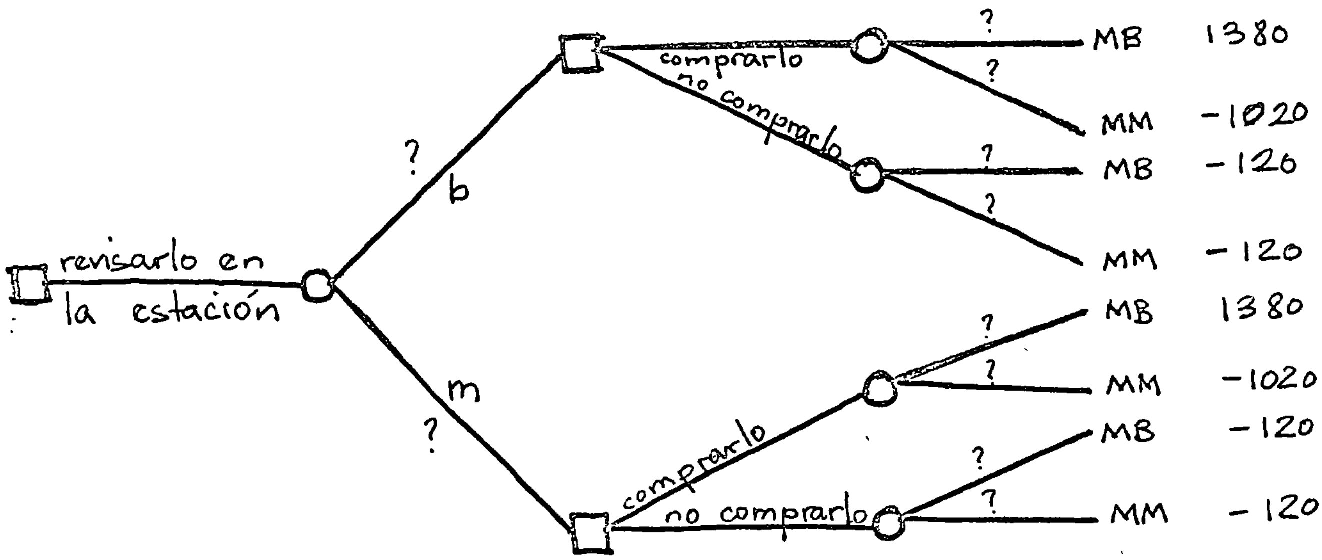


$P\{b\} = 0.63 + 0.03 = 0.66$

$P\{m\} = 0.27 + 0.07 = 0.34$

$P\{MB | b\} = \dfrac{0.63}{0.66} \doteq 0.95$

etc.



Para la obtención de los pagos esperados correspondientes a la decisión de mandar el coche a revisión en la estación de servicio deseo conocer:

a) La probabilidad de que en la estación den un diagnóstico de "motor bueno" o de "motor malo"

$$
\begin{aligned}
P\{b\} &= P\{b, MB\} + P\{b, MM\} \\
&= P\{b|MB\}P\{MB\} + P\{b|MM\}P\{MM\} \\
&= 0.9 \times 0.7 + 0.1 \times 0.3 \\
&= 0.63 + 0.03
\end{aligned}
$$

$\longrightarrow$ $P\{b\} = 0.66$

$$
\begin{aligned}
P\{m\} &= P\{m, MB\} + P\{m, MM\} \\
&= P\{m|MB\}P\{MB\} + P\{m|MM\}P\{MM\} \\
&= 0.1 \times 0.7 + 0.9 \times 0.3 \\
&= 0.07 + 0.27
\end{aligned}
$$

$\longrightarrow$ $P\{m\} = 0.34$

b) La probabilidad de que el motor esté bién dado que en la estación dijeron que está bién:

$$P\{MB|b\} = \frac{P\{MB,b\}}{P\{b\}} = \frac{P\{b,MB\}}{P\{b\}} =$$

$$= \frac{P\{b|MB\}P\{MB\}}{P\{b\}} =$$

$$= \frac{0.9 \times 0.7}{0.66} = \frac{0.63}{0.66} = \frac{21}{22} \doteq 0.95$$

La probabilidad de que el motor esté mal dado que dijeron que está bien:

$$P\{MM|b\} = \frac{P\{MM,b\}}{P\{b\}} = \frac{P\{b|MM\}P\{MM\}}{P\{b\}} =$$

$$= \frac{0.1 \times 0.3}{0.66} = \frac{0.03}{0.66} = \frac{1}{22} \doteq 0.05$$

Además:

$$P\{MB|m\} = \frac{P\{MB,m\}}{P\{m\}} = \frac{P\{m|MB\}P\{MB\}}{P\{m\}} =$$

$$= \frac{0.1 \times 0.7}{0.34} = \frac{0.07}{0.34} = \frac{7}{34} \doteq 0.2$$

$$P\{MM|m\} = \frac{P\{MM,m\}}{P\{m\}} = \frac{P\{m|MM\}P\{MM\}}{P\{m\}} =$$

$$= \frac{0.9 \times 0.3}{0.34} = \frac{0.27}{0.34} = \frac{27}{34} \doteq 0.8$$

Recapitulación : antes y después : el teorema de BAYES

	ANTES	DESPUES	
		b	m
$P\{MB\}$	0.7	0.95	0.2
$P\{MM\}$	0.3	0.05	0.8

	Evento	Apriori	Verosimilitud	Prob. Conjuntas	Aposteriori
	a_i	$P_1\{a_i\}$	$P\{l\mid a_i\}$	$P\{l,a_i\}=P\{l\mid a_i\}P\{a_i\}$	$P_2\{a_i\}$
$l=b$	MB	0.7	0.9	0.63	$0.63/0.66 \doteq 0.95$
	MM	0.3	0.1	0.03	$0.03/0.66 \doteq 0.05$
Σ		1.0	1.0	$P\{l=b\}=$ 0.66	1.00
$l=m$	MB	0.7	0.1	0.07	$0.07/0.34 \doteq 0.2$
	MM	0.3	0.9	0.27	$0.27/0.34 \doteq 0.8$
Σ		1.0	1.0	$P\{l=m\}=$ 0.34	1.0

$$* \longrightarrow \quad P_2\{a_i\}=P\{a_i\mid l\}=\frac{P\{l\mid a_i\}\,P\{a_i\}}{\sum_j \{l\mid a_j\}\,P\{a_j\}}$$

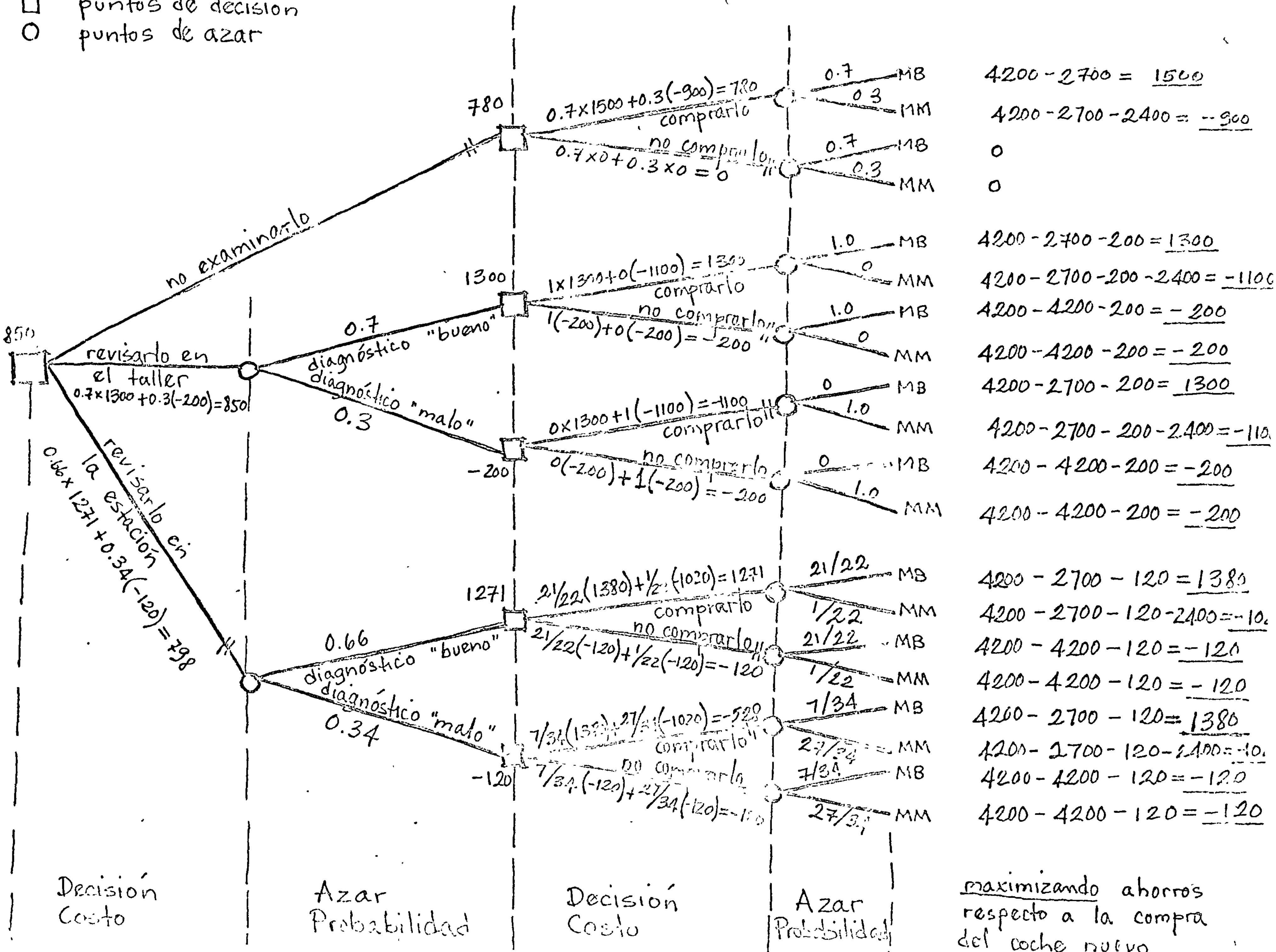
Para $a_i = MB$ y $l=b$:

$$P_2\{MB\}=P\{MB\mid b\}=\frac{P\{MB,b\}}{P\{b\}}=\frac{P\{MB,b\}}{P\{MB,b\}+P\{MM,b\}}=$$

$$=\frac{P\{b\mid MB\}P\{MB\}}{P\{b\mid MB\}P\{MB\}+P\{b\mid MM\}P\{MM\}}=$$

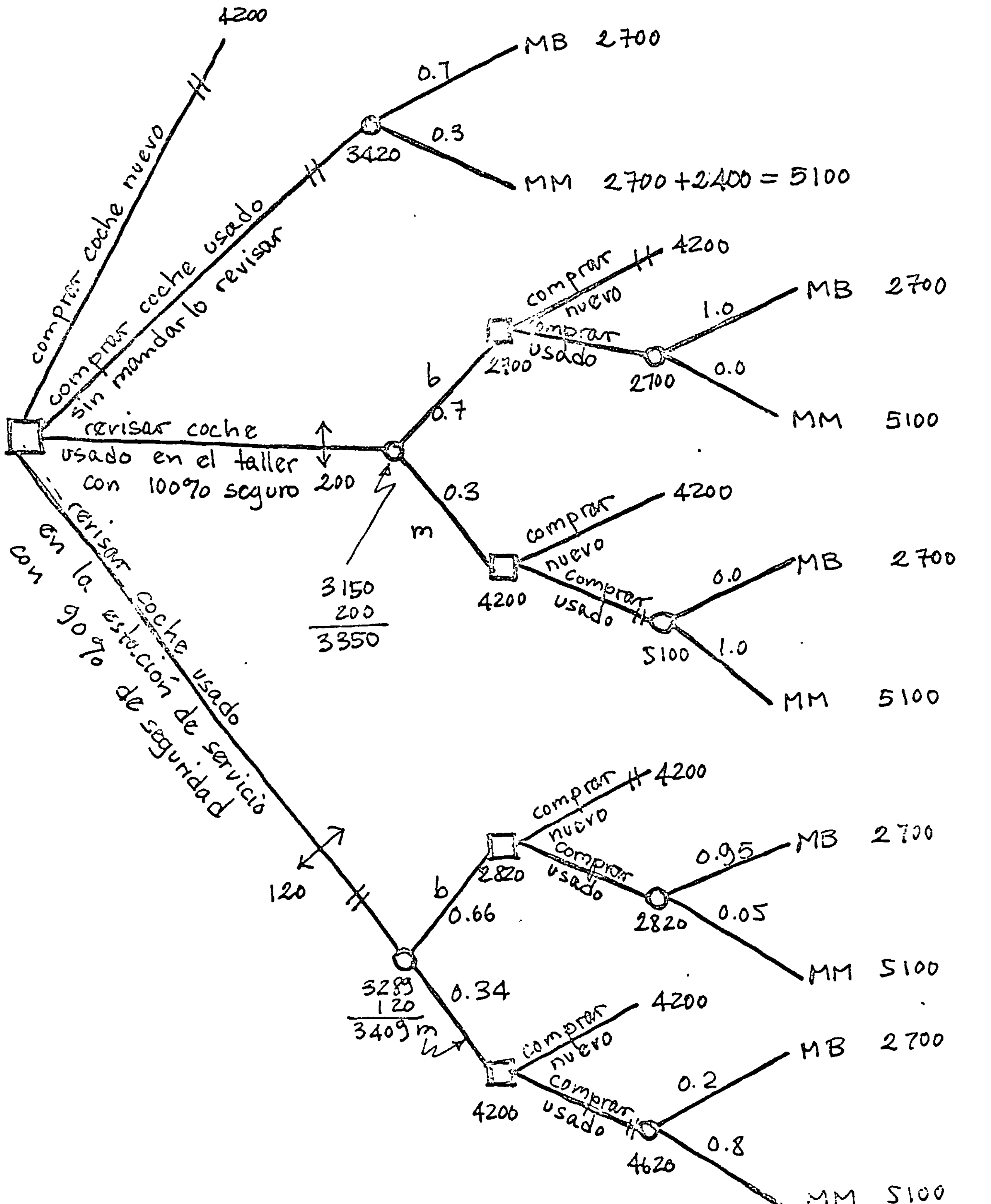
$$=\frac{0.9\times0.7}{0.9\times0.7+0.1\times0.3}=\frac{0.63}{0.63+0.03}=\frac{0.63}{0.66}\doteq 0.95$$

MB: motor bueno
MM: motor malo
□ puntos de decisión
O puntos de azar



Minimizando
costo total

Ahorros: 4200 − 4200 = 0
4200 − 3420 = 780
4200 − 3350 = 850 ←
4200 − 3409 = 791



TEORIA DE LA UTILIDAD

Enrique de la Mora A. Sept. 72

— Axiomas de la Teoría de la Utilidad :

Supóngase que A, B y C son premios o resultados de una decisión.

Notación : $A > B$: A es preferido a B
$A \sim B$: A es indiferente a B

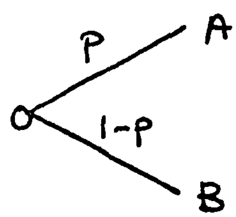
Axioma 1 :

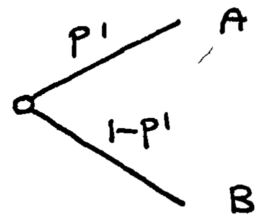
Las preferencias se pueden obtener tanto para premios como para loterías : las preferencias son transitivas :

si $A > B$ y $B > C$ entonces $A > C$
si $A \sim B$ y $B \sim C$ entonces $A \sim C$

Axioma 2 :

Si $A > B$, entonces $(P, A ; 1-P, B) > (P', A ; 1-P', B)$ si y solo si $P > P'$

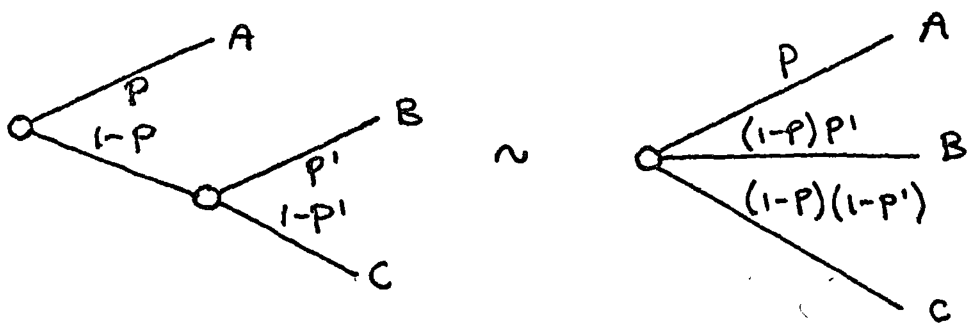




Axioma 3:

No hay un premio intrínseco en el juego:

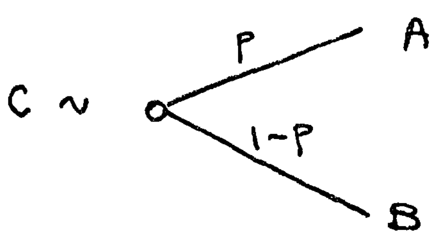
$$(p, A; 1-p, (p', B; 1-p', C)) \sim (p, A; (1-p)p', B; (1-p)(1-p'), C)$$



lotería compuesta   lotería simple

Axioma 4:

Si $A > C > B$, existe una probabilidad p: $0 < p < 1$ de manera que $C \sim (p, A; 1-p, B)$



– Función de utilidad

Existe una función de utilidad $u(\ )$ que asigna un número a cada premio, con las siguientes características:

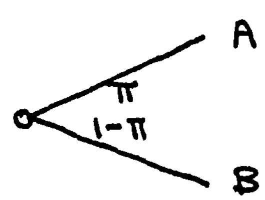
a) $u(A) > u(B)$ si y solo si $A > B$

b) si $C \sim (P, A; 1-P, B)$, entonces

$$u(C) = Pu(A) + (1-P)u(B)$$

– Obtención de la función de utilidad.

Considere un boleto $\boxed{\pi}$ donde $0 \leq \pi \leq 1$ que dá acceso a una lotería del tipo



Donde A y B son premios perfectamente definidos y A es cláramente preferible a B.

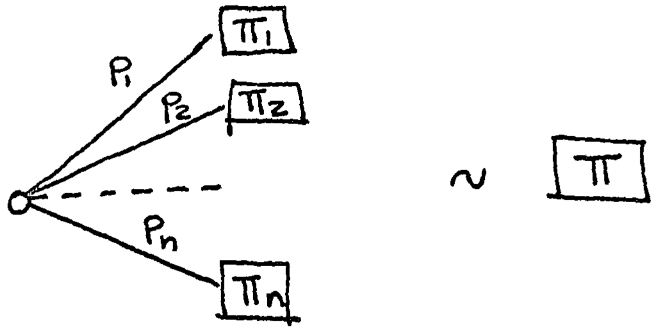
Se puede demostrar (ref. 1) que cualquier tomador de decisiones es indiferente entre una lotería con las siguientes características:

$$P\{\text{tener un boleto } \boxed{\pi_1}\} = P_1$$
$$P\{\text{tener un boleto } \boxed{\pi_2}\} = P_2$$
$$\vdots$$
$$P\{\text{tener un boleto } \boxed{\pi_n}\} = P_n$$

y el tener un solo boleto $\boxed{\pi}$ donde:

$$\pi = p_1\pi_1 + p_2\pi_2 + \cdots + p_n\pi_n$$

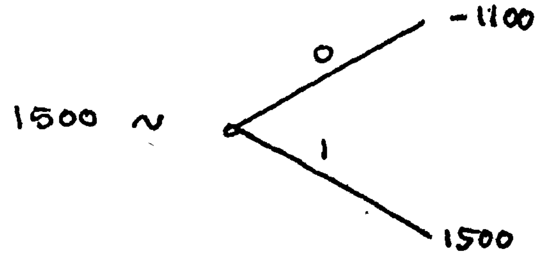
Esquemáticamente:

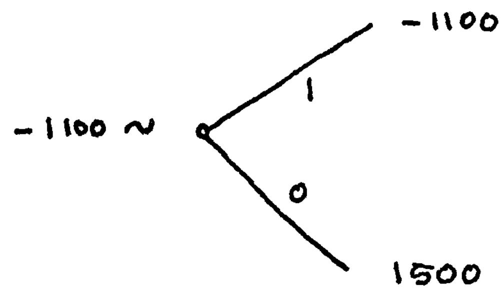


En referencia al problema de la clase anterior, considere Ud. a un tomador de decisiones que tiene aversión al riesgo. Para construir su función de utilidad, lo primero que debe hacer es escoger dos valores monetarios, por ejemplo -1100 y 1500 que cubran el rango completo de sus premios y pérdidas. Se asigna $A = 1500$
$B = -1100$

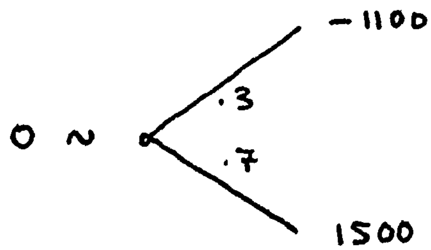
A continuación se pregunta y define la cantidad \$ x que le gustaría tener con certeza en vez de un boleto $\boxed{\pi}$. Sus respuestas crean una correspondencia entre x y π que puede ser graficada en un plano (x, π).

Evidentemente :

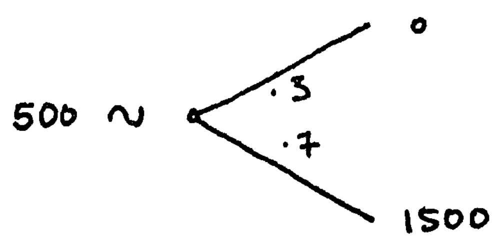




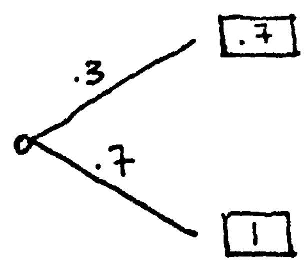
Suponga Ud. que :



y que

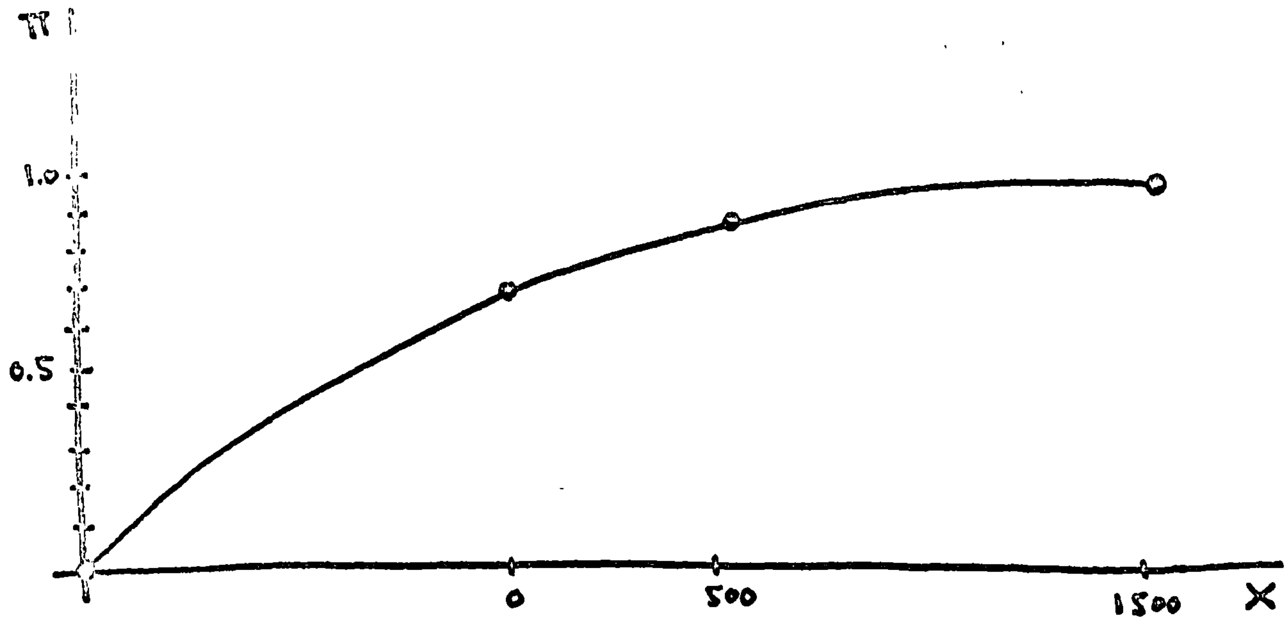


Entonces :



$0.21 + 0.7 = 0.91$

Se tienen cuatro coordenadas (x, π) de la función de utilidad: (1500, 1.0), (-1100, 0.0), (0, 0.7) y (500, 0.9). La función tiene las propiedades de continua y monotónica creciente.



De la gráfica anterior se pueden obtener, para cada x, los valores de π con los que se puede analizar el árbol de decisiones tomando en cuenta la aversión al riesgo particular de la persona que toma las decisiones.

Esto se muestra en la siguiente figura en la que se considera que se tienen, de la gráfica anterior, las siguientes coordenadas: (1500, 1), (-900, .18), (0, 0.7), (1300, 0.98), (-1100, 0), (-200, 0.58), (1380, 0.99), (-1020, 0.07), (-120, 0.62), (100, 0.75), (350, 0.86), (300, 0.83).

Solución del problema de tarea:

Inicialmente se sabe que:

$$P\{A\} = 0.5$$
$$P\{B\} = 0.5$$

Sea C el evento de sacar de la caja 8 blancas y 4 negras. en cualquier orden. Se puede calcular:

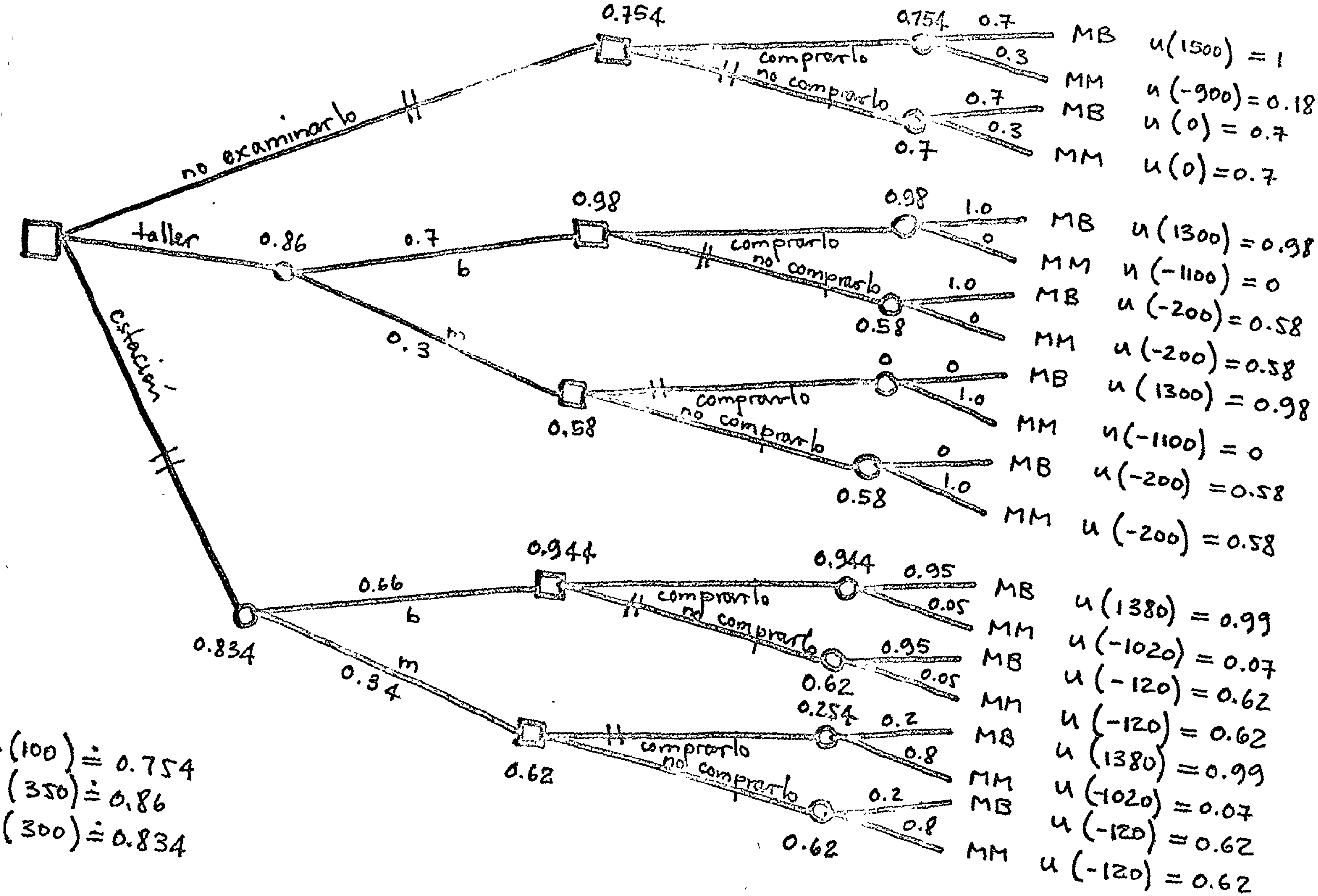
$$P\{C \mid A\} = (0.7)^8 (0.3)^4$$
$$P\{C \mid B\} = (0.7)^4 (0.3)^8$$

Por el teorema de Bayes:

$$P\{A \mid C\} = \frac{P\{C|A\} P\{A\}}{P\{C|B\} P\{B\} + P\{C|A\} P\{A\}}$$

$$= \frac{(0.7)^8 (0.3)^4 (0.5)}{(.7)^8 (.3)^4 (.5) + (.3)^8 (.7)^4 (.5)} = 0.967$$

Arbol de Decisiones
con aversión al riesgo

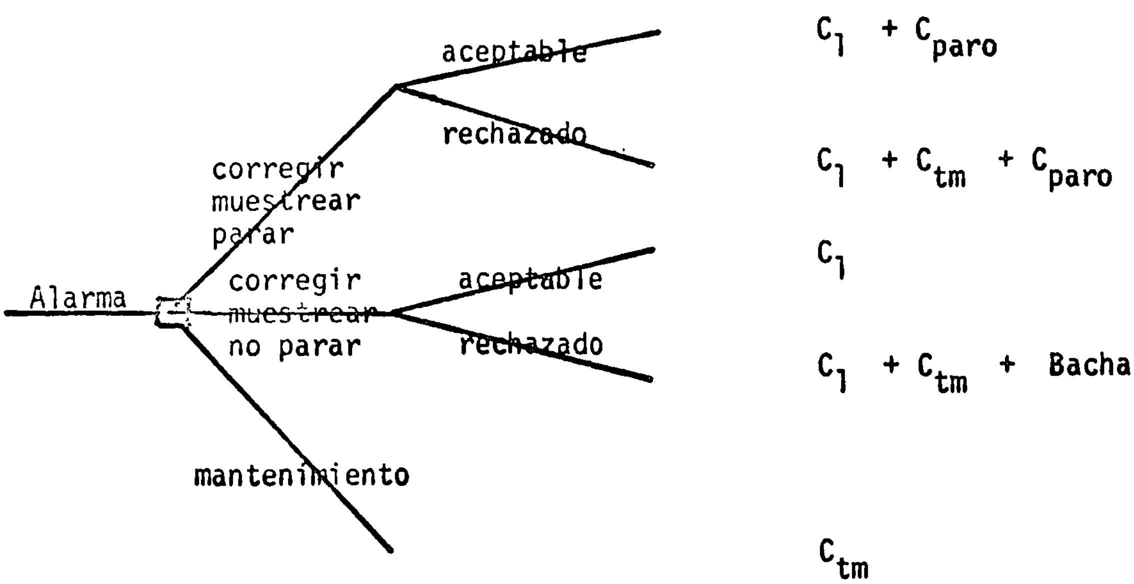


$u(100) \doteq 0.754$
$u(350) \doteq 0.86$
$u(300) \doteq 0.834$

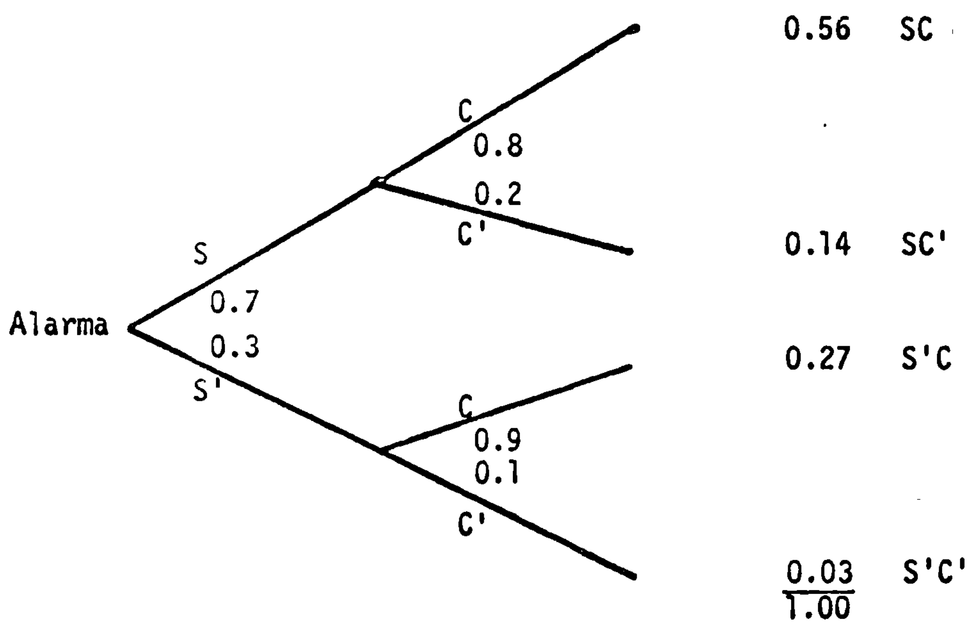
producir material aceptable.

Se busca establecer la regla de decisión de mínimo costo.

ARBOL DE DECISION:



ARBOL DE PROBABILIDAD



$P(SC) = P(C/S)\,P(S)$

$P(C) = 0.83$

$P(C') = 0.17$

9

4.- Inspección y Mantenimiento.

El "Análisis de Decisiones" permite establecer políticas de Inspección-Mantenimiento.

Ejemplo:

En un proceso de producción se tienen detectores de desajustes. La confiabilidad del detector es tal que al dar alarma del síntoma S la probabilidad de que sea falsa es 0.3.

Además de no hacer nada después del alarma el operador tiene tres cursos de acción a seguir:

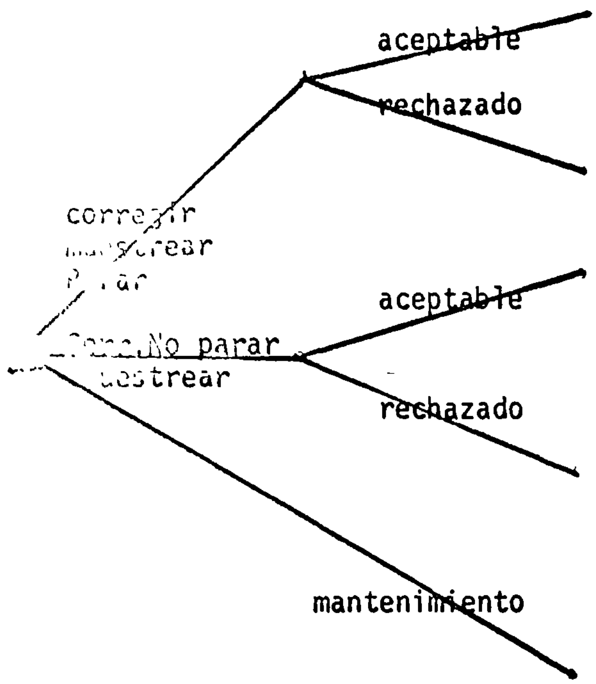
1) Hacer un ajuste menor a costo C_1, tomar una muestra y continuar la producción mientras se prueba la muestra.

2) Hacer el ajuste menor y parar mientras se hace la prueba.

3) Dar mantenimiento con tiempo de servicio aleatorio y costo C_{tm}.

La naturaleza del ajuste es tal que si la alarma fué falsa el proceso seguirá produciendo material aceptable con probabilidad 0.9, y si la alarma fué acertada el ajuste tiene sólo 0.8 de probabilidad de corregir el proceso y

ARBOL DE DECISION



	Probabilidad	Costo Esperado
aceptable	0.83	(50+200)x.83 = 208
rechazado	0.17	(50+400+200)x.17=110
		318
aceptable	0.83	(50)x.83 = 41.5
rechazado	0.17	(50+400+1000)x.17=246.5
		288.0
mantenimiento		400
		400

Se tiene:

$C_1 = 50$

$C_{tm} = 400$

$C_{paro} = 200$

Bacha = 1000

La alternativa de decisión de mínimo costo es: hacer ajuste menor y no parar durante la prueba.

Costo esperado = 288.0 + 41.5

BIBLIOGRAFIA

Teoría de decisiones y teoría de la utilidad:

1) Decision Analysis - Howard Raiffa - Addison Wesley
2) An Introduction to Bayesian Statistical Decision Processes - Bruce W. Morgan - Prentice Hall

Aplicaciones a la ingeniería:

3) Decision Analysis for Engineers - Benjamin

Teoría de juegos y otros temas:

4) Introduction to Operations Research - Hillier and Lieberman

En español, diferentes temas:

5) Analisis de Sistemas - IMPOS

EL COSTO EN LAS INVESTIGACIONES MUESTRALES

I. CONSIDERACIONES PRELIMINARES

La sorprendente popularidad que ha logrado últimamente el muestreo estadístico, implica serios riesgos. Por una parte se está llegando a una aplicación indiscriminada de las técnicas, cualquiera que sea la investigación, y se atribuye a los resultados una injustificada validez. Por otra parte, los presupuestos asignados a las investigaciones por muestreo son, en general exiguos e insuficientes. En no pocas investigaciones, se decide utilizar el muestreo sólo porque implica un costo substancialmente menor. En general, esto es cierto, pero no como para verificar una correspondencia proporcional entre tamaño de población y costo del censo, a tamaño de muestra y costo de la muestra. Se justifica esta preocupación cuando se piensa que muchas investigaciones de muestreo no deberían llevarse a cabo, por no obedecer a los requisitos mínimos de planificación y análisis. Lo que es peor, muchas instituciones y centros de investigación utilizan estimaciones ajenas que pueden resultar muy lejanas a la realidad.

A veces se utilizan estimaciones obtenidas por muestreo, que si bien en un tiempo pudieron tener una validez aceptable, ya han perdido actualidad. Esas estimaciones han sido y seguirán siendo válidas para la época en que fueron diseñadas y realizadas, pero sólo la periodicidad en su obtención permite obtener ventajas auténticas, y garantizar resultados correctos.

Cuando se discuten alternativas de investigación, no sólo debe atenderse al aspecto de los costos sino tambien de la oportunidad y rapidez en la obtención de resultados, precisión, posibilidades concretas de obtención de la información, etc.

Ademas, es necesario recordar que, en general, el costo por unidad investigada, es mayor en una muestra que en un censo, ya que la preparación del diseño muestral se suele dedicar mucho mayor esfuerzo y tiempo que en la preparación de un censo, en términos relativos.

II. ELEMENTOS DEL COSTO

Es importante identificar las etapas que se deben cumplir en una investigación muestral, a fin de asociarles los desembolsos directos que origina cada una de ellas. Una clasificación detallada, es la siguiente:

1. Fijación de objetivos
2. Determinación del marco de la muestra
3. Elección del diseño de la muestra
4. Recopilación de otros antecedentes relacionados con la investigación
5. Determinación del tamaño de muestra
6. Primera estimación del costo probable

..

7. Selección de las unidades de muestreo
8. Establecimiento de controles previos
9. Entrenamiento de enumeradores
10. Promoción entre la población informante
11. Diseño e impresión de cuestionarios
12. Organización de trabajo en el terreno
13. Control de la calidad de la información colectada
14. Sistematización de los datos obtenidos
15. Análisis de las informaciones y estimaciones
16. Publicación de los resultados
17. Determinación del costo efectivo

En algunas etapas, no habrá dificultad en asignar directamente la parte del costo total que le corresponda; en otras será necesario proceder a un prorrateo que garantice una distribución equitativa de los costos. La ventaja que puede lograrse de una correcta asignación y análisis de costos, es la de disponer de mejores estimaciones de los costos de investigaciones futuras; además se tendrán ideas claras del costo de cada etapa y de la necesidad de obtener el máximo provecho de los desembolsos ocasionados. Los presupuestos que se elaboren en base a esas estimaciones anteriores, deberán ser contrastados posteriormente con los costos efectivos. En lo posible, se deben establecer las diferencias entre lo presupuestado y la realidad, para cada una de las etapas, y tambien analizar las causas de esas diferencias a fin de obtener pautas que permitan el máximo rendimiento futuro de los recursos invertidos.

Como se verá más adelante, es necesario determinar cuales etapas y qué conceptos dentro de cada una, implican costos fijos y cuáles costos variables. Esto dará mayor flexibilidad para manejar alternativas de tamaño de muestra, en comparación a los mayores gastos que originasen.

No hay que olvidar, que además del costo monetario, se precisa tener ideas de otros conceptos relacionados, tales como el tiempo que mediará entre la iniciación de la investigación y la entrega de resultados, los rangos de tamaños de muestra entre los que no hay variaciones de costos fijos, los costos adicionales por imprevisiones, etc.

Para fines de ilustración, se presenta una tabla simplificada, que facilita la toma de decisiones en circunstancias semejantes:

CUADRO DE ANALISIS

Tamaño de muestra	Desviación máxima como % de la $\bar{x}$ poblacional	Costo Fijo estimado	Costo Variable por unidad	Costo total	Duración de la investigación (en días)
800	15 %	2000	10	10000	4
1200	13 %	2000	10	14000	6
2000	10 %	2000	10	22000	8
4000	8 %	3000	11	47000	12
7000	5 %	3000	11	80000	15
8000	3 %	3000	11	91000	17
35000	0 %	10000	12	430000	50

..

Los datos del cuadro, no obedecen a cálculos estrictos; se trata simplemente de aproximaciones ilustrativas.

El tamaño de la "Muestra" mayor puede corresponder al de un censo.

III. COSTOS VERSUS PRECISION

Desde otro punto de vista, es factible distinguir dos categorías dentro del costo total: el costo fijo, invariante ante cualquier tamaño de muestra, dentro de ciertos márgenes razonables, y el costo variable, dependiente directamente del número de unidades de la muestra. Se tiene entonces la relación aditiva

Costo total = Costo fijo + Costo variable,

o en símbolos:

$$C = C_o + C_v$$

Dado que C_v, depende del número de unidades de la muestra, se puede expresar lo anterior, en la forma:

$$C = C_o + n_1 \, C_1 \qquad (1)$$

en que

n_1: es el tamaño de la muestra, y

C_1: es el costo medio variable

Despejando n_1, se tiene

$$n_1 = \frac{C - C_o}{C_1}$$

Esto significa que se ha determinado el tamaño de la muestra, atendiendo al presupuesto disponible para la investigación, (C), y los costos fijos y variables (C_o y C_1).

Por otra parte, en el diseño aleatorio simple se utiliza la siguiente fórmula de aproximación del tamaño de muestra:

$$n_o = \frac{t^2 \, S^2}{d^2} \qquad (2)$$

en que t es el coeficiente de confianza (determinado por el nivel de confianza), d el semiancho del intervalo de confianza, y S^2 la varianza en la población*:

$$S^2 = \frac{\sum_{i=1} (Y_i - \bar{Y})^2}{N - 1}$$

Por lo general, esta fórmula del tamaño de muestra proporciona un valor distinto del que resulta de la fórmula de costos, ya que es

*Fórmula de Cochran

ta implica especificaciones técnicas de muestreo, y aquella se refiere a las disponibilidades monetarias. En algunas ocasiones cuando se asigna un presupuesto excesivo, el tamaño de muestra dado por la fórmula de costos será mayor que el dado por las especificaciones de precisión y confianza; pero en general ocurrirá lo contrario. En ambos casos es preciso elegir entre alternativas:

a) Caso en que el $n_1 < n_o$

Dada la escasez de recursos, los presupuestos para investigaciones por muestreo son exiguos. Además los costos de realizar muestreos esporádicos suelen ser altos. Este es el caso más frecuente, en que las disponibilidades de presupuesto determinan un tamaño de muestra insuficiente para cumplir con las especificaciones técnicas. Por lo tanto, será necesario analizar cuales son las especificaciones que se satisfacen con este tamaño de muestra n_1.

Observando la fórmula:

$$n_o = \frac{t^2 \; S^2}{d^2}$$

se tiene que:

a) S^2 es un parámetro poblacional; b) t^2 es un coeficiente cuya variación es muy restringida, ya que en general se eligen valores entre 1,6 y 2,5 si se pretenden niveles de confianza superiores al 90%. Por consiguiente el único elemento susceptible de cierta variabilidad, es $\underline{d}$, el semiancho del intervalo de confianza.

Esto nos lleva a calcular cual es el valor $\underline{d}$ que resultaría de aplicar en esa investigación el tamaño de muestra n_1. Procedemos a igualar ambas expresiones, como sigue:

$$\frac{C - C_o}{C_1} = \frac{t^2 \; . \; S^2}{d^2}$$

de donde, despejando el valor de $\underline{d}$:

$$d_1 = t \; S \; . \left(\frac{C_1}{C - C_o} \right)^{1/2} = t \; S \; . \left(\frac{C_1}{C_v} \right)^{1/2}$$

Este valor d_1 será mayor que el valor $\underline{d}$ obtenido anteriormente, por lo que habrá que considerar si el incremento $d_1 - d$ es significativo. Puede ser que las estimaciones resultantes pierdan utilidad práctica, porque su rango de variabilidad es muy grande, o alternativamente que ese valor no implique perder precisión y por consiguiente que la investigación puede llevarse a efecto sin mucho riesgo.

En caso de que d_1 sea considerablemente mayor que $\underline{d}$, será necesario que se destine un mayor presupuesto a la investigación. El costo total sería

$$C = C_o + n_o C_1$$

en que n_o se obtiene de la fórmula (1) y se admite el supuesto de que los costos fijos no cambian, aunque ha aumentado el valor de n . Si fuera muy dificultoso obtener una mayor disponibilidad presupuestaria, habrá que pensar en cambiar el diseño muestral por otro que resulte más económico para esa investigación. Por último, si esto tampoco fuera posible, será preferible no realizar la investigación por muestreo.

b) Caso en el que $n_1 > n_o$

Esta situación se comenta sólo como ilustración, ya que en la práctica tendrá carácter puramente accidental. Es el caso en que los recursos asignados para la investigación por muestreo, determinan un tamaño de muestra mayor que el que resulta de considerar la precisión y la confianza estipuladas. Aquí se procede a analizar cual sería el valor de d_1, tomando esa muestra mayor, esto es:

$$d_1 = t\,S\left(\frac{C_1}{C - C_o}\right)^{1/2} = t\,S\left(\frac{C_1}{C_v}\right)^{1/2}$$

Como d_1 es menor que d, el problema por considerar es, si la disminución del intervalo justifica el aumento en los costos. Es poco probable que esto ocurra, ya que al fijar el valor de d, habremos considerado el aspecto "precisión" en su justa dimensión. Una mayor precisión, si bien garantiza mayor exactitud, no se justifica a cualquier costo.

Naturalmente, esta consideración de los dos casos se ha hecho desde un punto de vista general, ya que habrá investigaciones en que la precisión sea tan importante, que el aspecto presupuestario sea relativamente secundario. En otras, la rapidez y oportunidad con que se requieren los resultados puede ser el aspecto dominante. Pero en la mayoría de las veces, la relación entre los recursos financieros y los requerimientos técnicos condicionará el desarrollo de la investigación por lo que es aconsejable evaluar la utilidad de la investigación y la validez de sus estimaciones, para compararlas con el costo total que involucran.

IV. CONSIDERACIONES FINALES

El objetivo de estos comentarios es el de divulgar las limitaciones del muestreo y sus alcances reales. La consideración del aspecto financiero tiene su origen en la observación de investigaciones por muestreo que se llevan a efecto con presupuestos insuficientes.

Es usual que al iniciar algún método nuevo de recolección y análisis de datos, se haga énfasis en sus ventajas. Pero el mal uso que se hace de la técnica, mueve a plantear las limitaciones y riesgos que acarrea una indiscriminada utilización de métodos, para los que se precisa un cabal conocimiento de la teoría y disponibilidad de recursos adecuados.

Cabe indicar que se han abordado estos aspectos, desde el punto de vista del muestreo aleatorio simple, porque este constituye la base de los diseños aleatorios, y su discusión tiene carácter general. Para otros diseños muestrales aleatorios, las consideraciones son similares y pueden llevarse a efecto sin que varíen notablemente las conclusiones presentadas.

ARIEL KLEIMAN

'cgs.

GLOSARIO

El presente *Glosario* se ha elaborado como una guía práctica para quienes no están familiarizados con los términos estadísticos. No pretende usar definiciones que sean lo suficientemente precisas para satisfacer a los estadígrafos profesionales. Aquellas personas que se interesen por conocer dichas definiciones pueden consultar el *Dictionary of Statistical Terms*, de Kendall y Buckland, Nueva York, Hafner Publishing Company, 1960.

A fin de facilitar al lector la comprensión de las definiciones, en éstas aparecen en letra cursiva y con inicial mayúscula aquellos términos que, a su vez, se encuentran definidos en el *Glosario*.*

Afijación óptima*	El uso de fórmulas matemáticas para determinar la mejor distribución de la muestra entre las diversas clases de un *Universo* estratificado.
Amplitud	La diferencia entre el valor máximo y el mínimo de una variable determinada en un grupo de observaciones. Se conoce también con el nombre de recorrido.
Amplitud de la precisión	Véase *Precisión*.
Amplitud media	El promedio aritmético de dos o más amplitudes. Véase *Amplitud*.
Atributo	La característica cualitativa de un elemento de una *Población*. Por ejemplo, la clasificación de comprobantes según hayan sido debidamente firmados o no. Compárese con *Variable*. *El Muestreo de atributos* es el uso de clasificaciones dicotómicas o binomiales, por ejemplo: "satisfactorio" o "insatisfactorio".
Campo	Véase *Universo*.
Características	Véase *Atributo*.
Censo	Un estudio completo del *Universo*, a diferencia de un estudio parcial del mismo a través de una *Muestra*.
Coeficiente de confianza	Es lo mismo que *Nivel de seguridad*.
Coeficiente de regresión	La estimación del cambio en la variable dependiente producido por la modificación, en una unidad, de la variable independiente.

* En la versión castellana del presente *Glosario*, se procuró adoptar una terminología estadística que fuese comprensible en todos los países de habla hispana. Para ello se consideró apropiado que la traducción se hiciera de acuerdo con la segunda edición del *Vocabulario Estadístico* elaborado por el Instituto Interamericano de Estadística y publicado en 1960 por la Organización de los Estados Americanos. El objeto de dicho vocabulario es fomentar el uso de una terminología interamericana uniforme. Por otra parte, aquellos términos que en el presente *Glosario* aparecen marcados con un asterisco son locuciones que tienen variantes terminológicas en países latinoamericanos Por ello se ha elaborado un Anexo al *Glosario*, en el que aparecen dichos términos acompañados de sus equivalentes en diversos países de América Latina.

Coeficiente de variación	La medida de la *Variabilidad* relativa en una *Distribución de frecuencias*. Se obtiene dividiendo la *Desviación estándar* entre la *Media*. Expresa la magnitud de la *Desviación estándar* con respecto a la *Media* de la *Población*. También se le llama *Desviación estándar relativa*.
Conjunto incompleto	Conjunto que no incluye la totalidad de lo que va a medirse. Por ejemplo, un conjunto del que una parte de los registros han sido retirados.
Curva acampanada *	Véase *Curva normal*
Curva normal	La curva, trazada en una gráfica, de los resultados de todas las muestras posibles. Generalmente será aplicable a todas las muestras, salvo las muy pequeñas o las de universos muy irregulares. También se le denomina *Curva acampanada* o Curva de Gauss. Véanse *Distribución de frecuencias* y *Muestreo estratificado*.
Desviación estándar *	Una medida de la variabilidad de una *Distribución de frecuencias*. Mientras mayor sea la variabilidad, mayor será el valor de la desviación estándar. Para obtenerla, se procede de la siguiente manera: se calcula la diferencia entre cada elemento de la *Población* y su *Media*, se eleva al cuadrado cada una de estas diferencias, se suman los cuadrados, se divide luego la suma así obtenida entre el número total de elementos y, finalmente, se extrae la raíz cuadrada de ese cociente.
Desviación estándar relativa	Es lo mismo que *Coeficiente de variación*.
Dígitos decimales aleatorios	Véase *Tablas de números aleatorios*.
Distribución	Véase *Distribución de frecuencias*.
Distribución asimétrica	Véase *Distribución de frecuencias*.
Distribución de frecuencias	La clasificación de datos numéricos de acuerdo con su tamaño o magnitud. Una *Población* cuyos elementos se clasifican de acuerdo con alguna característica cuantitativa (valor monetario de los pedidos, facturas, etc.), podría describirse mediante una distribución de frecuencias. Una distribución simétrica es una distribución de frecuencias que puede representarse por medio de una *Curva normal (acampanada)*. Una distribución asimétrica es una distribución de frecuencias cuya curva se extiende más hacia una dirección que hacia otra. Una distribución asimétrica sería, por ejemplo, aquella que comprendiera 500 facturas, por un total de $ 1 000 000, de las cuales 5 fueran por $ 100 000 cada una. Si se examinasen *todas* las facturas de alto valor —y, por tanto, se les excluyera de la distribución asimétrica—, las facturas restantes, suponiendo que no existieran otros valores extremos, reflejarían una *Distribución normal* y se prestarían para el muestreo aleatorio.
Distribución de Gauss	Véase *Curva normal*.

Distribución normal	Véase *Curva normal.*
Distribución simétrica	Véase *Distribución de frecuencias.*
Enunciado de confiabilidad	El grado en que se controla la diferencia entre el resultado de la muestra y el valor del *Universo* (es decir, la *Variación*). El enunciado de confiabilidad se compone de *1)* el *Nivel de confianza* y *2)* la *Precisión.* La precisión no se puede considerar separadamente del nivel de confianza.
Error*	Una discrepancia que tenga significación en la auditoría y que no haya sido descubierta y corregida en el proceso del sistema de control interno establecido.
Error de muestreo	Error que se produce exclusivamente a consecuencia del uso de muestras. Este tipo de error puede estimarse matemáticamente. Compárese con *Errores ajenos al muestreo.*
Errores ajenos al muestreo	Las equivocaciones cometidas por el auditor al inspeccionar o examinar los elementos de la muestra.
Estadística*	Una medida determinada con base en una muestra. Compárese con *Parámetro.* Una estadística es el valor de la característica en estudio derivado de una muestra.
Estimación de frecuencia	El método utilizado para determinar, por *Muestreo,* la tasa de ocurrencia de ciertos *Atributos* dentro de los límites prescritos de *Precisión* y *Nivel de confianza;* es decir, la determinación, con base en una muestra, de cuán a menudo ha aparecido un atributo específico. Véase *Atributo.*
Estimación del valor en unidades monetarias	La estimación del valor promedio de un grupo de elementos mediante el empleo de una muestra, con una seguridad —equivalente al *Nivel de confianza*— de que la *Media* de dicha muestra estará dentro de ciertos límites, de magnitud especificada, a uno y otro lado del verdadero valor medio que se hubiera obtenido promediando todos los elementos del *Campo.* Puede usarse para estimar el valor monetario de los errores, de la suma total de cuentas por cobrar, etc. El procedimiento de estimación no necesariamente debe servirse de unidades monetarias, puede ser aplicado a un promedio de cualquier valor, como, por ejemplo, la antigüedad promedio de las cuentas por cobrar. Véanse *Variable* y *Estimación de variables.*
Estimación de variables	Método estadístico para estimar valores, sean monetarios o de otra clase. En el muestreo de *Variables,* se miden o evalúan (en dólares, libras, días, etc.) las unidades de muestreo elegidas, y con base en las mediciones realizadas se calcula alguna medida estadística para estimar el valor del *Parámetro* del *Universo.* Véase *Estimación del valor en unidades monetarias.*
Estimación estadística	Estimación de los *Parámetros* (valor verdadero de las características en estudio) del *Universo,* a partir de una muestra estadística.

Estimación por ampliación Véase *Procedimientos de estimación.*

Estimación por diferencia Véase *Procedimientos de estimación.*

Estimación por intervalo La estimación del *Parámetro* de un *Universo* mediante la especificación de una amplitud de valores definida por un límite superior y uno inferior, y dentro de la cual se asegura que se encuentra el verdadero valor del *Parámetro*. Véase *Precisión*. Compárese con *Estimación por puntos.*

Estimación por puntos * La estimación del *Parámetro* de un *Universo*, mediante la asignación del mejor valor singular estimado del parámetro. Compárese con *Estimación por intervalo.*

Estimación por razón Véase *Procedimientos de estimación.*

Estimación por regresión lineal Véase *Procedimientos de estimación.*

Evaluaciones interinas * La evaluación de la confiabilidad del resultado de una muestra en cualquier momento de su recopilación, a condición de que los elementos hayan sido seleccionados por medios aleatorios para someterlos a una auditoría.

Fracción de muestreo Véase *Muestreo estratificado proporcional.*

Generador de números aleatorios Un método para obtener números aleatorios, ya sea en tarjetas perforadas o en cuadros impresos, mediante el procesamiento electrónico de datos (PED). Se han formulado programas para proporcionar números aleatorios dentro de la amplitud especificada por el auditor y para tenerlos clasificados en secuencia ascendente.

Hipótesis Una suposición acerca de una *Población*, suposición que se decide aceptar o rechazar en vista de los resultados observados en la muestra.

Intervalo de confianza La amplitud dentro de la cual se espera encontrar la media del *Universo*, con el grado de certeza especificado en el *Nivel de seguridad* (confianza). Es similar al factor de tolerancia (por ejemplo: más o menos 4 %) presente en cualquier medida. También se le llama *Intervalo de tolerancia* o *Precisión.*

Intervalo de tolerancia Véase *Precisión.*

Marco * Una lista de todos los elementos del *Universo*. En algunos casos es posible que el marco de muestreo no incluya todos los elementos del *Universo*, como, por ejemplo, cuando éstos se encuentran en tránsito o de alguna otra manera no están disponibles para su selección. A menos que las diferencias entre el marco y los datos del *Universo* sean insignificantes, es posible que los resultados de la muestra no sean suficientemente representativos de dicho universo.

Marco de muestreo Véase *Marco.*

Media — Véase *Media aritmética.*

Media aritmética — La suma de los valores de la *Población* dividida entre el número de elementos de la misma. Por lo tanto, si la *Población* consta de cinco cuentas con saldos de \$ 2, \$ 5, \$ 12, \$ 14 y \$ 17, el saldo medio de estas cuentas sería su total (\$ 50) dividido entre su número (5), o sea, \$ 10. La media es una medida de la *Tendencia central* de una *Distribución de frecuencias.*

Mediana — Un valor que divide una serie ordenada, en forma tal que por lo menos la mitad de los elementos son iguales o mayores que dicho valor, y cuando menos la mitad de dichos elementos son iguales o menores que el mismo.

Moda (o Modo) — El valor en torno al cual tienden a concentrarse los elementos; el valor más frecuente o más común en una *Distribución de frecuencias.* El valor de la moda corresponderá al valor de la abscisa del punto máximo de una curva que represente a una *Distribución de frecuencias.*

Muestra — El conjunto de elementos de un *Universo* que se seleccionan para examen.

Muestra con exclusión — Un método de muestreo que consiste en eliminar las unidades más pequeñas de un *Universo* y extraer una muestra solamente de las restantes. Aunque no es estrictamente un tipo de muestreo probabilístico, algunas veces es sumamente económico, y puede usarse en algunas situaciones si se tiene suficiente cuidado.

Muestra probabilística — Véase *Selección aleatoria.*

Muestras aleatorias interpenetrantes — Una derivación del *Muestreo aleatorio simple;* en vez de sólo un resultado independiente basado en la muestra, es posible tener muchos. Un método consiste en dividir el *Universo* en distintas zonas, aproximadamente de la misma magnitud, hacer selecciones aleatorias en cada una y tener tantas submuestras como zonas. Por ejemplo, un *Universo* de 100 000 se divide en 10 zonas; se hace una selección de 10 elementos de cada una; la primera selección de *cada* zona da por resultado una submuestra de 100 elementos que *penetra* las 10 zonas. La siguiente selección da otra submuestra penetrante de 100, y así hasta que se hayan obtenido 10 conjuntos de muestras penetrantes. La técnica que se usa para medir la concordancia entre las submuestras, a fin de identificar cualesquiera resultados absurdos o no confiables. Compárese con *Muestreo replicado.*

Muestreo — La selección y el análisis de un número finito de elementos a fin de obtener información acerca del *Universo* del que fueron seleccionados.

Muestreo aleatorio simple * — Método por el cual se seleccionan, enteramente al azar, los elementos de la muestra, de entre todos los que pertenecen al *Universo.* Este procedimiento brinda a cada elemento del universo igual oportunidad de ser seleccionado al extraerse la muestra. Un método común consiste en utilizar las *Tablas de números aleatorios.*

Muestreo científico — Un procedimiento de muestreo en el que el diseño y la selección de la muestra se efectúan de acuerdo con la teoría matemática, en especial la teoría de las probabilidades.

Muestreo de aceptación — Plan de muestreo utilizado para distinguir entre lotes de elementos buenos y lotes malos; definidos estos últimos como los que contienen elementos defectuosos en un porcentaje superior al especificado.

Muestreo de atributos — Véase *Atributo.*

Muestreo de conglomerados — El método de muestreo por el cual se ordena el *Universo* en grupos o "conglomerados" de elementos. El primer paso consiste en hacer una selección aleatoria de los conglomerados que se van a incluir en la muestra. Luego, se pueden muestrear los elementos de los conglomerados seleccionados (*Muestreo polietápico*), o bien se puede hacer un *Censo* de ellos. Por ejemplo, al investigar un gran número de cajas de herramientas, es posible reducir el costo de la auditoría mediante el empleo del muestreo por conglomerados. Primero, se hace una selección aleatoria de las cajas, y después una selección, también aleatoria, de los registros de herramientas que hay dentro de las cajas seleccionadas. El muestreo de conglomerados se usa generalmente para obtener los resultados más precisos con un presupuesto fijo, pero no los más exactos en términos de números de muestra.

Muestreo de descubrimiento * — Una técnica especializada para localizar casos de ciertas clases de irregularidades, tales como fallas al seguir procedimientos básicos de control interno, o evidencias de mal manejo y de posible fraude. Se utilizan cuadros especiales de probabilidades. El muestreo de descubrimiento no tiene por fin dar una base para evaluar el contenido de error en toda la *Población.* Proporciona el riesgo calculado, implícito en el descubrimiento de un error (o de ninguno) en una población. También se le llama *Muestreo exploratorio.*

Muestreo de estimación — La determinación, dentro de límites estipulados y con un riesgo especificado, de la frecuencia con que ocurren eventos tales como errores, violaciones al control interno, etc. En otras palabras, el método de muestreo utilizado para permitir al auditor llegar a una conclusión respecto de toda la *Población.* Comprende *Estimaciones de frecuencia* (*Muestreo de atributos*) y *Estimación del valor en unidades monetarias* (*Muestreo de variables*). El muestreo de estimación es el método utilizado más a menudo por el auditor.

Muestreo de suspensión o continuación — Una técnica introducida por la Fuerza Aérea de los Estados Unidos para hacer evaluaciones basadas en los resultados de una submuestra o de una muestra aleatoria simple reducida. Se han elaborado cuadros especiales para indicar el riesgo implícito y el estado probable del *Universo,* con base en la aceptación de una muestra reducida. Al usar los cuadros, el auditor deberá determinar: *1)* la tasa máxima aceptable de error —considerando las normas existentes o la experiencia—, y *2)* qué grado de seguridad desea tener de que sus conclusiones, basadas en la muestra y en la tasa aceptable de error, sean correctas.

Muestreo dirigido

La selección de aquellos elementos que, a juicio de quien la lleva a cabo, son los más adecuados para sus propósitos de auditoría. El muestreo dirigido difiere del *Muestreo estadístico* en un aspecto importante: no puede demostrarse matemáticamente el grado en que la muestra dirigida representa al universo del cual fue seleccionada. Por tanto, el muestreo dirigido se utiliza cuando no es esencial contar con una determinación precisa del estado probable del universo, o cuando no es posible, práctico o necesario usar el muestreo estadístico.

Muestreo doble

Un método para efectuar la selección inicial de una muestra grande, de la cual es posible obtener información que puede usarse como base para la selección de una muestra más pequeña y detallada. Dicho método se usa en la realización de encuestas porque es rápido y relativamente barato. Por ejemplo, el auditor puede investigar 500 elementos para obtener "grupos naturales" de información fácilmente analizables; luego, tomar una muestra de esos 500 para obtener información más detallada. Este método también se utiliza cuando se desconoce cierta información acerca del *Universo* y solamente puede obtenerse de una muestra. Compárese con el *Plan de muestreo doble.*

Muestreo en cadena *

Véase *Muestreo sistemático.*

Muestreo en dos etapas y polietápico *

Subdivisiones de muestras. Este método se usa en encuestas. Por ejemplo, las unidades primarias podrían consistir en una muestra de los estados de un país; las unidades secundarias, en una muestra de ciudades de los estados seleccionados; y las unidades terciarias, en una muestra de manzanas de dichas ciudades, etcétera.

Muostreo estadístico

La selección, con una base científica, y el análisis de un número finito de elementos con el objeto de obtener información acerca del *Universo* del que fueron seleccionados.

Muestreo estratificado

Un método de muestreo según el cual primeramente se separan los elementos del *Universo* en dos o más clases llamadas estratos. Los elementos contenidos dentro de cada estrato son, en general, comparables y cada estrato se muestra de manera independiente. Este método es apropiado cuando hay una gran discrepancia entre los elementos de la muestra respecto a las características que se medirán. Por ejemplo, supóngase que en un *Universo* de 1 000 facturas, 20 de ellas, que suman $ 200 000, representan la mitad del valor monetario del *Universo* total. Obviamente, deberá examinarse todo el estrato compuesto de estas 20 facturas, en tanto que el resto podrá ser muestreado. Véase *Distribución de frecuencias.*

Muestreo estratificado no proporcional

La selección de un número igual de elementos, de estratos desiguales, a diferencia del *Muestreo estratificado proporcional.* Este procedimiento da por resultado muestras a las que posteriormente deberá aplicarse un factor de ponderación para que queden sin *Sesgo.* Por ejemplo, en un *Universo* de 360 (formado de estratos que contienen 60, 120 y 180) y un tamaño de muestra de 40 para cada estrato, los resultados de la muestra deberán multiplicarse por el recíproco de la razón de la muestra al tamaño del estrato. El recíproco se determina de la siguiente manera:

Estrato	*Tamaño del estrato*	*Tamaño de la muestra*	*Razón*	*Recíproco*
A	60	40	40/60 ó 2/3	3/2
B	120	40	40/120 ó 1/3	3/1
C	180	40	40/180 ó 2/9	9/2

Luego se aplican los recíprocos a los resultados de muestra, ya sea que se estén midiendo unidades monetarias, errores u otras características.

Muestreo estratificado proporcional

Un método por el cual, de cada estrato, se selecciona una muestra que sea una proporción igual a la que guarda la muestra total respecto del *Universo*. En un universo de 360 (formado de estratos que contienen 60, 120 y 180 elementos) y con un tamaño de muestra de 120, las muestras de estrato serían de 1/3 de cada uno de los tres tamaños de estrato: 20, 40 y 60. También se le conoce como *Fracción de muestreo*. Compárese con *Muestreo estratificado no proporcional*.

Muestreo exploratorio

Es lo mismo que *Muestreo de descubrimiento*.

Muestreo objetivo

Muestreo basado en métodos científicos.

Muestreo polietápico *

Muestreo realizado por etapas, en cada una de las cuales las unidades de muestreo son submuestreadas con base en las unidades mayores escogidas en la etapa anterior.

Muestreo por intervalos

Un método por el cual se seleccionan elementos del *Universo* en tal forma que hay un intervalo uniforme entre los elementos muestrales. El primer elemento de la serie debe escogerse al azar. También se conoce como *Selección aleatoria restringida* y *Muestreo sistemático*.

Muestreo por números aleatorios

Selección aleatoria realizada mediante el uso de *Tablas de números aleatorios*.

Muestreo proporcional al tamaño

Un método para asegurarse de que las unidades más grandes de una *Población* tendrán mayores probabilidades de ser seleccionadas que las unidades más pequeñas. Supongamos que se desea seleccionar una caja de herramientas, de entre 5 cajas, para estudiar ciertas características de población. Supongamos también que el número de herramientas varía en la forma que se indica en el siguiente cuadro:

Cajas de herramientas	*Número de herramientas*	*Números consecutivos asignados para muestreo*
A	5 000	1 2 3 4 5
B	4 000	6 7 8 9
C	3 000	10 11 12
D	2 000	13 14
E	1 000	15

A fin de hacer variar la probabilidad de selección de acuerdo con el número de herramientas que hay en cada caja, se asignan números consecutivos a cada una en proporción al número de herramientas que contiene, según se indica en la tercera columna del cuadro. Entonces se selecciona el primer número entre 1 y 15 de de una *Tabla de números aleatorios* para determinar cuál será la caja que se estudiará.

Muestreo replicado — Un método según el cual se examinan grupos de muestras aleatorias y se comparan las tasas de exactitud (o error). Si éstas son similares, los elementos probados se consideran representativos del *Universo*. Compárese con *Muestras aleatorias interpenetrantes.*

Muestreo serial* — Véase *Muestreo sistemático.*

Muestreo sistemático — Un método en virtud del cual se seleccionan elementos del *Universo* en forma tal que exista un intervalo uniforme entre cada elemento de la muestra y el que le precede. El primer elemento de la serie debe escogerse en forma aleatoria. También se denomina *Selección aleatoria restringida* y *Muestreo por intervalos.*

Muestreo tipificado — Véase *Muestreo sistemático.*

NCA — *Nivel de calidad aceptable.*

Nivel de calidad aceptable — La calidad considerada conveniente para los fines de aceptar o rechazar una *Población,* mediante el uso de un plan formal de muestreo. Por ejemplo: 2 % o menos de error en las unidades muestreadas.

Nivel de confianza — Es una estimación del grado de certeza de que la media del *Universo* se encontrará dentro del *Intervalo de confianza.* También se le conoce como *Nivel de seguridad.* Los niveles de confianza de uso más común son 90 % y 95 %. Por ejemplo, al utilizar un nivel de confianza de 95 %, si todas las muestras posibles de un universo fueran tomadas en la misma forma y bajo las mismas condiciones, se obtendrían los mismos resultados 95 veces de cada 100.

Nivel de seguridad — Véase *Nivel de confianza.*

Número de aceptación — El máximo número de elementos defectuosos tolerado en un plan de *Muestreo de aceptación.*

Ordenamiento — La disposición de valores numéricos del menor al mayor, o viceversa. Por ejemplo un ordenamiento de tasas semanales de salarios: $ 140, $ 142, $ 142, $ 147, $ 150, $ 150, $ 150, $ 153, etcétera.

Parámetro — Una medida calculada sobre la base de un *Universo* o que describe a éste. Compárese con *Estadística.* Un parámetro es el valor de la característica en estudio obtenida de un *Censo* del cien por ciento. Los valores calculados con base en una muestra extraída del universo sólo pueden ser estimaciones o aproximaciones de los verdaderos parámetros. Por ejemplo, supóngase un universo de 1 000 solicitudes de compra de las cuales 30 en realidad no hubieran sido debidamente aprobadas, lo cual constituye la característica en estudio. Estas 30 representan el parámetro, o sea la verdadera característica, del universo. (Por supuesto, generalmente se desconoce el parámetro.) Supóngase, además, que se examina una muestra de 158 (*Nivel de confianza* de 95 %, Precisión de ± 2 %, tasa esperada de ocurrencia de no más de 2 %). Hagamos la suposición adicional de que la muestra tiene una tasa real de error de 2 % (la *Estadística*). El parámetro estimado para todo el universo se proyectaría entonces entre 0 y 40 errores; en tanto que el verdadero parámetro es 30.

Periodo de muestreo — El lapso de tiempo cubierto por la muestra. Cuando la muestra se toma de un periodo restringido (una semana, un mes) tal vez los resultados no se apliquen al resto del periodo cubierto por la auditoría.

Plan de muestreo doble — Un método por el cual se obtiene un máximo de dos muestras antes de tomar una decisión. Compárese con el *Plan de muestreo monoetápico* y el *Plan de muestreo secuencial.*

Plan de muestreo monoetápico — La selección de una muestra de tamaño fijo tomada de un grupo definido de elementos, y la aceptación o rechazo del grupo, con base en el número de errores encontrados en la muestra.

Plan de muestreo múltiple — Véase *Plan de muestreo secuencial.*

Plan de muestreo secuencial* — A menudo denominado muestreo múltiple. Es una ampliación del *Muestreo doble*. Consiste en tomar una serie de muestras, hasta que sean concluyentes las pruebas de la calidad del trabajo. Siempre que la suma de los errores obtenidos en las muestras sea suficiente para indicar en forma definitiva si el trabajo es bueno o deficiente, se toma de inmediato la decisión de aceptarlo o rechazarlo. Si el número de errores no establece cabalmente la tasa de error, se obtienen nuevas muestras hasta que se pueda tomar una decisión.

Población — Véase *Universo.*

Precisión — La amplitud dentro de la cual se encontrará la media del *Universo*, con el grado de certeza especificado por el *Nivel de confianza*. Es similar al factor de tolerancia (más o menos 4 %, por ejemplo) que se encuentra en cualquier medida. También se le denomina *Intervalo de tolerancia* o *Intervalo de confianza.*

Precisión de muestreo — Véase *Precisión.*

Probabilidad (Nivel de probabilidad) — La razón de la frecuencia de ciertos eventos a la frecuencia de todos los eventos posibles en una serie o conjunto. Dicho en otras palabras, el número de veces que puede ocurrir un evento de una manera específica, en comparación con el número de veces que puede ocurrir en todas las formas posibles. Generalmente se expresa como una razón decimal, que puede convertirse a un porcentaje multiplicándola por 100.

Procedimientos de estimación — Los métodos usados para estimar el volumen o extensión del *Universo* por medio de muestras. Estos métodos comprenden *Estimación por amplia-*

ción, Estimación por razones, Estimación por diferencia y *Estimación por regresión lineal.* Véanse ejemplos y procedimientos en el Apéndice Técnico, a partir de la página 111.

Promedio — Un valor típico que tiende a resumir o describir la masa de datos. El promedio sirve de base para medir o evaluar valores extremos o desusados.

Prueba dirigida — El tipo de muestreo utilizado para realzar una situación indeseable o aislar una discrepancia con el menor esfuerzo. Véase *Muestreo de descubrimiento o exploratorio.* Sin embargo, este tipo de muestreo nunca debe dejar la impresión de que los resultados obtenidos son representativos de todo el *Universo.*

Prueba en bloques — Una prueba conforme a la cual las muestras se basan en periodos de tiempo o agrupamientos consecutivos; por ejemplo, toda la facturación preparada en un mes determinado, o las cuentas por cobrar de todos los clientes cuyos nombres empiezan con "E" y "M". Cualquier conclusión a que se llegue respecto a las transacciones en otros meses o con otras letras es meramente subjetiva.

Representativa — Se dice de una muestra. Si ésta es representativa del *Universo* del que se obtiene, sus características son las mismas que las de dicho universo, dentro de los límites fijados de *Nivel de confianza* y de *Precisión.*

Riesgo — Véase *Precisión.*

Riesgo de muestreo — Siempre que una decisión o una conclusión esté basada en un resultado muestral, existirán riesgos concomitantes de que la decisión sea incorrecta. En el muestreo estadístico, el grado de riesgo puede determinarse matemáticamente. Véase *Nivel de confianza.*

Selección aleatoria — Una selección que está regida totalmente por las leyes del azar y en la que cada uno de los elementos de la *Población* deberá tener igual oportunidad de ser escogido. En una población estratificada, en la que algunos de los estratos se muestrean con mayor intensidad que otros, cada elemento de la población total no tiene necesariamente una oportunidad "igual" de selección, pero las técnicas de selección aleatoria deberán proporcionar una oportunidad de selección "conocida". Una muestra aleatoria es la que se obtiene por selección aleatoria y, por lo tanto, es representativa de los resultados que podrían esperarse si se tomaran más muestras.

Selección aleatoria con prueba subjetiva — La selección de elementos al azar realizada por un auditor, cuando, para prueba, escoge uno de ellos sin base científica, pero desconociendo de antemano cuáles serán sus atributos. Algunos estadígrafos consideran que con este método de selección es escasa la posibilidad de que se produzca un sesgo, pero que solamente deberá ser utilizado si el auditor es consciente de lo que significa lo aleatorio.

Selección aleatoria irrestricta — Véase *Selección aleatoria.*

Selección aleatoria restringida — Véase *Muestreo sistemático.*

Sesgo * — La existencia de un factor selectivo que influirá en la determinación del contenido de la muestra en una forma particular. Generalmente el sesgo es un efecto que impide que un resultado estadístico sea representativo, al distorsionarlo sistemáticamente, a diferencia de un error aleatorio, ya que éste sólo puede producir una distorsión ocasional que tiende a equilibrarse en el promedio. Un ejemplo de la posibilidad de introducir un sesgo en los resultados de muestras es la obtención de una muestra de una nómina en la que figurase en cada décima línea un ayudante de capataz. De esta manera, una muestra sistemática con un intervalo de muestreo igual a 10, o algún múltiplo de 10, incluiría ya sea a todos los ayudantes de capataz o bien a ninguno de ellos. Otro ejemplo sería la selección para prueba de solamente aquellos elementos a los que es fácil probar o localizar.

Sin respuesta — Situación definida por el hecho de que no se localiza o investiga un elemento seleccionado para prueba. Esto permite que se introduzca un *Sesgo* en la muestra. Una vez que un elemento ha sido designado para examen, deben usarse todos los medios posibles para obtener una respuesta. Al confirmar las cuentas por cobrar, por ejemplo, el método Hansen-Hurwitz estipula que el auditor: *1)* debe pedir respuestas precisas a las solicitudes de confirmación y *2)* deberá tomar una muestra estadística de las faltas de respuesta y seguir trabajando intensamente sobre ellas o verificar la exactitud de los saldos por medio de otras técnicas equivalentes de auditoría.

Tabla de intervalos de tolerancia — Una tabla que se utiliza para evaluar resultados muestrales cuando la tasa de error que revela el muestreo difiere de la tasa de error esperada que se consideró al determinar inicialmente el tamaño de la muestra.

Tablas de números aleatorios — Tablas de dígitos dispuestos de tal manera que el auditor puede usarlas con toda confianza para la selección de muestras, con la certeza de que han sido mezclados concienzudamente (es decir, dispuestos en forma aleatoria). También se denominan *Dígitos decimales aleatorios.* Algunas tablas utilizadas comúnmente son: la *Tabla de 105 000 dígitos decimales aleatorios,* calculada por la Interstate Commerce Commission (Comisión de Comercio Interestatal) y *Un millón de dígitos aleatorios,* calculados por The Rand Corporation. Véase *Generador de números aleatorios.*

Tamaño muestral promedio (TMP) — El tamaño medio (esperado) de la muestra que es necesario en un plan de *Muestreo de aceptación* antes de que se adopte la decisión de aceptar o rechazar.

Tendencia central — Un valor singular que representa la masa total de observaciones. Entre las medidas de tendencia central se encuentran la *Media aritmética, la Mediana* y la *Moda.*

Truncamiento * — La suspensión del muestreo en un punto arbitrario. Se hace para evitar muestras demasiado grandes en el *Muestreo secuencial.*

Unidad elemental	La unidad singular sometida a prueba. Véase *Unidad muestral.*
Unidad muestral	La unidad seleccionada para examen. Si inicialmente se selecciona un grupo o conglomerado, entonces éste es la unidad muestral. Los elementos singulares del conglomerado se denominan *Unidades elementales.*
Unidad para muestreo	Unidad del *Universo* sujeto a examen, extraída para su inclusión en una muestra. Puede ser una unidad primaria de la cual se obtengan nuevas unidades de muestra, o un elemento singular extraído directamente del universo para su inclusión en la muestra final.
Universo	El conjunto o totalidad de los elementos o unidades acerca de los cuales se desea obtener información. También se le denomina *Población* o *Campo.*
Universo finito	Universo constituido por un número limitado de elementos; por ejemplo, la cantidad de cuentas por cobrar en cierta fecha. Compárese con *Universo infinito.*
Universo heterogéneo	Universo compuesto de elementos de naturaleza diferente. Tal universo deberá probarse por medio del *Muestreo estratificado* o ser examinado por completo.
Universo homogéneo	Universo formado por elementos que son muy similares. Tal universo se presta al *Muestreo sistemático.*
Universo infinito	Universo que contiene un número ilimitado de elementos. Por ejemplo, la operación de correr asientos realizada indefinidamente. Compárese con *Universo finito.*
Variabilidad	Una medida concebida para describir la diseminación o dispersión de una *Distribución de frecuencias.*
Variable	La característica cuantitativa de un elemento de la *Población;* por ejemplo, el plazo vigente de las cuentas por cobrar o el monto del saldo de una cuenta. Véase *Estimación del valor en unidades monetarias.* Compárese con *Atributo.*
Variación	Véase *Coeficiente de variación.*
Varianza *	El cuadrado (la segunda potencia) de la *Desviación estándar.*

Ajuste de Curvas.

Uno de los principales objetivos de muchos problemas [illegible] la Administración de Negocios es hacer predicciones en base a ecuaciones matemáticas.

Ejemplo: predicciones de ventas.

Usualmente estas predicciones requieren que la fórmula encontrada y que relaciona la variable dependiente (cuyo valor uno quiere predecir) con una ó más variables independientes

Caso 1 — La variable dependiente (por predecir) está en función de una sóla variable independiente.

Muchos problemas nos dan grandes variaciones debidas al azar, para los valores de y suponiendo X fija.

Ejemplo: de variac. de long. de especímenes en un horno debido a " de temperatura de la atmósfera interna.

De esos ejemplos puede decirse que y es una v.a cuya distribución depende de X. En muchos de tales casos nos interesará las relaciones entre X y la media de las distribuciones correspondientes de las y's.

Por lo pronto consideraremos que X es fija (no aleatoria)

Considerémos el caso cuando la curva de regresión de y sobre X es lineal, esto es, cuando la media de las distrib. de las y's está dada por: $\alpha + \beta X$

En general un valor observado de y diferirá de la media y denotaremos esa diferencia por ε

$$\Rightarrow y = \alpha + \beta x + \varepsilon$$

[illegible] sobre α de tal [illegible]

[illegible]

i	1	2	3	4	5	6	7	8	9	10
X_i	0	20	40	60	80	100	120	140	160	180
y_i	0.01	0.12	0.24	0.38	0.51	0.67	0.84	1.01	1.15	1.31

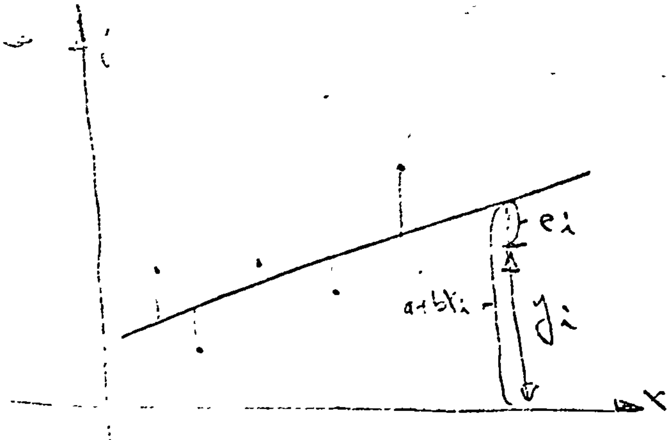
X = VENTAS DEL MES i

y = UTILIDADES " "

De la gráfica del pizarrón podemos ver que la relación es lineal.

α y β determinan completamente la recta de regresión de y sobre X y la estimación de α y β es equivalente a encontrar la ecuación de la línea recta que mejor se ajusta a los puntos.

Podríamos trazar la línea a ojo y estimar para nuestro ojo más bien los parámetros, pero para el caso de la figura fi



esto ya no es tan fácil y ni tan evidente.

Trataremos de ajustar la "mejor" curva y veremos que entendemos por mejor:

Si predecimos y por la ecuación: $y' = a+bx$

(a y b fijos), $\Rightarrow$ e_i el error de predecir el valor de y correspondiente ~~a y_i~~ a un valor dado de X_i, es.

$$y_i - y'_i = e_i$$

La ecuación: $y' = a + bx$, nos da una estimación (3) de la recta de regresión cuya ecuación actual pero desconocida es

$$y = \alpha + \beta x$$

e_i - error actual al predecir y_i y está dado por:

$$e_i = y_i - y_i'$$

Determinaremos a y b de tal suerte que los errores estimados sean en algún sentido lo más pequeño posible. (

Puesto que no podemos minimizarlos cada uno separadamente se antojaría tomar su suma:

$$\sum_{i=1}^{n} e_i$$

tan cercana como sea posible a 0 -

Sin embargo como la Σ puede hacerse $= a$ 0 escogiendo todas aquellas líneas que nos den errores que se anulen minimizaremos en vez de la Σ, la Σ de los cuadrados de $e_i \Rightarrow$ escogeremos a y b de tal suerte que

$$\sum_{i=1}^{n} \left[y_i - (a + bx_i)\right]^2$$

sea mínimo.

Tal criterio llamado de los mínimos cuadrados ... da valores de a y b (estimadores de α y β) que tienen muchas propiedades deseables, algunas de las que mencionaremos más adelante.

Una condición necesaria para tener un mínimo es que las derivadas parciales con respecto a a y b sean nulas

$$\Rightarrow \quad 2\sum_{i=1}^{n} \left[y_i - (a + bx_i)\right](-1) = 0$$

$$2\sum_{i=1}^{n} \left[y_i - (a + bx_i)\right](-x_i) = 0$$

Escribiendo estas ecuaciones de otra forma

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Para el ejemplo:

$$\sum_{i=1}^{n} x_i = 860 \qquad \sum_{i=1}^{n} x_i^2 = 98.800$$

$$\sum_{i=1}^{n} y_i = 165.2 \qquad \sum_{i=1}^{n} x_i y_i = 18469.0$$

Las ecuaciones normales son:

$$165.2 = 10a + 860b$$
$$18469.0 = 860a + 98.800\,b$$

Resolviendo estas dos ecuaciones tenemos:

$$a = 1.90 \quad , b = 0.17$$

La ecuación de la línea recta que nos da el mejor ajuste desde el punto de vista de los mínimos cuadrados es:

$$y' = 1.90 + 0.17x$$

Esta ecuación puede usarse para predecir cuál será el valor de y para $x = 80$

$$y' = 1.90 + (0.17)\cdot 80 = 15.5$$

Es imposible hasta ahora hacer ninguna afirmación acerca de la bondad de la ~~hipótesis~~ predicción, a menos que hagamos algunas hipótesis acerca de la ley de distribución de los y's y acerca de la verdadera naturaleza de la regresión, o sea que si para una x dada la verdadera media es de la forma $\alpha + \beta x$.

Se demuestra que a y b son estimadores insesgados de α y β, y además mediante el Teorema de Gauss-Markov que dice que entre todos los insesgados ... y que ...

[illegible] de los estimadores de [illegible] insesgados son los que tienen la eficiencia [illegible] pequeña.
La demostración está en el libro de F.A.: Graybill.

INFERENCIAS BASADAS EN LOS ESTIMADORES OBTENIDOS POR MÍNIMOS CUADRADOS

Hipótesis:

i) La regresión es lineal

ii) Las n var. a. que tienen los valores y_i $(i=1,2,\ldots,n)$ están independientemente distribuidas con dist. normal de medias $\alpha+\beta x_i$ y variancia común σ^2.

Si escribimos

$$y_i = \alpha+\beta x_i + \epsilon_i$$

se sigue de las hipótesis anteriores que los ϵ_i son valores de v.a. independiente y normalmente distribuidos y variancia común σ^2.

Estas hipótesis se requieren únicamente para el análisis de bondad del ajuste, mas no para obtener los estimadores de α y β por medio de los mínimos cuadrados.

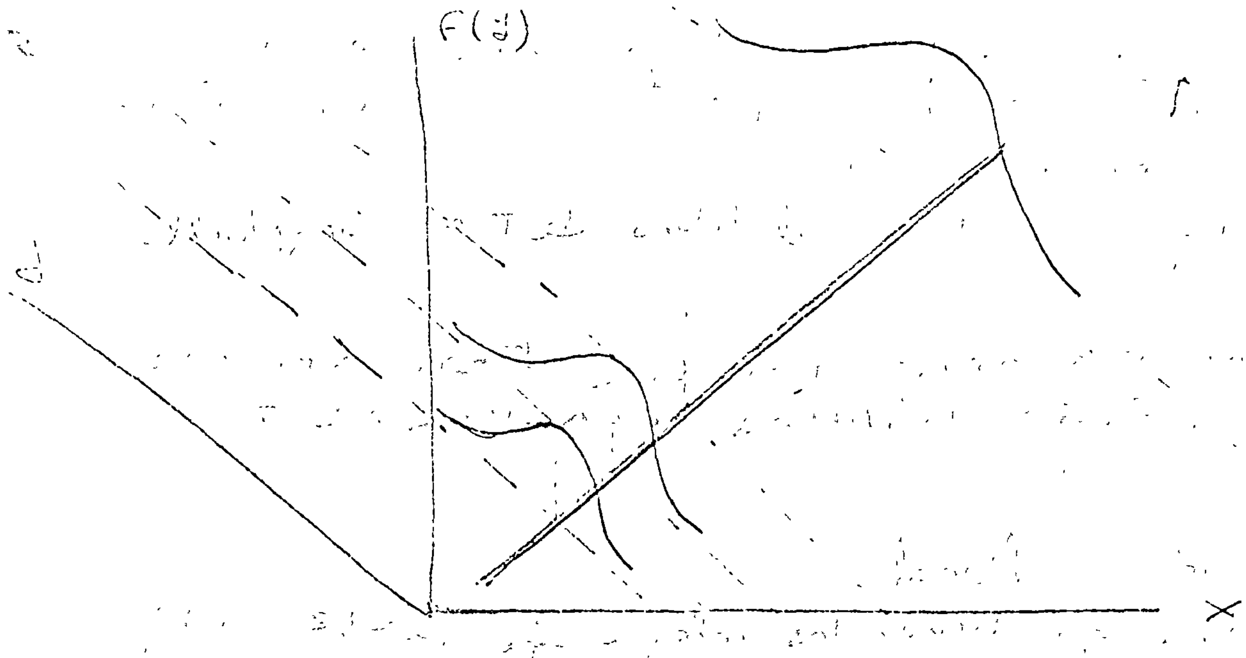
$$S_{xx} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2$$

$$S_{xy} = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Usando esa notación puede demostrarse:

$$a = \bar{y} - b\,\bar{x} \qquad , \qquad b = \frac{S_{xy}}{[illegible]}$$



Nótese [illegible] de las varianzas S_x^2 y S_y^2:

$$S_x^2 = \frac{S_{xx}}{n(n-1)} \quad y \quad S_y^2 = \frac{S_{yy}}{n(n-1)}$$

La varianza σ^2 se define en términos de las variaciones verticales de los puntos a la línea de mínimos cuadrados; [illegible] la desviación i-ésima será:

$$y_i - y_i' = y_i - (a + b x_i)$$

y el estimador de σ^2 será:

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left[y_i - (a + b x_i) \right]^2$$

[illegible] error estándar de estimación

$$S_e^2 = \frac{S_{xx} S_{yy} - (S_{xy})^2}{n(n-2)\, S_{xx}}$$

[illegible] $n-2$ [illegible]

[illegible] puede considerarse que cumple hipótesis del principio

$(n-2)\, S_e^2/\sigma^2$

es un valor de una v.a con dist. χ^2 y $n-2$ grados de libertad.

Además los parámetros:

$$t = \frac{a-\alpha}{S_e}\sqrt{\frac{n\, S_{xx}}{S_{xx} + (n\bar{x})^2}}$$

$$t = \frac{b-\beta}{S_e}\sqrt{\frac{S_{xx}}{n}}$$

tienen distribución t con $n-2$ grados de libertad.

Usando estos 2 parámetros podemos calcular los intervalos de confianza para a y b como estimadores de α y β:

$$\cancel{a}\, a \pm t_{\alpha/2} \cdot S_e \sqrt{\frac{S_{xx} + (n\bar{x})^2}{n\, S_{xx}}}$$

$$\cancel{b}\, b \pm t_{\alpha/2} \cdot S_e \sqrt{\frac{n}{S_{xx}}}$$

Ejemplo:

$\sum_{i=1}^{n} y_i^2 = 3563.18$

$S_{xx} = 10\,(98\,800) - 860^2 = 248{,}400$

$S_{yy} = 10\,(2563.18) - (165.2)^2 = 8373.76$

$$S_{xy} = 10(12469) - 300(165.2) = 42618$$

$$\Rightarrow \quad S_e^2 = \frac{(248400)(8213.76) - (42618.0)^2}{(10)(8)(248,400)}$$

Puesto que $t_{0.025} = 2.306$ para $10 - 2 = 8$ grados de libertad, tenemos el siguiente 0.95 intervalo de confianza para α y β

$$1.50 \pm (2.306)(3.59)\sqrt{\frac{248\,400 + 300^2}{10(248,400)}}$$

ó

$$-3.32 < \alpha < 7.12$$

y

$$0.17 \pm 2.306 \cdot 3.59\sqrt{\frac{10}{248,400}}$$

ó

$$0.12 < \beta < 0.22$$

H_0: $\beta = \beta_0$

$$t = \frac{(b - \beta_0)}{S_e}\sqrt{\frac{S_{xx}}{n}}$$

Hipótesis Alternativa	Rechazar Ho si
$\beta < \beta_0$	$t < -t_\alpha$
$\beta > \beta_0$	$t > t_\alpha$
$\beta \neq \beta_0$	$t < -t_{\alpha/2}$ ó $t > t_{\alpha/2}$

t_α se obtiene de las tablas para $n-2$ grados de libertad.

... ($\beta = 0$) ⇒ la recta de regresión es horizontal y la media de "y" no depende linealmente de X.

$H_0: \beta = 0$ y $H_1: \beta \neq 0$, usando los cálculos anteriores tenemos:

$$t = \frac{0.17 - 0}{3.59}\sqrt{\frac{248\,000}{10}} = 7.46$$

y puesto que $t_{0.025}$ para 8 grados de libertad es $2.306 < 7.46$ ⇒ rechazamos H_0 con un 0.05 de nivel de significancia ⇒ concluimos que existe una relación lineal entre X y Y.

Si queremos conocer ahora la media de Y para un valor dado de X, digamos X_0 ⇒ la cantidad que queremos estimar es $\alpha + \beta X_0$ y podría parecernos razonable usar como estimador $a + bX_0$, donde a y b son los valores obtenidos por el método de mínimos cuadrados.

Puede demostrarse que ese estimador tiene una varianza dada por

$$\sigma^2\left[\frac{1}{n} + \frac{n(X_0 - \bar{X})^2}{S_{xx}}\right]$$

... de confianza $(1-\alpha)$ ...

$$\boxed{(a + b x_0) \pm t_{\alpha/2} \cdot S_e \sqrt{\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{S_{xx}}}}$$

donde $t_{\alpha/2}$ se obtiene de las tablas de distribución t de student para $n-2$ grados de libertad.

Si deseamos predecir y para $x = x_0$ $\Rightarrow$ ... se conoce y es el valor de una v.a. con media $\alpha + \beta x_0$ y varianza σ^2. Sin embargo ... debemos considerar la variabilidad ... donde y, a y b son v.a. ... la fórmula siguiente

$$\boxed{(a + b x_0) \pm t_{\alpha/2} \cdot S_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{S_{xx}}}}$$

donde el número de grados de libertad para $t_{\alpha/2}$ es $n-2$

Ejemplo:
$x = 80 \Rightarrow$ intervalo al 0.95 de confianza para $\alpha + 80\beta$ es:

$$15.5 \pm (2.306)(3.59)\sqrt{\frac{1}{10} + \frac{10(80-86)^2}{248400}}$$

$$12.9 < \alpha + 80\beta < 18.1$$

Correspondientemente, los límites de 11
0.95 de predicción, para el valor futuro
de y cuando $x = 80$ están dados
por:

$$15.5 \pm (2.306)(2.59)\sqrt{1+\frac{1}{10}+\frac{10(80-86)^2}{213,100}}$$

ó $\boxed{15.5 \pm 8.7}$

Notamos que no obstante que la media de
la distribución de y's cuando $x=80$
puede estimarse muy cercanamente, el valor
[illegible] de [illegible] predicción no puede predecir
[illegible] buena precisión
De la fórmula vemos que aún cuando $n \to \infty$
[illegible] no tiende a cero. El ancho
del intervalo no tiende a [illegible] Se a
[illegible], [illegible] depende más que nada
[illegible] la variabilidad de los datos
[illegible] que si queremos extra-
polar [illegible] así mismo que [illegible] intervalo se ensan-
chan los límites
[illegible] considerablemente
Por ejemplo, los límites de un intervalo con
0.95 de confianza para $x = 210$ están
dados por:

$$42.7 \pm (2.306)(2.59)\sqrt{1+\frac{1}{10}+\frac{10(210-86)^2}{213,100}}$$

ó 42.7 ± 11.8 ó sea $30.9 \le y \le 54.5$
[illegible] que la regresión lineal sigue

Regresión Curvilinea.

Hasta ahora hemos estudiado el caso de la regresión lineal, o sea, para cualquier X dado, la media de la distribución de las y's está dada por $\alpha + \beta x$.

Estudiaremos ahora el caso para el que dado X la media de la distribución de las y's esté dada por:

$$\beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

ó

$$\alpha \cdot \beta^x \quad ; \quad \log y' = \log \alpha + x \log \beta$$

Veamos primeramente el segundo caso. De la transformación hecha vemos que tenemos también una línea recta y podemos tener estimadores de $\log \alpha$ y $\log \beta$ y por tanto de α y β aplicando el método visto antes.

Para dar un ejemplo tenemos:

X	y	log y
5	92.7	1.967
10	58.3	1.766
15	59.5	1.775
20	41.7	1.620
25	45.6	1.659
30	31.8	1.502
35	38.3	1.583
40	19.9	1.299

$\sum x = 180$; $\sum x^2 = 5100$

$\sum \log y = 13.171$; $\sum x \log y = 280.420$

Las ecuaciones normales son:

$$13.171 = 8 (\log a) + 180 (\log b)$$

$$280.420 = 180 (\log a) + 5100 (\log b)$$

Resolviendo el sistema tenemos:

$$\log a = 1.986$$
$$\log b = -0.0151$$

Así: $a = 96.8$ y $b = 0.966$ y la curva tiene por ecuación: $y' = 96.8 (0.966)^x$

que puede escribirse como:

$$y' = 96.8 \cdot 10^{-0.051x} \quad ó \quad y' = 96.8 \; e^{-0.0346x}$$

ya que: $0.966 = 10^{-0.0151} = e^{-0.0346}$

Así si queremos hacer una predicción para $X = 12$ se tendrá:

$$\log y = 1.986 - 0.0151 \times 12 = 1.805$$

Otras transformaciones muy usuales son las que aparecen al tratar de ajustar los siguientes tipos de funciones:

$$y' = \frac{1}{\alpha + \beta x}$$

y la función potencia: $y = \alpha . X^{\beta}$

La primera de ellas representa una relación lineal entre X_i y $1/y_i$:

$$\frac{1}{y_i} = \alpha + \beta X$$

y la segunda representa una relación lineal entre $\log X_i$ y $\log y_i$:

$$\log y' = \log \alpha + \beta \log X$$

Ahora bien, si no existe una clara indicación acerca de la forma funcional de regresión de y sobre X usualmente supondremos que puede expanderse mediante una Serie de Taylor y que los primeros términos de dicha expansión nos darán una buena aproximación y entonces ajustamos nuestros datos a un polinomio de la forma:

$$y' = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^p$$

donde el grado se determina mediante inspección de los datos, ó mediante (métodos científicos que discutiremos enseguida):

Para estimar los coeficientes β_i deberemos minimizar

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots \beta_p X_i^p \right) \right]^2$$

En otras palabras, aplicamos el criterio de los mínimos cuadrados y minimizamos la suma de los cuadrados de las distancias verticales de los puntos a la curva.

Si b_i son los estimadores de β_i, derivando obtenemos las siguientes $p+1$ ecuaciones:

$$\frac{\partial S}{\partial b_0} = 2\left[\sum y_i - \sum\left(b_0 + b_1 X_i + b_2 X_i^2 + \dots + b_p X_i^p\right)\right] = 0$$

$$\sum y = n b_0 + b_1 \sum X_i + b_2 \sum X_i^2 + \dots + b_p \sum X_i^p$$

$$\frac{\partial S}{\partial b_1} = 2\sum\left[y_i - \left(b_0 + b_1 X_i + b_2 X_i^2 + \dots + b_p X_i^p\right)\right] X_i$$

$$\sum xy = b_0 \sum X + b_1 \sum X^2 + b_2 \sum X^3 + \dots + b_p \sum X^{p+1}$$

etc.

Arreglando en forma matricial:

$$\begin{pmatrix} n & \sum X & \sum X^2 & \dots & \sum X^p \\ \sum X & \sum x^2 & \sum X^3 & \dots & \sum X^{p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X^p & \sum X^{p+1} & \sum X^{p+2} & \dots & \sum X^{2p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum X y \\ \vdots \\ \sum X^p y \end{pmatrix}$$

Ejemplo de ajuste de polinomios.

X	0	1	2	3	4	5	6	7	8	
y	12.0	10.5	10.0	8.0	7.0	8.0	7.5	8.5	9.0	

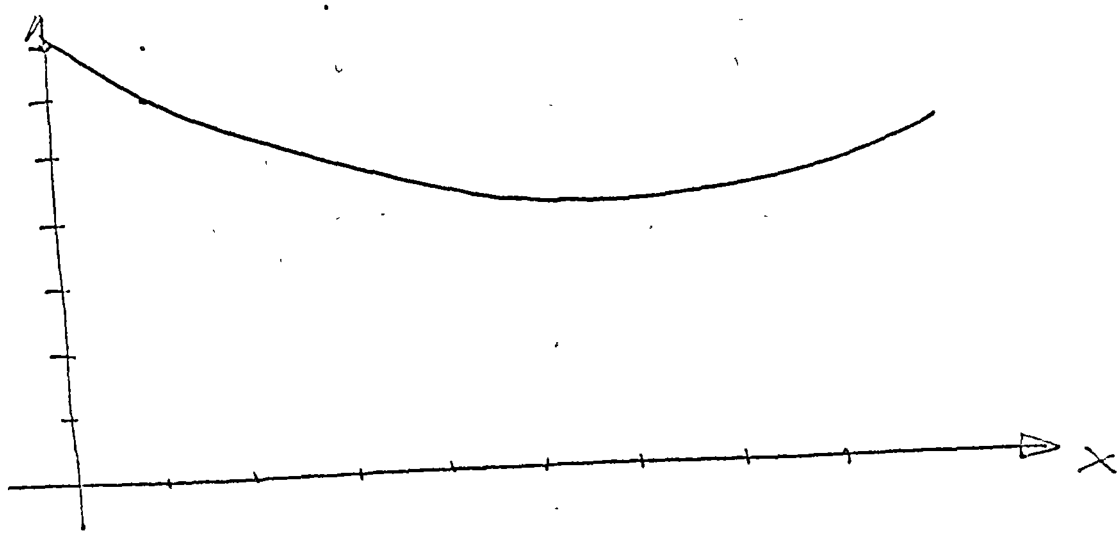
$\Sigma X = 36$; $\Sigma X^2 = 204$; $\Sigma X^3 = 1296$

$\Sigma X^4 = 8772$; $\Sigma y = 80.5$; $\Sigma Xy = 299.0$

$\Sigma X^2 y = 1697.0$

Obteniendo: $b_0 = 12.2$; $b_1 = -1.85$; $b_2 = 0.183$

$\Rightarrow y' = 12.2 - 1.85X + 0.183X^2$



REGRESIÓN MÚLTIPLE.

Antes de pasar a extender los métodos de las secciones anteriores a problemas que involucran más de una variable independiente conviene aclarar que las curvas obtenidas (y las superficies que obtendremos) no sólo se usan para hacer predicciones. Se usan también para resolver problemas de optimización, o sea para determinar para qué valor o valores de la variable (o variables) independiente, la variable dependiente alcanza un máximo o un mínimo.

En la Regresión Múltiple tratamos con datos consistentes de $n(r+1)$ valores:

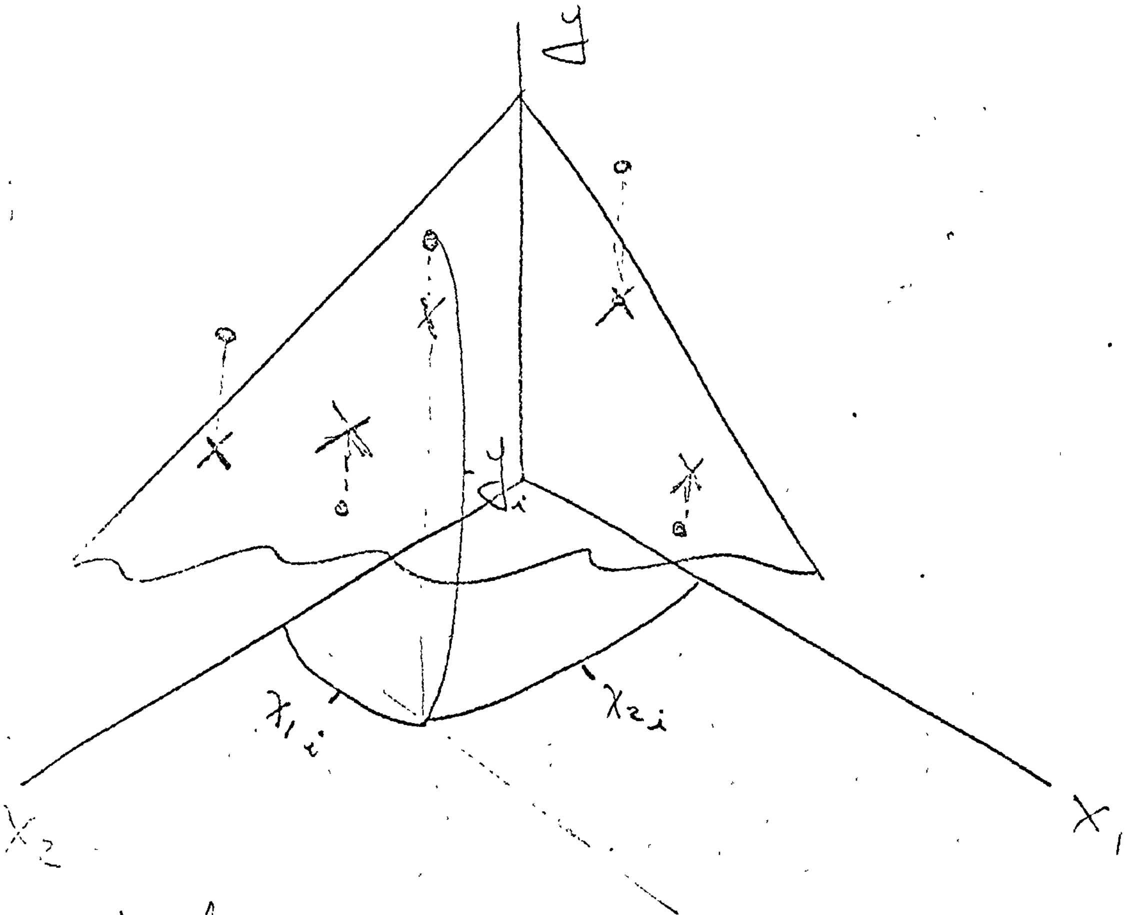
$(x_{1i}, x_{2i}, \ldots, x_{ri}, y_i)$, donde las X se suponen conocidas sin error, mientras que las y's son valores de v.a.

Datos de este tipo se usan para, por ejemplo, para determinar el efecto que sobre el precio de un producto tienen: el nivel de producción de la fábrica, el nivel de consumo y los inventarios.

Tratemos primeramente el caso lineal, o sea cuando para cualquier conjunto de valores $x_1, x_2, \ldots, x_r$; la media de la distribución de las y's está dada por:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r$$

[illegible] var. ind. es el problema de ajustar un plano a un conjunto de n puntos con coordenadas (X_{1i}, X_{2i}, y_i)



Aplicando el método de los mínimos cuadrados para minimizar la suma de los cuadrados de las distancias verticales de los puntos al plano:

$$\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})\right]^2$$

Las ecuaciones normales resultantes son:

$$\Sigma y = n b_0 + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

$$\Sigma X_1 y = b_0 \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2$$

$$\Sigma X_2 y = b_0 \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2$$

Nótese que la notación es abreviada, p. ejemplo:

$$\sum X_1 = \sum_{i=1}^{n} X_{1i} \quad ; \quad \sum X_1 X_2 = \sum_{i=1}^{n} X_{1i} X_{2i}$$

$$\sum X_1 y = \sum_{i=1}^{n} X_{1i} y_i$$

Ejemplo.

y Numero de vueltas	X_1 % del elemento A	X_2 % del elemento B
38	1	5
40	2	5
85	3	5
59	4	5
40	1	10
60	2	10
68	3	10
53	4	10
31	1	15
35	2	15
42	3	15
59	4	15
18	1	20
34	2	20
29	3	20
42	4	20

$\sum X_1 = 40$; $\sum X_2 = 200$; $\sum X_1^2 = 120$

$\sum X_1 X_2 = 500$; $\sum X_2^2 = 3000$; $\sum y = 733$; $\sum X_1 y = 1989$

$\sum X_2 y = 8285$

Sustituyendo en las ecuaciones normales:

$$723 = 16 b_0 + 40 b_1 + 200 b_2$$

$$1989 = 40 b_0 + 120 b_1 + 500 b_2$$

$$8285 = 200 b_0 + 500 b_1 + 3000 b_2$$

$$y' = 48.2 + 7.83 X_1 - 1.76 X_2$$

CORRELACIÓN:

Hasta ahora hemos estudiado los problemas en donde la variable o variables se han supuesto conocidas sin error. Aunque este es el caso en muchas situaciones experimentales, existen también problemas donde tanto las X_i como las y_i son variables aleatorias. Este es el caso, por ejemplo, de las relaciones entre: precipitación vs gasto, resistencia a la tensión vs dureza, etc, impureza en el aire vs incidencia de ciertas enfermedades, etc.

Problemas como estos se conocen como problemas de análisis de correlación.

Los puntos dados (X_i, y_i) $(i=1,2,\ldots,n)$ son pares de valores de v.a. cuya densidad de probabilidad conjunta es $f(x,y)$.

Veamos que entendemos por densidad de prob. conjunta.

Tenemos: $P(A \cap B) = P(A) \cdot P(B/A)$

$$f(x,y) = g(x)\, f(y/x)$$

donde $f(y/x)$ define la densidad de probabilidad condicional de la la segunda v.a para X fija. Aquí $g(x)$ es la densidad marginal.

DENSIDADES DE PROBABILIDAD CONJUNTAS.

Si $X_1, X_2, \ldots X_K$ son los valores de K v.a., nos referiremos a la función $f(x_1, x_2, \ldots, X_K)$ como la densidad de probab. conjunta de estas v.a. si la probabilidad de que

$a_1 \leq X_1 \leq b_1$ y $a_2 \leq X_2 \leq b_2$, ... y, $a_K \leq X_K \leq b_K$

está dada por la integral múltiple

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} \cdots \int_{a_K}^{b_K} f(x_1, x_2, \ldots, X_K)\, dX_1 dX_2 \ldots dX_K$$

Nótese que si $f(x_1, x_2, \ldots, X_K) \geq 0$ para todos los valores de $X_1, X_2, \ldots, X_K$ para los que la densidad de probabilidad está definida y

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots X_K)\, dX_1 dX_2 \ldots dX_K = 1$$

esta definición es consistente con la definición de función densidad de probabilidad.

Si la función $F(x_1, X_2, \ldots X_K)$ nos representa valor $\leq x_1$, la segunda v.a tome un valor $\leq x_2$ ... y la Késima v.a. tome un valor $\leq x_K$, nos referiremos a la función F como la función de distribución conjunta de las K v.a.

Ejemplo: Consideremos la densidad de prob. conjunta dada por:

$$f(x_1, x_2) = \begin{cases} 6e^{-2x_1-3x_2} & \text{para } x_1>0,\ x_2>0 \\ 0 & \text{para otro caso.} \end{cases}$$

$$P(1 \leq x_1 \leq 2 \ y \ 2 \leq x_2 \leq 3) = \int_1^2\int_2^3 6e^{-2x_1-3x_2}\,dx_1\,dx_2 = (e^{-2}-e^{-4})(e^{-6}-e^{-9})$$

$$= 0.0003$$

$$P(0 \leq x_1 \leq 2 \ y \ 2 \leq x_2 \leq \infty) = \int_0^2\int_2^\infty 6e^{-2x_1-3x_2}\,dx_1\,dx_2 = (1-e^{-4})e^{-6} = 0.0025$$

$$F(x_1, x_2) = \begin{cases} \int_0^{x_1}\int_0^{x_2} 6e^{-2u-3v}\,du\,dv & x_1, x_2 > 0 \\ 0 & 0 \text{ en otro caso.} \end{cases}$$

Definición. Diremos que K v.a. son independientes si y sólo si:

$$F(x_1, x_2, \ldots, x_k) = F_1(x_1) \cdot F_2(x_2) \cdot \ldots \cdot F_K(x_K)$$

Donde $F_i(x_i)$ para $i = 1, 2, \ldots, K$ es el valor correspondiente de la función de distribución de la iésima v.a.

De allí se sigue que:

$$f(x_1, \ldots x_K) = f_1(x_1)\, f_2(x_2) \ldots f_K(x_K)$$

donde $f_i(x_i)$ para $i = 1, \ldots K$ es el valor correspondiente de la densidad de probab. de la iésima v.a.

En general, la densidad de probabilidad de la iésima v.a. puede obtenerse a partir de la densidad de prob. conjunta, o sea:

$$f_i(X_i)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,x_2,\ldots,x_k)\,dx_1\ldots dx_{i-1}\,dx_{i+1}\cdots dx_k$$

en donde $f_i(x_i)$ se llama la densidad marginal de la iésima v.a. Así por ejemplo, la densidad marginal está dada para el ejemplo anterior, y para la 1ª variable por:

$$f_1(x_1)=\int_0^{\infty}6e^{-2x_1-3x_2}\,dx_2=2e^{-2x_1}$$

para $x_1>0$ y $f_1(x_1)=0$ en otro caso.

Si una v.a. toma el valor $g(x_1,x_2,\ldots x_k)$ cuando k v.a. toman los valores $x_1,x_2,\ldots,x_k$ $\Rightarrow$ el promedio o media de la 1ª v.a. se define por la integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(x_1,x_2,\ldots,x_k)\,f(x_1,x_2\ldots x_k)\,dx_1\,dx_2\ldots dx_k$$

Por ejemplo, si nos interesara el valor medio del producto de 2 v.a. en el ejemplo numérico anterior, obtendríamos:

$$\int_0^{\infty}\int_0^{\infty}x_1x_2\,6e^{-2x_1-3x_2}\,dx_1\,dx_2=\frac{1}{6}$$

Hacemos las mismas hipótesis para $f(y|x)$ que las supuestas en la primera parte, o sea, para cualquier X se supondrá que $f(y|x)$ es una dist. normal con media $\alpha+\beta X$ y variancia σ^2. Entonces, la regresión de y sobre X es lineal y la variancia de la densidad condicional no depende de X, más aún, supondremos que la densidad marginal $g(x)$ es normal con media μ_1 y variancia σ_1^2.

Sustituyendo las expresiones para $f(y|x)$ y $g(x)$, obtenemos:

$$f(x,y) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-(x-\mu_1)^2/2\sigma_1^2} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[y-(\alpha+\beta x)]^2/2\sigma^2}$$

$$= \frac{1}{2\pi\sigma_1\sigma} e^{-\left\{[y-(\alpha+\beta x)]^2/2\sigma^2 + (x-\mu_1)^2/2\sigma_1^2\right\}}$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$

Nótese que esta distribución de probabilidad conjunta involucra 5 parámetros $\mu_1, \sigma_1, \alpha, \beta$ y σ.

Por razones de simetría y otras consideraciones que explicaremos adelante es costumbre expresar la función de densidad normal bivariada $[f(x,y)]$ en términos de $\mu_1, \sigma_1, \mu_2, \sigma_2$ y ρ. Aquí μ_2 y σ_2^2 son la media y la variancia de la distribución marginal $h(y)$, mientras que ρ, llamado coeficiente

IPESA INGENIERIA Y PROCESAMIENTO ELECTRONICO. S. A.
AV. MORELOS 110-207 MEXICO I, D. F.
46-03-00 46-61-41

de correlación se define como:

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_2^2}$$

con ρ tomando valores positivos cuando β es positivo ($y < 0$ cuando $\beta < 0$).

Puede demostrarse que:

$$\mu_2 = \alpha + \beta\mu_1 \quad y \quad \sigma_2^2 = \sigma^2 + \beta^2\sigma_1^2$$

sustituidos en la expresión anterior para $f(x, y)$ y obtenemos:

$$f(x, y) = \frac{1}{2\pi \cdot \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \, e^{-\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] / 2(1-\rho^2)}$$

para $-\infty < X < \infty$ ... y ... $-\infty < y < \infty$

Respecto al coeficiente de correlación ρ, nótese que $-1 \leq \rho \leq +1$, pues:

$$\sigma_2^2 = \sigma^2 + \beta^2\sigma_1^2$$

y así $\sigma_2^2 \geq \sigma^2$

Más aún, ρ puede ser igual a 1 o −1 sólo cuando $\sigma^2 = 0$, que representa el caso degenerado cuando toda la probabilidad se concentra a lo largo de la línea $y = \alpha + \beta x$ y se tiene entonces un una relación lineal perfecta entre las 2 v.a.; esto es, para un valor dado de X, y debe ser $= \alpha + \beta X$

El coeficiente de correlación es igual a cero $\Leftrightarrow$

$$\sigma^2 = \sigma_2^2$$

y se sigue de: $\sigma_2^2 = \sigma^2 + \beta^2 \sigma_1^2$

que eso sólo es posible cuando $\beta = 0$

$\Rightarrow \rho = 0$ implica que la recta de regresión de y sobre X es horizontal y ∴ que el conocimiento de X no ayuda a la predicción de y.

$\Rightarrow \rho = \pm 1 \Rightarrow$ correlación lineal perfecta entre la 2 v.a

$\rho = 0 \Rightarrow f(x,y) = f(x)\, f(y)$ (Las 2 v.a son independientes)

- Para valores de ρ entre 0 y 1: $0 < \rho < 1$

interpretamos ρ refiriéndonos a la identidad

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_2^2} = \frac{\sigma_2^2 - \sigma^2}{\sigma_2^2}$$

Puesto que σ^2 es una medida de la variación de y's cuando X es conocida mientras que σ_2^2 es una medida de la variación de las y's cuando X es desconocida

Así ρ^2 nos dice la proporción de las variaciones de las y's que pueden ser atribuidas a la relación lineal con X

Dada una muestra aleatoria de tamaño n (esto es, n pares de valores) (x_i, y_i), se acostumbra estimar ρ en términos del coeficiente de correlación de la muestra

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

donde S_{xy}, S_{xx} y S_{yy} se definen como antes

Ejemplo:

X Esfuerzo	y Modulo de Elasticidad
8.76	889
7.08	808
11.35	1513
5.25	550
10.91	1008
13.56	1952
6.48	908
11.83	1654
15.08	1502
7.13	1216
9.08	1315
3.86	497

$\Sigma x = 110.38$ $\qquad \Sigma y = 13\,812$ $\qquad \Sigma xy = 141673.33$

$\Sigma x^2 = 1143.8090$ $\qquad \Sigma y^2 = 18132296$

$$S_{xx} = 12(1143.8090) - (110.38)^2 = 1541.9636$$

$$S_{xy} = 12(141\,673.33) - (110.38)(13182) = 175\,511.40$$

$$S_{yy} = 12(18132,296) - (13182)^2 = 26\,816\,208$$

$$r = \frac{175\,511.40}{\sqrt{(1541.9636)(26\,816\,208)}} = 0.86$$

Lo que significa que $100r^2 = 74\%$ de la variación en el módulo de elasticidad puede atribuirse a diferencias en el esfuerzo.

Siempre que un valor de r se base en datos de la muestra es recomendable ejecutar una prueba de significancia (una prueba con hipótesis nula $\rho=0$) con objeto de determinar cuando podemos tener un valor alto en la muestra y sin embargo no tener una relación entre las 2 variables.

Para muestras aleatorias extraídas de una población binormal, esta prueba se basa en el parámetro

$$\frac{1}{2}\ln\frac{1+r}{1-r}$$

Puede demostrarse que la distribución muestral de ese parámetro es aproximadamente normal con media

$$\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$$

y variancia $\frac{1}{n-3}$

$\Rightarrow$ probamos la hipótesis $\rho=0$ con el parámetro

$$Z = \frac{\sqrt{n-3}}{2}\ln\frac{1+r}{1-r}$$

cuya distribución muestral es aproximadamente la distribución normal estandarizada.

Regresando al ejemplo:

$$Z = \frac{\sqrt{9}}{2}\ln\frac{1+0.86}{} = 3.88$$