UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN
COORDINACIÓN DE MATEMÁTICAS

ESTADISTICA II

-- GUÍA DE ESTUDIO --

Tema: 1 Teoría elemental del muestreo
Presentación e introducción
1. Introducción al muestreo
2. Diferentes tipos de muestreo
3. Estimación de parámetros

Tema II. Distribuciones muestrales e intervalos de confianza para la media poblacional.

1. Distribución de muestreo de la media
2. Medición de la precisión de un estimado a través de la muestra.
3. Distribución de las diferencias de las medias de dos muestras independientes.
4. Intervalos de confianza para la proporción.
5. Distribución de las diferencias de las proporciones de dos muestras independientes.
6. Determinación del tamaño de la muestra.

Tema III. Pruebas de hipótesis.

1. Etapas básicas en pruebas de hipótesis.
2. Pruebas de hipótesis según el tamaño de la muestra.
3. Pruebas de hipótesis para la media de una y dos poblaciones.

Tema IV. Estadística no paramétrica.

1. Características de las pruebas no paramétricas.
2. La distribución JI cuadrada.
3. Pruebas de bondad de ajuste.
4. Tablas de contingencia.
5. Pruebas de los signos.
6. Prueba de rachas.

Tema V. Análisis de regresión lineal.

1. Análisis de regresión simple.
2. Método de mínimos cuadrados.
3. Inferencias relativas a la pendiente de la recta de regresión.
4. Predicción de un valor particular de “y” para un valor dado de “x”.
5. Coeficiente de correlación y coeficiente de determinación.
6. Examen

Tema VI. series de tiempo.

1. Análisis de tendencia.
2. Variación cíclica.
3. Variación temporal.
4. Variación irregular.
5. Análisis en predicciones.

Introducción.

Predicción[1]:

El deseo de predecir el futuro es una característica inherente al ser humano. No obstante, la necesidad de hacer predicciones fiables[2] en los negocios va más allá de la curiosidad. Así por ejemplo:

- Las decisiones de inversión deben tomarse mucho antes de que un nuevo producto salga al mercado, por tanto es muy deseable tener predicciones sobre cómo será la situación del mercado en el futuro próximo.
- Para productos ya establecidos, hacer predicciones sobre las ventas a corto plazo es importante para establecer los niveles óptimos de acumulación de inventarios y producción.
- Para tomar una decisión sobre aumentar o no el nivel de pasivos de una empresa, es importante predecir los tipos de interés en el futuro.

---

[1] Paul Newbold. “***Estadística para los negocios y la economía***”. Editorial: Prentice Hall. P.p 4

[2] fiable.
1. adj. Dícese de la persona a quien se puede fiar, o de quien se puede responder; por ext., se aplica también a las cosas que ofrecen seguridad. (DRAE)

- Para formular una política económica razonable, los gobiernos necesitan predicciones sobre cuál sería el producto interno bruto (**PIB**) el desempleo y la inflación bajo varias políticas diferentes.

Básicamente, las predicciones de valores futuros suelen obtenerse a partir del descubrimiento de regularidades en el comportamiento en el pasado. Por esta razón, es necesario disponer de datos sobre el comportamiento, tanto de la variable a predecir, como de otras variables relacionadas. El análisis de esta información puede sugerir tendencias en el futuro.

Toma de decisiones en un entorno de incertidumbre[3].

En cualquier tipo de negocio, deben tomarse constantemente decisiones en un entorno en el que la persona que debe decidir no conoce con seguridad el comportamiento futuro de los factores que podrían afectar el resultado que se obtendría bajo varias opciones posibles a considerar. Por ejemplo:

- Cuando un fabricante presenta una oferta para un contrato, no está completamente seguro de los costos futuros que le ocasionará hacer frente a su oferta. Es más, tampoco conocerá las ofertas de sus competidores. A pesar de esta incertidumbre[4], la decisión debe tomarse. Y

- Cuando un inversor decide cómo equilibrar su cartera de acciones, bonos y otros instrumentos financieros, no conoce los movimientos futuros del mercado. Puede tener alguna idea sobre futuros desarrollos, pero no puede predecir con exactitud qué ocurrirá.

Estos ejemplos demuestran que, en los negocios, en el momento de decidir entre diferentes opciones, resultan de vital importancia las técnicas para tratar la incertidumbre.

En las presentes notas, veremos una serie de técnicas útiles a la hora de analizar información numérica. Su objetivo es ayudar a comprender los entornos con incertidumbre, de forma que puedan tomarse mejores decisiones.

Hay que hacer hincapié, no obstante, en que estas técnicas son únicamente herramientas útiles para el administrador. No pretenden ser sustitutos de la familiaridad con el entorno que se consigue con años de trabajo y experiencia, sino más bien ayudas para agudizar dicha familiaridad.

Por tanto, a pesar de que un análisis técnico profundo de la información numérica será, en ocasiones, de mucho valor, no se aprovechará al máximo si no se utiliza en combinación con la experiencia que se obtiene de estudiar las característica del entorno en el que se trabaja. De hecho, los métodos estadísticos resultan de mayor utilidad en la gestión[5] cuando se combinan con la experiencia en el entorno de los negocios.

---

[3] Paul Newbold. "***Estadística para los negocios y la economía***". Editorial: Prentice Hall. P.p 4

[4] incertidumbre.
1. f. Falta de certidumbre; duda, perplejidad. (DRAE)

[5] gestión. (DRAE)
Del lat. gestio, -onis.
1. f. Acción y efecto de gestionar.
2. [f.]Acción y efecto de administrar.
de negocios.
1. Der. Cuasi contrato que se origina por el cuidado de intereses ajenos sin mandato de su dueño.

Capítulo 1 Teoría elemental del Muestreo.

Presentación e introducción

1. introducción al muestreo
2. diferentes tipos de muestreo
3. estimación de parámetros

Teoría del muestreo[6].

> La teoría del muestreo estudia la relación entre una población y las muestras tomadas de ella.

La teoría del Muestreo es de gran utilidad en muchos campos. Por ejemplo, hoy por hoy, todos sabemos que Vivimos[7] en un mundo en el que la mayoría de los países hacen enérgicos esfuerzos para aumentar el nivel de vida de sus poblaciones, con el fin de lograr el desarrollo equilibrado se elaboran planes detallados y se ejecutan en la medida de los posible. Para elaborar esos planes de una manera científica es necesario disponer de los hechos básicos en términos numéricos para las diferentes regiones del país y para éste como un todo.

Los recursos de los países pequeños no bastan para recopilar datos, año tras año de cada persona, empresa o institución. Por fortuna como sabemos ahora, no es indispensable incluir cada una de las unidades del universo para llegar a una cifra aceptable para el total.

Así, tenemos la aplicación de los métodos de muestreo a problemas prácticos que enfrentan los países en desarrollo, lo que ha sido de igual importancia para la creación de sistemas nacionales de Estadística. Por lo general, esos países no tienen una tradición larga de censos o de recopilaciones periódicas similares que puedan usar como contexto de su muestreo. Por lo tanto, en esas condiciones la aplicación de métodos de muestreo que produzcan resultados aceptables requiere de gran ingenio.

Una **muestra** cuidadosamente diseñada puede proporcionar la información necesaria para establecer los lineamientos que requiere un país[8], a un costo que este último podría muy posiblemente absorber.

Es decir, la teoría del muestreo se utiliza para **estimar** magnitudes desconocidas de una población, tales como la media y la varianza, llamadas a menudo **parámetros de la población** o simplemente **parámetros**, a partir del conocimiento de esas magnitudes sobre muestras, que se llaman **estadísticos de la muestra** o simplemente **estadísticos**.

La teoría del muestreo es útil también para determinar si las diferencias observadas entre dos muestras son debidas a variaciones fortuitas o si son

---

[6] Murray R. Spiegel. **"Estadística"** serie Schaum. Editorial McGraw-Hill. P.p 186
[7] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 9
[8] **NOTA:** La oficina de Estadística de las Naciones Unidas se dedica entre otras funciones a encontrar la forma y los medios de ayudar a los gobiernos nacionales en la obtención de los datos estadísticos tan indispensables para la planificación del desarrollo económico y social, controlar la ejecución real de los programas y evaluar los resultados.

realmente significativas. Tales cuestiones aparecen, por ejemplo, al probar un nuevo suero como tratamiento de una enfermedad o al decidir si un proceso de producción es mejor que otro. Las respuestas implican el uso de los llamados **contrastes (o tests) de hipótesis y de significación**, que son importantes en la **teoría de las decisiones**.

En general, un estudio de las inferencias hechas sobre una población a partir de muestras suyas, con indicación de la precisión de tales inferencias, se llama **Inferencia Estadística**.

FUNDAMENTOS DE LA TEORÍA DEL MUESTREO[9].

La teoría de las probabilidades es el fundamento de los métodos de muestreo y no oculto este hecho. Un buen conocimiento de:

- Álgebra
- Cálculo y
- Probabilidades

Desde el punto de vista de la matemática y de:

- Los métodos generales de estadística y de la
- Teoría básica de las estimaciones

Desde el punto de vista estadístico es esencial para un entendimiento adecuado del desarrollo riguroso de la Teoría del Muestreo.

Variables aleatorias[10].

Supongamos que hay un experimento aleatorio que genera un espacio muestral con sus puntos muestrales $E_{1,}$ $E_{2,...}$ y probabilidades asociadas $P_r(E_1)$, $P_r(E_2)$,...ahora se definirá una función en este espacio muestral.

Supongamos que hay una regla por la cual un número "U" está asociado con cada punto del espacio muestral. De conformidad con dicha regla, asignamos los números reales: $U_{1,}$ $U_{2,}$ ... a los puntos $E_{1,}$ $E_{2,...,}$ respectivamente. Reuniendo todos los puntos con los cuales está asociado el numero "U_i " formamos el evento $U = u_i$ que interpretamos como: "la variable aleatoria "U" toma el valor de "u_i ".

El conjunto de relaciones:

$$P_r(U = u_i) = g(u_i), \qquad \sum g(u_i) = 1, \qquad (i = 1, 2,...)$$

Define la distribución de probabilidad de la variable aleatoria "U".

[9] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 11
[10] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 16

El teorema del límite central[11]:

La aplicación del teorema del límite central o teorema central del límite a la distribución muestral de las medias de muestras, que vimos con anterioridad, permite utilizar la distribución de probabilidad normal para crear intervalos de confianza para la media de la población.

El teorema del límite central afirma que, para grandes muestras aleatorias, la distribución muestral de las medias de muestras está más próxima a una distribución de probabilidad normal. La aproximación es más precisa para muestras grandes. Ésta es una de las conclusiones más útiles en Estadística. Es posible razonar sobre la distribución muestral de las medias de muestras sin contar con información alguna sobre la forma de la distribución original de la que se toma la muestra. En otras palabras, el teorema del límite central es válido para todas las distribuciones.

El enunciado formal del teorema del límite central es el siguiente:

Teorema del límite central.

Si en cualquier población se seleccionan muestras de un tamaño específico, la distribución muestral de las medias de muestras es aproximadamente un distribución normal. Esta aproximación mejora con muestras de mayor tamaño.

ley de los grandes números[12]:

la ley de los grandes números sugiere que la probabilidad de una desviación significativa de un valor de probabilidad determinado empíricamente[13], a partir de una determinado teóricamente, es menor cuanto más grande sea el número de repeticiones del experimento[14].

Muestreo[15].

Es el proceso para obtener información acerca del conjunto de una población o universo examinando solo una parte del mismo.

---

[11] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 234

[12] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA primera reimpresión, México 1998. p.p 158

[13] empírico, ca. Del lat. empiricus, y este del gr. mpeirikÕj, que se rige por la experiencia.
1. adj. Relativo a la experiencia o fundado en ella. (DRAE)

[14] Jacob Bernoulli (1654-1705) fue uno de los primeros que estudiaron la probabilidad matemática. En su libro "Ars Conjectandi" (1713) apareceríó la primera proposición de la ley de los grandes números. En su honor, su nombre va asociado a varios conceptos matemáticos, como los experimentos de Bernoulli en probabilidad, los números de Bernoulli en la teoría de los números y la lemniscata de Bernoulli en cálculo. (nota tomada del libro: "Probabilidad y Estadística" de Stephen S. Willoughby. Publicaciones culturales, s.a. p.p 100.

[15] Notas tomadas durante el curso: El muestreo Estadístico aplicado a la Auditoria; impartido por el M.C José Refugio Ruiz Piña. (dirección general de asuntos del personal académico. Programa de actualización académica para profesores de licenciatura. Octubre del 2001)

ENCUESTA POR MUESTREO[16].

LA FUNCIÓN DEL MÉTODO DE MUESTREO[17].

En la actualidad se ha llegado a considerar la encuesta por muestreo como un instrumento organizado para encontrar hechos. Su importancia para la civilización moderna radica en que puede utilizarse para resumir, a fin de orientar a la administración, hechos que de otra manera serían inaccesibles debido: a la lejanía y oscuridad de las personas o de las otras unidades de que se trate, o a su gran número. La encuesta por muestreo permite que se tomen decisiones que tienen en cuenta los factores significativos de los problemas que se procura resolver.

Como instrumento para descubrir hechos, **la encuesta por muestreo no se ocupa principalmente de la interpretación económica o sociológica de los hechos que demuestra**, aunque debiera proporcionar información adecuada para esas interpretaciones. Más bien se ocupa de la adecuada representación de los hechos individuales registrados y de su recopilación y resumen.

Un muestreo[18] aleatorio simple es un proceso en el cual cada muestra posible de un tamaño dado tiene la misma probabilidad de ser seleccionada. Obtener una muestra verdaderamente aleatoria, o al menos aproximadamente aleatoria, requiere de cierto raciocinio y esfuerzo. Una muestra aleatoria no es una muestra casual o desordenada. La población objetivo se debe identificar. En principio, se debería elaborar una lista de todos los elementos de la población y seleccionar aleatoriamente aquellos que estarán incluidos en la muestra, utilizando una tabla de números aleatorios.

El muestreo[19] se debe considerar siempre que se quiera tener información y el costo (en dinero, en trabajo o en tiempo) de obtener la información completa es excesivo.

Metodología de muestreo[20].

Es importante hacer notar que la Metodología de muestreo debe quedar plasmada en los papeles de trabajo correspondientes.

---

[16] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 36

[17] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 36

[18] Hildebrand David. K y Lyman Ott R. "Estadística aplicada a la administración y a la economía" editorial: Addison Wesley Longman. P.p. 230

[19] ibid. P.p. 231

[20] Notas tomadas durante el curso: "El muestreo estadístico aplicado a la auditoria" impartido por el Maestro en Ciencias: José Refugio Ruiz Piña. Octubre del 2001.

1. **definir el objetivo.**

2. **seleccionar el plan de muestreo adecuado.**
   a) Muestreo de atributos.
   b) Muestreo de suspensión o continuación.
   c) Muestreo de variables.
   d) Muestreo de descubrimiento.
   e) Muestreo dirigido.

3. **definir el nivel de confianza y precisión deseada.**
   a) Determinar el tamaño de la muestra.

4. **seleccionar la muestra.**
   a) Aleatoria.
   b) Sistemática o intervalos.
   c) Estratificada.
   d) Conglomerados.
   e) Automatizada.

5. **realizar las pruebas.**

6. **determinar estadísticos.**

7. **evaluar resultados.**

Ejemplo de Muestreo de Atributos[21]:

La función del muestreo de atributos, es determinar "cuantos elementos" existen. Y se utiliza para estimar la frecuencia probable con la cual ocurre un determinado evento.
Donde este evento puede ser una clase de error u otro atributo de la población. Es aplicable cuándo el propósito de una auditoria puede lograrse mediante una respuesta de "sí" o "no", "bueno" o "malo", "blanco" o "negro", etc.

Ejemplo:

Supóngase que el propósito de la prueba de auditoria estriba en determinar cuántos errores se cometieron en números de identificación asignados a la facturación corriente de la empresa en la que laboramos.

El auditor determina que el tamaño de la población es de 10 000 facturas y que la tasa de error esperada no deberá ser mayor del 5%. Desea contar con un nivel de confianza del 90% de que los resultados de sus pruebas se hallan

[21] "Folleto de curso: "el muestreo estadístico aplicado a la auditoria" ; impartido por el M.C José Refugio Ruiz Pina. Octubre del 2004. pag 25.

En este caso particular, el tamaño de la muestra resulta de 140, i.e n = 140, obtenido de la tabla correspondiente al cuadro A5 (pagina 55). Por lo tanto es necesario revisar de acuerdo con este criterio, un total de 140 facturas, lo cual necesariamente implica la obtención mediante el procedimiento necesario de los 140 elementos que forman la muestra.

5. realizar las pruebas.

Los datos del problema son:

N = 10 000 facturas
M.e = 10%
N.C = 90%
P = 0.5
Q = 0.5

Recuerda que el muestreo de atributos nos sirve para determinar si las cosas se están haciendo bien de acuerdo a cierto criterio establecido.

Y dado que la muestra debe ser aleatoria, entonces, la fórmula a utilizar es[1]:

$$N.F = \left[Li + (Ls - Li)(N.A)\right] Entero$$

en donde:

N.F = número o folio de la factura a revisar
Li = límite inferior de la población (en este caso es 1)
Ls = límite superior de la población (en este caso es 10 000)
N.A = número aleatorio comprendido entre 0 y 1
Entero = significa que de toda la operación, lo único que nos interesa es la parte entera, desdeñando la parte decimal.

El N.A (número aleatorio) lo podemos elegir de una tabla de números aleatorios. Para su elección podemos utilizar el criterio de:

- los números de serie de un billete cualquier denominación.
- Los números de serie de un boleto del sistema de transporte colectivo "metro"
- Los números que vienen en una tarjeta de crédito o débito, etc.

Por lo tanto, de las 10 paginas que forman la tabla de números aleatorios, elegimos una al azar. (pag. 83) y de ella en base a los número obtenidos del billete, del boleto o de la tarjeta, elegimos una fila y una columna, en nuestro caso podría ser por ejemplo, la fila 133 columna 4 de donde obtenemos el número aleatorio: 09244, que al ser considerado decimal nos queda como: 0.09244

[1] Ésta fórmula se basa en una distribución de tipo uniforme (discreta)

Una vez elegido el número aleatorio de inicio (también llamado “semilla”) podemos elegir continuar hacia:

- Arriba
- Abajo
- En diagonal derecha
- En diagonal izquierda, etc.

De la semilla. Hasta completar el número de facturas que sea necesario revisar de acuerdo al tamaño de la muestra.

Y procedemos a sustituir datos en la fórmula. Así:

$$N.F = [1 + (10000 - 1)(0.09244)]Entero$$

N.F = 925.31

Por lo tanto, teniendo en consideración sólo la parte entera, debemos revisar la factura número: 925

6. determinar estadísticos.

A continuación hacemos una tabla para poder visualizar mejor cuáles serán las facturas a revisar: (continuando hacia debajo de la semilla plantada)

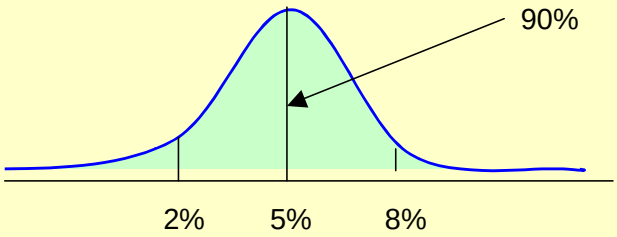
# de experimento	# aleatorio	# de factura a revisar
01	0.09244	925
02	0.11592	1160
03	0.32402	324
04	0.12021	1202
05	.	.
06	.	.
.	.	.
.	.	.
.		
.		

7. **evaluar resultados.** Si una vez que hemos obtenido la muestra completa (las 140 facturas) revisamos la misma y encontramos 8 facturas mal elaboradas, entonces, este número de facturas representa de acuerdo a una "regla de tres":

140----------100%
8----------X %

de donde podemos observar que: **X = 5.7[1]%**

y si este valor lo graficamos en el intervalo de precisión, podemos ver que este valor (x=5.7%) se encuentra dentro de dicho intervalo de precisión construido



luego entonces, el auditor estará seguro de que 90 veces de cada 100 la tasa de error se encuentra entre el 2% y el 8%; i.e que no es necesario tomar alguna medida correctiva.

Nota:

Si revisamos la tabla del cuadro F.3 (pag. 93) de los límites de precisión revisados con base en la tasa de error hallada en la muestra, tendríamos que los límites de la precisión son:

Si n = 140, N = 10 000 y la tasa de error de la muestra es de 5.7% ≈ 5%

$Li_{(prec)}$= 2.4%
$Ls_{(prec)}$=9.3%

O bien del cuadro F.4 (pag. 94) para n = 140, N = 10 000 y una tasa de error de la muestra de 5.7% ≈ 10%, los límites del intervalo de precisión son:

$Li_{(prec)}$= 6.2%
$Ls_{(prec)}$=15.2%

[1] si la tasa encontrada en la muestra es mayor del 5% y esta fuera de los límites de la precisión, es común reevaluar la precisión o recalcular el tamaño de muestra (aumentarlo), utilizando para ambos casos los cuadros F. (nota: esto incluso es aplicable para una tasa encontrada de 0%). Por lo tanto si usted desea verse conservador y no andar recalculando el tamaño de la muestra o la precisión, puede utilizar como alternativa el calculo del tamaño de la muestra mediante la fórmula genérica. Cuyo resultado casi siempre es mayor que el de las tablas.

Por lo cual, lo más que podemos decir es que: contablemente hablando, un intervalo de precisión de 2% a 8% es razonable[1]. Y que después del estudio realizado, la facturación actual se encuentra dentro de los límites razonables de la tasa de error estipulada.

Pero, si el tamaño de la muestra lo determinamos de acuerdo con la fórmula genérica para este caso, tendremos que:

Para la determinación del tamaño de la muestra se requiere[22]:

1. tamaño del universo.
2. tasa de error esperada.
3. homogeneidad-heterogeneidad del fenómeno.
4. precisión o margen de error.
5. exactitud o nivel de confianza.
6. número de estratos.
7. etapas de muestreo.
8. conglomeración de unidades.
9. estado del marco muestral.
10. efectividad de la muestra.
11. técnica de recolección de datos. Y
12. recursos disponibles.

Dentro de la teoría del muestreo y probabilidad existen diversos procedimientos para el cálculo de los tamaños de la muestra: todos ellos consideran los elementos que hemos enumerado. A continuación se presenta una fórmula genérica para el cálculo del tamaño de muestra. Las variables que considera la fórmula son los siguientes:

Variable	Descripción
n	Tamaño de la muestra
N	Tamaño del universo
P	Probabilidad de ocurrencia (homogeneidad del fenómeno)
Q	Probabilidad de no ocurrencia (1-p)
Me	Margen de error o precisión. Expresado como probabilidad.
Nc	Nivel de confianza o exactitud. Expresado como valor **z** que determina el área de probabilidad buscada.

La fórmula utilizada es la siguiente:

$$n = \frac{NPQ}{\left[\frac{Me^2}{Nc^2}(N-1)\right] + PQ}$$

por ejemplo: supongamos que queremos calcular el tamaño de una muestra para el siguiente caso:

Variable	Descripción
n	?
N	3,000,000
P	Desconocemos la probabilidad de ocurrencia. Por esta razón asumimos el mayor punto de incertidumbre, que es de 50%, que al ser expresada como probabilidad queda como: 0.5
q	1 – 0.5 = 0.5
Me	+/- 5% de margen de error. Que expresado como probabilidad queda

[2] s. Jesús. “Técnicas de Investigación en sociedad, cultura y comunicación” editorial: Addison Wesley Longman

Así, para nuestro ejemplo en cuestión tendríamos que

N = 10 000 facturas
M.e ≈ 10%
N.C = 90%
P = 0.5
Q = 0.5

que sustituidos en la fórmula correspondiente nos darían:

$$n = \frac{(10{,}000)(0.5)(0.5)}{\left[\frac{(0.05)^2}{(1.96)^2}(10{,}000-1)\right]+(0.5)(0.5)}$$

de donde finalmente vemos que es necesario revisar un total de: **369.98** facturas.

Finalmente podemos observar que utilizando las tablas para determinar el tamaño de la muestra, esta resulta ser más pequeña que la que resulta de calcularla utilizando la fórmula genérica; ambos procedimientos son válidos y su uso se restringe a las condiciones imperantes al momento de realizar el estudio.

Cabe mencionar que en la actualidad existen numerosos paquetes de software que son utilizados principalmente por los despachos contables para determinar tanto el tamaño de la muestra como los elementos a revisar dependiendo del tipo de muestreo seleccionado.

Ejercicio propuesto[1]:

Supóngase una prueba en la cual el auditor espere una tasa de error no mayor del 5% en una población de 20 000 artículos, una precisión de $\pm$ 3% y un nivel de confianza del 95%. Determine el tamaño de la muestra y todos los elementos muestrales.

1 Tomado del curso: “el muestreo estadístico aplicado a la auditoria”. Impartido por el M.C José Refugio Ruiz Piña. Octuble del 2001. pag. 26

TIPOS DE MUESTREO:

MUESTREO APLICANDO CRITERIO[23].

Para la obtención de información sobre la base de una muestra, el estadístico (persona que estudia la estadística) de encuestas rechaza de antemano ciertos procedimientos. Esto ocurre cuando no es posible encontrar un método objetivo para diferenciar un procedimiento de otro. Por ejemplo:

- Se podría obtener información sin mucho gasto preguntando a personas expertas en un determinado campo. Sin duda esos expertos tendrán opiniones diferentes, y no hay ningún método objetivo para diferenciar entre sus opiniones.
- Otro procedimiento que pertenece a esta categoría consiste en limitar el muestreo a unidades que parecen ser representativas de la población que se considera[24]. Se obtiene información sobre esas unidades y con base en la misma se hacen estimaciones sobre las características de la población. También en este caso el criterio de la persona que selecciona la muestra es importante, porque personas diferentes tendrán criterios diferentes.

No hay un método objetivo para preferir un criterio u otro. No podemos predecir el tipo de distribuciones de los resultados producidos por un gran número de seleccionadores de muestra que aplican su criterio, ni podemos predecir cómo diferirán del denominado "verdadero" valor que se busca. **No conocemos ningún método objetivo para medir la confianza que debe tenerse en los resultados cuando la muestra es seleccionada por criterio**. La razón es que con esos métodos nos e conoce la probabilidad de que una determinada unidad sea seleccionada en el muestreo. Por lo tanto, no podemos estimar la distribución de frecuencia de las estimaciones de este procedimiento (muestreo por criterio). En ausencia de información sobre cómo diferirán las diferentes muestras entre si, el error de muestreo no puede determinarse objetivamente.

---

[23] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 36

[24] un ejemplo es el **muestreo por cuotas**, en el que los entrevistadores quedan en libertad de seleccionar sus informantes siempre que la muestra se refiera a "x" número de hombres y "x" número de mujeres, a determinado número de personas con elevados ingresos y otro tanto de bajos ingresos, etc.)

MUESTREO PROBABILISTICO[25].

El panorama cambia tan pronto como empezamos a utilizar un procedimiento de muestreo en el que todas las unidades pertenecientes a una población tienen una probabilidad conocida (que no es igual a cero) de ser seleccionadas en la muestra. Con la ayuda de la teoría de las probabilidades estamos entonces en la posición de determinar la distribución de frecuencia de las estimaciones derivables del procedimiento de muestreo y de estimación. Podemos calcular la proporción de estimaciones dentro de un intervalo especificado en torno al llamado valor "verdadero" buscado. Sabemos los resultados que producirá la repetición de un determinado procedimiento de muestreo, y esto nos permite diferenciar entre los diversos procedimientos. Además, lo que es muy importante, puede obtenerse una medida de la variación muestral (la medida en que las estimaciones de la muestra diferirán del promedio) de una manera objetiva a partir de la muestra misma. Todo el **corpus** de la Teoría de la Probabilidad y la Inferencia Estadística (basada en la primera) están disponibles para desprender conclusiones válidas a partir de la muestra.

[25] Raj, Des. "Teoría del Muestreo". Fondo de cultura económica. P.p 41

Muestreo aleatorio simple[26]:

El tipo que más se utiliza es un muestreo aleatorio simple.

Muestreo aleatorio simple:

Consiste en una muestra seleccionada de modo que cada uno de los elementos o personas en la población tengan las mismas probabilidades de ser incluidos.

Para ilustrar el muestreo aleatorio simple, suponga que una población consta de 845 empleados y se ha de seleccionar una muestra de 52 empleados de esa población.

Una forma de asegurar que todos los empleados tengan la misma oportunidad de ser elegidos es escribir primero el nombre de cada no de ellos en un papel y colocar todos los papeles en una urna. Luego de haberlos revuelto en forma minuciosa, se hace la primera selección tomando un papel de la caja sin mirarla. Este proceso se repite hasta haber seleccionado la muestra de 52 personas.

Un método más conveniente de seleccionar una muestra aleatoria es usar el número de identificación de cada empleado y una **tabla de números aleatorios**.

Un muestreo[27] aleatorio simple es un proceso en el cual cada muestra posible de un tamaño dado tiene la misma probabilidad de ser seleccionada. Obtener una muestra verdaderamente aleatoria, o al menos aproximadamente aleatoria, requiere de cierto raciocinio y esfuerzo. Una muestra aleatoria no es una muestra casual o desordenada. La población objetivo se debe identificar. En principio, se debería elaborar una lista de todos los elementos de la población y seleccionar aleatoriamente aquellos que estarán incluidos en la muestra, utilizando una tabla de números aleatorios. **Ejemplo 6.1 pag. 230 (del Hildebrand)**

---

[26] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 223

[27] Hildebrand, David, K. & Lyman Ott. R. "Estadística aplicada a la administración y a la economía". Editorial: Addison wesley Longman. P.p. 230

Muestreo[28] aleatorio Sistemático[29]:

El procedimiento del muestreo aleatorio simple puede ser difícil en ciertos casos. Por ejemplo, suponga que la población que nos interesa consiste de 2000 facturas que se localizan en cajones. Tomar una muestra aleatoria sencilla requeriría primero numerar las facturas, del 0001 al 1999. utilizando una tabla de números aleatorios, se seleccionaría luego una muestra de, por ejemplo 100 números, luego, en los cajones deberá localizarse una factura que concuerde con cada uno de estos 100 números. Esta tarea puede requerir mucho tiempo. En lugar de ello, se podría seleccionar una **muestra aleatoria sistemática**[30] recorriendo simplemente los cajones, contando las facturas y tomando todas las que hagan el número 20 del grupo, para su estudio. Así, la primera factura debería elegirse utilizando un proceso aleatorio: por ejemplo, una tabla de números aleatorios. Si se eligió la décima factura como punto de partida, la muestra consistiría en las facturas: décima, trigésima, quincuagésima, septuagésima, etc

Debido a que el primer número se elige al azar, todos tienen la misma probabilidad de seleccionarse para la muestra. Por lo tanto, se trata de un muestreo probabilístico.

En resumen:

Para un muestreo aleatorio sistemático.

Se acomodan los elementos o personas de la población en cierta forma. Se selecciona un punto de partida aleatorio y luego se toma cada K-ésimo miembro para formar la muestra.

[28] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 225
[29] En un mestreo sistemático, el primer artículo se elige al azar.
[30] En ciertas circunstancias, una muestra sistemática podrá producir resultados sesgados.

Muestreo aleatorio estratificado[31]:

Otro tipo de muestreo probabilístico es el muestreo aleatorio estratificado[32].

Muestreo aleatorio estratificado.

Se divide una población en subgrupos llamados estratos, y se selecciona una muestra de cada uno de ellos.

Una vez que la población se divide en estratos, es posible seleccionar una **muestra proporcional** o **no proporcional**. Como el nombre lo implica, un procedimiento de muestreo proporcional requiere que el número de artículos de cada estrato esté en la misma proporción que en la población.

Por ejemplo, el problema podría ser estudiar los gastos de publicidad de las 352 empresas Mexicanas más grandes. Suponga que el objetivo des estudio consiste en determinar si las empresas con altos rendimientos sobre su inversión (una medición de la rentabilidad) han gastado una mayor proporción de su presupuesto de ventas en publicidad que las empresas que tienen un menor rendimiento o incluso un déficit.

Suponga que las 352 empresas se dividieron en 5 estratos y si seleccionamos una muestra de 50 empresas, entonces se deberían incluir:

Estrato	Rentabilidad	# empresas	# muestreado	?
1	30% y más	8	**1**	(8/352)(50)
2	De 20 a 30%	35	**5**	(35/352)(50)
3	De 10 a 20%	189	**27**	(189/352)(50)
4	De 0 a 10%	115	**16**	(115/352)(50)
5	Déficit	5	**1**	(5/352)(50)
	Total	352	**50**	

En una muestra estratificada no proporcional, la cantidad de artículos que se seleccionan en cada estrato no guarda proporción con los números respectivos en la población.

En algunos casos, el muestreo estratificado tiene la ventaja de poder reflejar con mayor precisión las características de la población que un muestreo aleatorio simple o sistemático.

Muestreo por Conglomerados[33]:

[31] Una muestra estratificada garantiza la representación de cada subgrupo.
[32] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 226
[33] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 227

Otro tipo de muestreo que es común, es el muestreo por conglomerados[34]. Muchas veces se le emplea para reducir el costo de realizar un muestreo de una población dispersa en una gran área geográfica. Suponga que se desea determinar el punto de vista de los industriales de toda la República Mexicana con respecto a las reformas fiscales del año 2002. la selección de una muestra aleatoria de los industriales de toda la República Mexicana y el contacto personal con cada uno de ellos serían muy onerosos en cuanto a tiempo y dinero. En lugar de ello, se podría emplear un muestreo por conglomerados subdividiendo la República Mexicana en unidades pequeñas, ya fueran estados o regiones. Muchas veces, éstas se conocen como **unidades primarias**. suponga que se subdividió a la República Mexicana en 12 unidades primarias y luego se escogió a cuatro de ellas, de esta forma, los esfuerzos se concentran en estas cuatro unidades, tomando una muestra aleatoria de los industriales de cada una de estas regiones y entrevistarlos (observe que se trata de una combinación del muestreo por conglomerados y el muestreo aleatorio simple).

[34] el muestreo por conglomerados reduce el costo del muestreo.

Muestreo con medidas estadísticas[35].

Un método eficaz de muestreo necesita, al menudo, algo más que objetividad: requiere de algún medio para establecer tamaños de muestra y evaluar matemáticamente los resultados obtenidos de ella. Esto se logra con una **muestra estadística** o **muestra probabilística**. Este tipo de muestra tendrá un comportamiento mensurable en función de las reglas de la teoría de la probabilidad.

Con una muestra estadística, es posible afirmar, con un determinado **grado de confianza**, que el resultado de la muestra no se aleja de las condiciones reales del universo, más allá de cierto **límite especificado**.

Ventajas[36].

- ✓ Los resultados de la muestra pueden ser justificados objetivamente.
- ✓ Proporciona un medio para conocer con anticipación el tamaño máximo necesario de la muestra.
- ✓ Suministra una estimación de la magnitud del riesgo de que la muestra pueda no ser representativa de toda la población.
- ✓ Puede ser más exacto que el que se realiza examinando cada uno de los elementos de una población grande.
- ✓ Las muestras estadísticas suelen ser más económicas que los tamaños de muestra tradicionales.
- ✓ Proporciona un medio de proyectar los resultados de las pruebas dentro de límites conocidos de confianza.

[35] Notas tomadas durante el curso: "El muestreo estadístico aplicado a la auditoria" impartido por el Maestro en Ciencias: José Refugio Ruiz Piña. Octubre del 2001.

[36] Notas tomadas durante el curso: "El muestreo estadístico aplicado a la auditoria" impartido por el Maestro en Ciencias: José Refugio Ruiz Piña. Octubre del 2001.

Conceptos básicos[37].

Nivel de confianza. Es el grado en el que se justifica estimar que una muestra aleatoria indica el verdadero valor del universo (dentro de una amplitud estipulada).

Por ejemplo: un 95% de N.C (nivel de confianza) quiere decir que hay 95 posibilidades entre 100 de que los resultados de la muestra representen las condiciones verdaderas del universo.

Precisión. Es la amplitud (expresada como más o menos un porcentaje determinado) dentro de la cual debe encontrarse la respuesta verdadera concerniente a las características (errores por ejemplo) de la población que se estudia, con un determinado nivel de confianza.

En otras palabras, es el grado de exactitud del supuesto de que el número de errores de la muestra se aplica proporcionalmente a la parte no muestreada de la población.

Por ejemplo. Sí con base en una prueba se afirma que la tasa de error proyectada en un universo dado es 5%, +-2%, se está diciendo que la tasa de error en la muestra examinada fue exactamente de 5%, en tanto que la precisión en la muestra (con un nivel de confianza especificado) era de +-2%. Es decir, la tasa puede ser tan pequeña como el 3% o tan grande como el 7%.

[37] Notas tomadas durante el curso: "El muestreo estadístico aplicado a la auditoria" impartido por el Maestro en Ciencias: José Refugio Ruiz Piña. Octubre del 2001.

Tipos de muestras[38]:

En general podemos decir que existes dos tipos de muestras, a saber:

- Muestras probabilísticas y
- Muestras no probabilísticas.

De ambas, aquellas que nos interesan son las muestras probabilísticas, pues los resultados de un muestreo no probabilístico pueden estar sesgados. Por lo tanto, nos surge la pregunta: ¿Qué es una muestra probabilística?

Muestra probabilística:

Es una muestra seleccionada de tal forma que cada artículo o persona dentro de la población tiene la misma probabilidad (distinta de cero) de ser incluida en la muestra.

No existe un método "mejor" para seleccionar una muestra probabilística de una población de interés. Quizá el método que se utilizó para seleccionar una muestra de facturas de un cajón no sea el idóneo para elegir una muestra nacional de votantes. Sin embargo, **todos los métodos de muestreo probabilistico tiene similar finalidad: permitir que el azar determine los artículos o personas que incluye la muestra.**

Muestras aleatorias y números aleatorios[39].

Para que las conclusiones de la teoría del muestreo y de la inferencia estadística sean válidas, las muestras deben ser "**representativas**" de la población. El análisis de los métodos de muestreo y problemas relacionados se llama el **diseño del experimento**.

Una forma de obtener una muestra representativa es mediante un **muestreo aleatorio**, de acuerdo con el cual, cada miembro de la población tiene la misma probabilidad de ser incluido en la muestra. Existen al menos dos métodos para lograr obtener una muestra representativa; a saber

1. El primer método consiste en asignar un número a cada elemento de la población, escribir dicho número en una papeleta, y realizar un sorteo justo con ellas en una urna.
2. Un método alternativo consiste en recurrir a una **tabla de números aleatorios**.

---

[38] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 222

[39] Murray R. Spiegel. "Estadística" serie Schaum. Editorial: Mc Graw-Hill. P.p 186

Error en el muestreo[40]:

En el análisis anterior se acentuó la importancia de seleccionar una muestra a fin de que todos los artículos de la muestra tengan la misma oportunidad de ser elegidos. Para lograr esto, se puede seleccionar:

- ✓ una muestra aleatoria simple;
- ✓ una muestra sistemática;
- ✓ una muestra estratificada;
- ✓ una muestra por conglomerados o
- ✓ una combinación de estos métodos.

Sin embargo, es improbable que la media de la muestra fuera **idéntica** a la media de la población. Así mismo, tal vez la desviación estándar u otra medición que se calcule con base en la muestra no sea **exactamente igual** al valor correspondiente de la población. Así, es posible que existan diferencias entre las estadísticas de la muestra, como la media o la desviación estándar de la muestra, y los parámetros de la población correspondientes. La diferencia entre un estadístico de la muestra y un parámetro de la población se conoce como **error de muestreo.**

Error de muestreo.

Es la diferencia entre un estadístico y el parámetro correspondiente.

[40] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 229

CAPITULO II. DISTRIBUCIONES MUESTRALES Y TEOREMA DEL LIMITE CENTRAL

II.1. DISTRIBUCIONES MUESTRALES.

Consideremos todas las muestras posibles de tamaño "N" en una población dada (con o sin reposición). Para cada muestra podemos calcular un estadístico (tal como la media o la desviación típica) que variará de muestra a muestra. De esta manera obtenemos una distribución del estadístico que se llama su **distribución de muestreo**.

Si por ejemplo, el estadístico utilizado es la media muestral, entonces la distribución se llamaría la **distribución de muestreo de medias**, o **distribución de muestreo de la media**. Análogamente, podríamos tener distribuciones de muestreo de la desviación típica, de la varianza, de la mediana, de las proporciones, etcétera.

Para cada distribución de muestreo podemos calcular la media, la desviación típica, etc. Así pues, podremos hablar de la media y la desviación típica de la distribución del muestreo de medias, etcétera.

II.2. DISTRIBUCIÓN MUESTRAL DE LAS MEDIAS DE LAS MUESTRAS[1]:

Una vez que se descubrió la posibilidad del error de muestreo cuando se utilizan los resultados de la muestra para estimar el parámetro de una población:

- ¿Cómo es posible hacer una predicción precisa sobre el éxito de una pasta de dientes de reciente desarrollo con base sólo en los resultados de la muestra?
- ¿De qué manera puede el departamento de control de calidad de una empresa que se dedica a la producción masiva liberar un embarque de microprocesadores, con base en una muestra de sólo diez unidades?
- ¿Cómo puede Gallup o Harris hacer una predicción precisa de una votación presidencial con base en una muestra de sólo 2000 votantes registrados, de una población de casi 90 millones de votantes?

Para responder a estas preguntas, se examina la distribución muestral de las medias de la muestra.

Al organizar las medias de todas las muestras posibles de un cierto tamaño en una distribución de probabilidad, se obtiene una **distribución muestral de las medias de las muestras.**

---

[1] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 230

Distribución muestral de las medias de las muestras.

Es la distribución de probabilidad de todas la medias posibles de las muestras de un tamaño de muestra dado.

Por ejemplo[2]:

El número de unidades producidas por un obrero que trabaja de lunes a sábado en una fábrica que produce "latas" para refresco es la siguiente: 80, 80, 76, 70, 70 y 68. Suponga que estos números constituyen la población de la cual se desea tomar una muestra de tamaño 3.

a) determine la media aritmética de estos números.
b) Determine la desviación estándar de los números.
c) Calcule el número de muestras de tamaño 3
d) Liste cada una de las muestras
e) Calcule la media de cada una de las muestras.
f) Encuentre la media de la distribución de las medias de las muestras.
g) Calcule la desviación estándar de las medias de las muestras.
h) Compare los resultados de los incisos a y f
i) Compare los resultados de los incisos b y g.

Problema tomado con ligeros cambios del libro: "Probabilidad y Estadística" de stephen S. Willoughby. Editorial: Publicaciones cultural s.a. p.p 126

[2] problema tomado con ligeros cambios del libro: "Probabilidad y Estadística" de Stephen S. Willoughby. Editorial: Publicaciones cultural s.a. p.p 126

Solución al problema propuesto:

a) para encontrar la media aritmética de los numero solicitada, procedemos a utilizar la fórmula correspondiente, tomando en consideración de que si se trata de una población, entonces el símbolo a utilizar es: μ, por lo tanto:

$$\mu = \frac{1}{N}\sum_{1}^{n} x_i$$

de donde sustituyendo datos tenemos que:

$$\mu = \frac{1}{6}[80+80+76+70+70+68]$$

$\mu = 74$ **solución al a)**

b) para el inciso b, es recomendable elaborar la tabla indicada a continuación:

# de experimento	Datos	media aritmética	Dato - media	(dato - media)elevado al cuadrado
i	x_i	μ	$(x_i - \mu)$	$(x_i - \mu)^2$
1	80	74	6	36
2	80	74	6	36
3	76	74	2	4
4	70	74	-4	16
5	70	74	-4	16
6	68	74	-6	36
Sumatoria	444		0	144

En esta tabla podemos observar que la sumatoria de la columna correspondiente a la diferencia del dato menos la media, es cero, por lo tanto, hasta ese punto nuestro proceso es correcto.

Finalmente para este inciso, aplicamos la fórmula correspondiente:

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

de donde sustituyendo valores tenemos que:

$$\sigma = \sqrt{\frac{1}{6}(144)}$$

σ = 4.9 **respuesta al b)**

c) para dar respuesta a este inciso, debemos aplicar la fórmula correspondiente al cálculo de combinaciones. Es decir:

$$C_r^n = \frac{n!}{r!\,(n-r)!}$$

de donde sustituyendo valores tenemos que:

$$C_r^n = \frac{6!}{3!(6-3)!}$$

$$C_r^n = \frac{6x5x4x3!}{3!(3x2x1)}$$

de donde fácilmente vemos que el número de combinaciones de 6 objetos tomados de 3 en 3 es:

$C_r^n = 20$ **respuesta al c)**

d) para dar respuesta a este inciso, es necesario realizar los siguientes pasos:

1. identificar cada uno de los datos. En nuestro caso, en virtud de que algunos datos se repiten se procede a identificarlos de las siguiente manera: **80_1, 80_2, 76, 70_1, 70_2, 68.**

2. a continuación, se colocan estos datos en forma horizontal, es decir de la siguiente forma:

   80_1, 80_2, 76, 70_1, 70_2, 68.

3. como siguiente punto, se elabora una tabla donde se colocaran todas las combinaciones obtenidas siguiendo el orden indicado a continuación: la primera terna o combinación se obtiene de los tres primero datos, es decir:

   Si los datos son: **80_1, 80_2, 76, 70_1, 70_2, 68.**

   Entonces, la primera terna es: **80_1, 80_2, 76,**

   para la segunda terna, se toman los dos primeros datos junto con el cuarto dato (es decir, no saltamos el tercer dato), por lo tanto:

   recordemos que los datos son: **80_1, 80_2, 76, 70_1, 70_2, 68.**
   Entonces, la segunda terna sería: **80_1, 80_2, 70_1.**

Para la tercera terna se hace lo mismo, sólo que en este caso utilizamos los dos primeros datos más el quinto dato, y así sucesivamente hasta que cubrimos todos los datos que se encuentran a la derecha de los dos primeros datos. Mediante este procedimiento, obtenemos las siguientes ternas:

Para este caso también recordemos que los datos son:

$$80_1, 80_2, 76, 70_1, 70_2, 68.$$

80_1	80_2	76
80_1	80_2	70_1
80_1	80_2	70_2
80_1	80_2	68

Continuando con este procedimiento, nos “saltamos” el segundo dato, continuando con el tercero y cuarto dato; es decir, la siguiente terna tendría la forma siguiente:

Recordemos que los datos son: $80_1, 80_2, 76, 70_1, 70_2, 68.$

$$80_1, 76, 70_1$$

siguiendo este procedimiento, podemos encontrar fácilmente las siguientes ternas:

tengamos siempre presente los datos: $80_1, 80_2, 76, 70_1, 70_2, 68.$

80_1	*76*	*70_1*
80_1	76	70_2
80_1	76	68
80_1	70_1	70_2
80_1	70_1	68
80_1	70_2	68

Una vez que hemos terminado con todas las posibles combinaciones que empiezan con el primer dato, nos continuamos de la misma forma para el segundo dato; mediante este procedimiento podemos encontrar todas las restantes combinaciones, que son:

En este caso también y por comodidad recordemos los datos:

$80_1, 80_2, 76, 70_1, 70_2, 68.$

80_2	76	70_1
80_2	76	70_2
80_2	76	68
80_2	70_1	70_2
80_2	70_1	68
80_2	70_2	68
76	70_1	70_2
76	70_1	68
76	70_2	68
70_1	70_2	68

e) para calcular la media de cada una de las muestras, conviene elaborar una tabla donde estén incluidas todas las muestras de tamaño 3 encontradas, por lo tanto elaboramos la siguiente tabla, donde fácilmente podemos calcular la media de cada una de las muestras requerida.

	M U E S T R A S			Media
1	80_1	80_2	76	78 2/3
2	80_1	80_2	70_1	76 2/3
3	80_1	80_2	70_2	76 2/3
4	80_1	80_2	68	76
5	80_1	76	70_1	75 1/3
6	80_1	76	70_2	75 1/3
7	80_1	76	68	74 2/3
8	80_1	70_1	70_2	73 1/3
9	80_1	70_1	68	72 2/3
10	80_1	70_2	68	72 2/3
11	80_2	76	70_1	75 1/3
12	80_2	76	70_2	75 1/3
13	80_2	76	68	74 2/3
14	80_2	70_1	70_2	73 1/3
15	80_2	70_1	68	72 2/3
16	80_2	70_2	68	72 2/3
17	76	70_1	70_2	72
18	76	70_1	68	71 1/3
19	76	70_2	68	71 1/3
20	70_1	70_2	68	69 1/3

f) si ahora consideramos el conjunto de todas las medias de las muestras como un nuevo conjunto al que podemos llamar **distribución de las medias de las muestras**, fácilmente podemos calcular la media de la distribución de las medias de las muestras, para lo cual procedemos a aplicar la formula correspondiente:

$$\mu_{\bar{x}} = \frac{1}{N}\sum_{1}^{n} x_i$$

de donde sustituyendo datos tenemos que:

$$\mu_{\bar{x}} = 69$$

g) Calcule la desviación estándar de las medias de las muestras.

Continuar con el ejercicio sobre el cálculo de la media de las medias. Willoughby Pág. 128 $\mu_{\bar{x}} = \bar{x}$

a) Compare los resultados de los incisos a y f
b) Compare los resultados de los incisos b y g.

Al desarrollar el ejercicio en el que calculamos la media de las medias, podemos observar en términos generales que:

- La media de las medias de la muestra es igual a la media de la población.
- La dispersión de la distribución de las medias de la muestra es menor a la dispersión en los valores de la población.
- La forma de la distribución muestral de las medias de muestras y la forma de la distribución de frecuencia de los valores de la población es diferente. La distribución de las medias de las muestra tiende a tener una forma de campana y a aproximarse a la distribución de probabilidad normal.

En resumen se tomaron todas las muestras aleatorias posibles de una población y para cada muestra se calculó un estadístico de muestra (la media). Debido a que cada muestra posible tiene la misma posibilidad de ser seleccionada, se puede determinar la probabilidad de que la media obtenida tenga un valor comprendido en un rango. La distribución de los valores de las medias obtenidas se conoce como distribución muestral de las medias de muestras.

Aunque en la práctica sólo se ve una muestra aleatoria específica, en teoría podría surgir cualquiera de las muestras. En consecuencia, el proceso de muestreo repetido genera la distribución muestral. Luego, la distribución muestral se utiliza para medir lo probable que podría ser obtener un resultado específico.

En este caso debemos tomar en consideración lo siguiente: Supongamos que se toman todas las posibles muestras de tamaño "**n**" sin reposición, de una población finita de tamaño $N > n$. Si denotamos la media y la desviación típica de la distribución de muestreo de medias por: $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$ y las de la población por μ y σ, respectivamente, entonces:

$$\mu_{\bar{x}} = \mu \qquad \text{y} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ \mu$$

si la población es infinita o si el muestreo es con reposición, los resultados anteriores se reducen a:

$$\mu_{\bar{x}} = \mu \qquad \text{y} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

para valores grandes de "**n**" ($n \geq 30$), la distribución de muestreo de medias es aproximadamente normal con media $\mu_{\bar{x}}$ y desviación típica $\sigma_{\bar{x}}$ independientemente de la población (en tanto en cuanto la media poblacional y la varianza sean finitas y el tamaño de la población sea al menos el doble que el de la muestra). Este resultado para una población infinita es un caso especial del **teorema central del límite** de la teoría avanzada de probabilidades, que afirma que la precisión de la aproximación mejora al crecer "**n**". Esto se indica en ocasiones diciendo que la distribución de muestreo es **asintóticamente normal**.

En caso de que la población esté normalmente distribuida, la distribución de muestreo de medias también lo está, incluso para pequeños valores de "n" (o sea, n<30).

II.3. DISTRIBUCIÓN MUESTRAL DE LA PROPORCIÓN.

Suponga el ejemplo de una comercializadora que pretende establecer un nuevo centro, y desea saber la proporción del consumidor potencial que compraría su principal producto que vende, para lo cual realiza un estudio de mercado, consultando de una encuesta de 30 participantes, quienes lo comprarían y quienes no, obteniéndose los siguientes resultados:

$x_1 = 1$	$x_7 = 1$	$x_{13} = 0$	$x_{19} = 1$	$x_{25} = 0$
$x_2 = 0$	$x_8 = 0$	$x_{14} = 1$	$x_{20} = 0$	$x_{26} = 0$
$x_3 = 0$	$x_9 = 0$	$x_{15} = 1$	$x_{21} = 1$	$x_{27} = 0$
$x_4 = 0$	$x_{10} = 0$	$x_{16} = 0$	$x_{22} = 1$	$x_{28} = 1$
$x_5 = 0$	$x_{11} = 0$	$x_{17} = 0$	$x_{23} = 1$	$x_{29} = 0$
$x_6 = 1$	$x_{12} = 0$	$x_{18} = 1$	$x_{24} = 0$	$x_{30} = 1$

Donde 1 significa que si está dispuesto a comprar el producto y 0 no está dispuesto a comprarlo.

En este caso, la proporción de la población P que compraría el producto, se puede estimar con $\bar{p}$ (proporción de la muestra que lo compraría), cuyo valor esperado sería $E(\bar{p}) = P$, y el error de $\bar{p}$ al estimar P es:

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$$

si la población es finita, y si la población es infinita o si el muestreo es con reposición, los resultados anteriores se reducen a:

$$\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}}$$

Es decir, de acuerdo al teorema del límite central, $\bar{p}$ muestral se comportará como una normal con media P (la verdadera proporción poblacional) y desviación estándar $\sigma_{\bar{p}}$.

En el ejemplo de la comercializadora se tiene que $\bar{p} = \frac{12}{30} = 0.40$.

Pero suponiendo que el verdadero parámetro de la población es P = 0.30, es decir sólo el 30% de la población lo compraría, entonces el promedio $\bar{p}$ estimará a P poblacional pero con un error igual a $\sigma_{\bar{p}}$ que en este caso es:

$$\sigma_{\bar{p}} = \sqrt{\frac{0.30(0.70)}{30}} = 0.1195$$

Y en este caso $\bar{p}$ muestral tendrá distribución normal con media P=0.30 y desviación estándar $\sigma_{\bar{p}} = 0.1195$.

II.4. DISTRIBUCIÓN t de Student, X^2 Ji-Cuadrada y F de Fisher.

Cuando se hace inferencia estadística, muchas de las veces es necesario determinar la distribución de los estadísticos muestrales como $\bar{x}$ o como S^2 (la varianza muestral):

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

Pues conociendo la distribución podremos hacer algunas estimaciones de los parámetros poblacionales como media, varianza, proporción, etc..

En este sentido la teoría de la estadística, así como la ley de los grandes números que nos dice que al sumar un número considerable de variables aleatorias, la suma se aproxima a una distribución normal, ambas nos dan una

respuesta respecto del modelo de distribución. Así que es necesario mencionar como es que se obtienen los modelo de distribución t , X^2 y F.

Los tres modelos de distribución anteriores, se infieren a partir de la distribución normal estándar (de media cero y varianza uno).

Distribución X^2

Considere X_1, X_2,…, X_n n-variables aleatorias normales estándar, las cuales tienen distribución particular y totalmente conocida. Entonces la variable Y:

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2$$

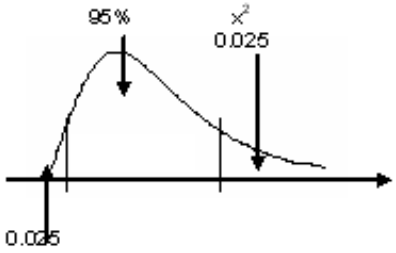
Tendrá distribución X^2 (Ji-Cuadrada) con n-grados de libertad.

¿Como se interpreta esto?.

Lo anterior equivale a un resultado probabilistico que nos dice que si se extrae una muestra aleatoria de tamaño n de una distribución normal estándar y con ella obtenemos la suma de sus cuadrados , el resultado tendrá una distribución totalmente conocida llamada X^2.

Los grados de libertad indican el número de variables que se están sumando y que se refiere a que tanta libertad tiene la variable *Y* para tomar valores, por ejemplo si n=20 y *Y*=10 entonces este último número puede provenir de infinidad de valores por ejemplo: $X_1^2 = X_2^2 = X_3^2 = \ldots = X_{20}^2 = 1$, o tantas combinaciones de las X_i las cuales pueden considerarse de 20 posibles valores distintos.

A saber la gráfica de un modelo X^2 es:



Así que un estadístico comúnmente utilizado para estimar la varianza de una población es:

$$\frac{(n-1)S^2}{\sigma^2}$$

El cual tiene una distribución X^2 con n-1 grados de libertad.

Este estadístico es útil porque su expresión únicamente tiene como parámetro desconocido a la varianza poblacional σ^2 por lo cual si se deseará determinar un intervalo de estimación para σ^2 se podrá hacer a través de dicho estadístico, como se verá en el siguiente capítulo.

Distribución *t-Student*

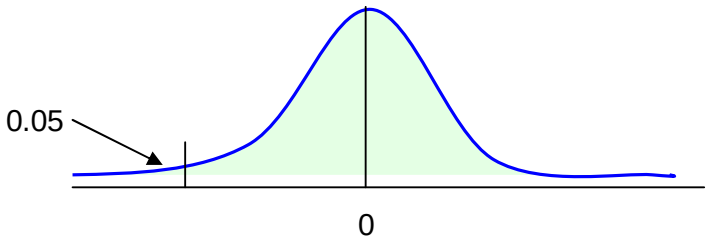
Esta distribución se deduce también a partir de la distribución normal estándar, como se indica a continuación:

Considere que se extrae una muestra aleatoria de una distribución normal estándar, obteniéndose X y X_1, X_2,…,X_n. Entonces la variables resultado Y:

$$Y = \frac{X}{\sqrt{X_1 + X_2 + ... + X_n} / \sqrt{n}}$$

tendrá distribución t de student con n-grados de libertad.

Así, para un valor particular de tamaño de muestra n, la distribución t presenta una gráfica en particular, que es:



Sin embargo si el tamaño de muestra es mayor a 30 entonces la distribución t es casi una normal estándar.

En este caso, un estadístico comúnmente utilizado para estimar la media de una población es:

$$\frac{\overline{X} - \mu}{S / \sqrt{n}}$$

El cual tiene una distribución t con n-1 grados de libertad.

Este estadístico es útil porque su expresión únicamente tiene como parámetro desconocido a la media poblacional μ.

La demostración se hace a partir del hecho que:

$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ tiene distribución normal estándar.

Y el estadístico

$\frac{(n-1)S^2}{\sigma^2}$ tiene distribución Ji-cuadrada con n-1 g.l.

Por lo cual el estadístico

$$\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/\sqrt{n-1}}}$$

tendrá distribución t-student (por fórmula de la t-student). Pero la ecuación anterior es igual a:

$$\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

Que es a lo que queríamos llegar.

Distribución F-*Fisher.*

El caso de la distribución F también se deduce a partir del modelo de distribución normal estándar, como sigue:

Sea X_1, X_2,…, X_n una muestra aleatoria de n-valores provenientes de una normal estándar y Y_1, Y_2, … ,Y_m m-variables valores también de una normal estándar, entonces el resultado aleatorio F como sigue:

$$F = \frac{(X_1^{\,2} + X_2^{\,2} + ... + X_n^{\,2})/n}{(Y_1^2 + Y_2^2 + ... + Y_m^{\,2})/m}$$

F tendrá una distribución F con n-grados de libertad en el numerador y m-grados de libertad en el denominador.

En este caso, si se desea hacer inferencias respecto a las varianzas de dos poblaciones (por ejemplo quién produce con menos error de manufactura, la empresa A o la empresa B). Para ello se calcula las varianzas muestrales extraídas de dos poblaciones distintas y el estadístico S_1^2 / S_2^2 se comportará como un modelo F, lo anterior se cumple siempre y cuando la varianza de la primera población (σ_1^2) sea igual a la varianza de la segunda población (σ_2^2), ya que el estadístico:

$$\frac{\frac{(n-1)S_1^{\ 2}}{\sigma^{\ 2}_1}}{\frac{(m-1)S_2^{\ 2}}{\sigma_2^{\ 2}}} = \frac{S_1^{\ 2}}{S_2^{\ 2}}$$

se comporta como una distribución F con n-grados de libertad en el numerador y m-grados de libertad en el denominador.

CAPITULO III. ESTIMACIÓN DE PARÁMETROS E INTERVALOS DE CONFIANZA

La acción directiva tiene en las crisis su campo natural de trabajo, pues es atributo del director el enfrentamiento de ellas, sea para prevenirlas, sea para resolverlas o bien sea para amortiguar sus consecuencias, pero sea cual fuere la situación de la alta dirección, el uso de los métodos cuantitativos es innegable; y entre ellos, la Estadística no es precisamente el de menor uso. Y claro está, que el análisis e interpretación de los estados financieros provenientes de la contabilidad es el terreno del cual parten las acciones directivas.

Los modelos estratégicos para el manejo de las empresas suelen estudiarse para su aplicación en situaciones de normalidad, siendo muy pocos los estudios que se refieren precisamente a las coyunturas[1] de turbulencia, complejidad e incluso caos indomeñable[2], siendo aquí, donde la maximización de los recursos de la empresa y la minimización de los costos, requiere con mayor fuerza de los métodos cuantitativos y en particular de la investigación de operaciones junto con la Estadística.

El estudio o tratamiento de empresas en crisis, es ahora inevitable, ya que los momentos de crisis son ahora más comunes y frecuentes que los momentos de normalidad, provocando que en nuestros tiempos la acción directiva puede resumirse sucintamente así: Acción de síntesis sobre las situaciones criticas.

El trabajo de dirección se caracteriza por no contar con reglas fijas conocidas, y cuyos resultados son inciertos (aunque implique la obligación de acertar).

III.1. Estimación de parámetros

En el último tema vimos como se puede emplear la teoría del muestreo para recabar información acerca de muestras aleatorias tomadas -de una población conocida[3]. Desde un punto de vista práctico, no obstante, suele resultar más importante ser capaz de inferir información sobre la población a partir de muestras suyas. Con tal situación trata la **inferencia estadística**, que usa los principios de la teoría del muestreo.
Un problema importante de la inferencia estadística es la **estimación de parámetros de la población**, o brevemente **parámetros** (tales como la media o la varianza de la población), de los correspondientes **estadísticos muestrales**, o simplemente **estadísticos** (tales como la media y la varianza de la muestra).

Consideremos este problema en nuestro presente tema.

---

[1] Combinación de factores y circunstancias que, para la decisión de un asunto importante, se presenta en una nación. (diccionario de la real academia española)
[2] indomable. (DRAE)
[3] Cuando la población de interés no es muy grande y se tiene acceso a ella, se puede calcular fácilmente los parámetros de la misma; sin embargo, en la mayoría de los casos es necesario estimar la media de la población y algunos otros parámetros. (Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 242)

Demos una definición técnica para poder continuar con el análisis. Utilicemos **"a"** como un símbolo genérico de un parámetro poblacional y, **"â"** para indicar una estimación de **"a"** basada en datos de la muestra. Una vez acordado esto podemos decir que:

Estimador.

Un estimador "â" de un parámetro "a" es una función de los valores muestrales aleatorios, que proporciona una estimación puntual de "a". Un estimador es en sí una variable aleatoria y por consiguiente tiene una distribución muestral teórica.

En donde se llama **estimador puntual**[4] al número (punto sobre la recta real), que se calcula a partir de una muestra dada y que sirve como una aproximación (estimación) del valor exacto desconocido del parámetro de la población. Es decir:

Estimador puntual:

Valor que se calcula a partir de la información de la muestra, y que se usa para estimar el parámetro de la población.

[4] Kreyszig. Erwin. "Matemáticas avanzadas para ingeniería". Editorial: Limusa. Vol. 2. p.p 958

Existe una distinción técnica entre un **estimador** como una función de variables aleatorias y una **estimación** como un único número. Tal distinción se refiere al proceso en sí (estimador) y el resultado de dicho proceso (la estimación.) Lo que en realidad importa de esta definición es que: nosotros solo podemos definir buenos procesos (estimadores), mas no garantizar buenos resultados (estimaciones).

Por ejemplo:

la media muestral (**ē**)* es el mejor estimador de una población normal (**μ**), sin embargo no podemos garantizar que el resultado sea óptimo todas las veces. Es decir, no podemos garantizar que, para cada muestra, la media muestral esté siempre más cerca de la media poblacional, que, digamos, la mediana muestral. Así, lo más que podemos hacer es encontrar estimadores que den buenos resultados en el límite.

* en realidad debe ser una "x barra".

Como una aproximación[5] de la media **μ** de una población, puede tomarse la media **ē**[6] de una muestra correspondiente, lo cual da la estimación: **û = ē,** para **μ**, es decir:

$$\hat{u} = \bar{e} = \frac{1}{n}\sum_{i=1}^{i=n} x_i \quad \text{----------(1)}$$

donde **n**= tamaño de la muestra.
Del mismo modo, una estimación para la varianza de una población, es la varianza de una muestra correspondiente; es decir:

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

evidentemente estos casos 1 y 2 son estimaciones de los parámetros para distribuciones en las que **μ** o bien la varianza aparecen explícitamente como parámetros, tales como las distribuciones Normal y de Poisson. Aquí, podemos mencionar que (1) es un caso muy especial del llamado **Método de los momentos**. En este método, los parámetros que van a estimarse se expresan en términos de los momentos de la distribución[7], en las fórmulas resultantes, esos momentos se reemplazan por los momentos correspondientes de la muestra. Esto proporciona las estimaciones deseadas. Aquí, el k-ésimo momento de una muestra $x_1, x_2, \ldots x_n$, es:

$$m_k = \frac{1}{n}\sum_{i=1}^{i=n}(x_i)^k$$

[5] Kreyszig. Erwin. "Matemáticas avanzadas para ingeniería" vol. 2. editorial: limusa. P.p 958
[6] considerése a este símbolo como la media aritmética de la muestra.
[7] Para mayor información consulte la sección 19.8 del libro: "Matemáticas avanzadas para ingeniería" de Erwin Kreyszig. Editorial: Limusa. Vol. 2.

Estimador insesgado

Un estimador **â** que es una función de datos muestrales, se conoce como: **Estimador insesgado** del parámetro poblacional **a** si su valor esperado es igual a **e**. Dicho de otra manera, **â** es un estimador insesgado del parámetro **a** sí:

$$E(\hat{a}) = a$$

La condición de que el estimador **â** es insesgado supone que el valor **promedio** de **â** es exactamente correcto.
Cuando es estimador es sesgado, la magnitud del sesgo viene dada por:

$$\text{Sesgo}(\hat{a}) = E(\hat{a}) - a$$

Si la media de las distribuciones de muestreo de un estadístico es igual que la del correspondiente parámetro de la población, el estadístico se llama un estimador sin sesgo del parámetro; si no, se llama un estimador sesgado. Los correspondientes valores de tales estadísticos se Llaman estimaciones sin sesgo y sesgadas, respectivamente.

Por ejemplo:

La media de las distribuciones de muestreo de medias $\mu_{\bar{x}}$ y μ, la media de la población. Por tanto, la media muestral $\mu_{\bar{x}}$ es una estimación sin sesgo de la media de la población μ.

En términos de Esperanzas, podríamos decir que un estadístico es insesgado si su esperanza es igual al correspondiente parámetro de población.

Estimador Eficiente

Se dice que un estimador es **el más eficiente** para un problema particular cuando tiene el error estándar más pequeño de todos los estimadores insesgados posibles.

Se utiliza la palabra eficiente porque, en una situación dada, el estimador hace el mejor uso posible de los datos muestrales. Y de acuerdo con la teoría estadística clásica, en términos generales se debe preferir el estimador insesgado más eficiente sobre cualquier otro. De aquí, más adelante veremos que las **Hipótesis** nos dicen cual es el estimador más eficiente de un cierto parámetro en un momento dado.

Así por ejemplo

si las distribuciones de muestreo de dos estadísticos tienen la misma media (o esperanza), el de menor varianza se llama un Estimador eficiente de la media, mientras que el otro se llama un estimador ineficiente. Los valores correspondientes de los estadísticos se llaman estimación eficiente y estimación ineficiente, respectivamente.

Si consideramos todos los posibles estadísticos cuyas distribuciones de muestreo tienen la misma media, aquel de varianza mínima se llama a veces el estimador de máxima eficiencia, o sea, el mejor estimador.

Ejemplo

Las distribuciones de muestreo de media y mediana tienen ambas la misma media, a saber, la media de la población. Sin embargo, la varianza de la distribución de muestreo de medias es menor que la varianza de la distribución de muestreo de medianas. Por tanto, la media muestral da una estimación eficiente de la media de la población, mientras la mediana de la muestra da una estimación ineficiente de ella.

En la práctica, las estimaciones ineficientes se usan con frecuencia a causa de la relativa sencillez con que se obtienen algunas de ellas.

De manera desafortunada, las declaraciones de eficiencia dependen fuertemente de algunos supuestos. Por ejemplo, cuando la distribución de la población no es normal, la media muestral no es siempre el estimador más eficiente. Con lo cual surge un tema de investigación en la teoría estadística, es el de los llamados **estimadores robustos**: estadísticos casi insesgados y casi eficientes para una gran variedad de distribuciones poblacionales. Semejantes estimadores todavía son motivo de estudio en la teoría estadística.

Estimador consistente

Un estimador es consistente si se aproxima al parámetro poblacional con probabilidad uno a medida que el tamaño de la muestra tiende a infinito.

Por ejemplo:

la media muestral $\mu_{\bar{x}}$ de una muestra aleatoria tiene valor esperado μ y un error estándar que se aproxima a cero a medida que "n" tiende a infinito. Por lo tanto, cuando el tamaño de la muestra tiende a infinito, la media muestral $\mu_{\bar{x}}$ se aproxima a μ tanto como se quiera. Y de acuerdo con la definición, la media muestral $\mu_{\bar{x}}$ es consistente.

Un estimador inconsistente es a todas luces un mal estimador y no es aconsejable dar una estimación imprecisa basada en una infinidad de datos, cosa que puede suceder si el sesgo de un estimador se aproxima a cero a medida que "n" tiende a infinito. Por ejemplo, utilizar el 25 percentil para estimar la mediana poblacional produciría un estimador inconsistente. También habría inconsistencia si el error estándar de un estimador no tiende a cero a medida que el tamaño muestral crece.

Por lo general, los estimadores inconsistentes son el resultado de alguna equivocación o, lo que es más probable, resultan del fracaso de una hipótesis clave.

Método de máxima verosimilitud[8]

Para responder a la pregunta: ¿Cómo se procede en cualquier situación de muestreo para encontrar un estimador de un parámetro?, la Estadística dice: Con el **método de máxima verosimilitud** de R. A. Fisher[9]. el cual es un procedimiento general para la selección de estimadores.

Hay varias razones por las que se quiere utilizar un estimador de máxima verosimilitud para un parámetro; aunque dichos estimadores no siempre son eficientes e insesgados, por lo general son la mejor opción que se tiene debido a las siguientes propiedades:

- A medida que se incrementa el tamaño muestral, el sesgo del estimador de máxima verosimilitud tiende a cero.
- Su error estándar se aproxima al mínimo error estándar posible. Y
- Su distribución muestral se aproxima a la normal.

Debido a estas propiedades, muchos investigadores están a favor del uso de los estimadores de máxima verosimilitud en gran cantidad de situaciones de muestreo.

Pero veamos con más detalle cómo podemos encontrar un estimador de máxima verosimilitud.

Por lo tanto, empecemos por entender qué es la función de verosimilitud.

FUNCIÓN DE VEROSIMILITUD[10]

Para explicarla[11], sea una **variable aleatoria**[12] discreta (o continua) "**Y**" cuya función de probabilidad (o densidad) $f_Y(y)$ depende de un solo parámetro **a** y tómese una muestra correspondiente de "**n**" valores independientes: $y_1, y_2, \ldots, y_n$. Entonces, en el caso discreto la función de verosimilitud **L** es la

[8] Hildebrand, David, K. & Lyman Ott. R. "Estadística aplicada a la administración y a la economía". Editorial: Addison Wesley Longman. P.p 285

[9] Sir Ronald Alyner Fisher (1890-1962) fue un especialista inglés en genética y estadística, experimento la necesidad de precisar los métodos estadísticos para interpretar datos cualitativos. En sus trabajos sobre pruebas de hipótesis, desarrollo aplicaciones de la distribución F, por lo que lleva su nombre. Esta distribución se utiliza para probar la varianza de pequeñas muestras de una población. (nota tomada del libro: "probabilidad y Estadística" de Stephen S. Willoughby. Editorial: Publicaciones cultural S.A. p.p 122)

[10] Hildebrand, David, K. & Lyman Ott. R. "Estadística aplicada a la administración y a la economía". Editorial: Addison Wesley Longman. P.p 286

[11] Kreyszig. Erwin. "Matemáticas avanzadas para ingeniería". Editorial: Limusa. Vol. 2. p.p 958

[12] ver anexo 1

probabilidad de observar los datos que de hecho se están observando, es decir:

$$L(y_1,y_2,\ldots,y_n,a) = P(y_1,y_2,\ldots y_n)$$

Que consideramos como una función del parámetro desconocido de la población **a.** Y si los datos se toman de una distribución continua, la distribución de probabilidad **P** se reemplaza por la función de densidad **f**, es decir:

$$L\ (y_1,y_2,\ldots y_n,\ a) = f(y_1,y_2,\ldots,y_n)$$

Suponiendo que los valores muestrales se toman independientemente, podemos obtener la probabilidad **P** o la densidad **f** como un producto, tal como se indica a continuación:

La probabilidad en el caso discreto de que una muestra de tamaño "**n**" consista de esos "**n**" valores es:

$$L(y_1,y_2,\ldots,y_n,\ a) = P(y_1)P(y_2)\ldots P(y_n)$$

Y en el caso continuo, la probabilidad de que la muestra consista de valores, en pequeños intervalos pertenecientes a la muestra es:

$$L(y_1,y_2,\ldots,y_n,\ a) = f(y_1)f(y_2)\ldots f(y_n)$$

Ya que $f(y_i)$ depende de **a,** la función "**L**" depende de $y_1,\ y_2,\ldots,y_n$ y **a.** Si consideramos además que: $y_1,\ y_2,\ldots,y_n$ son dados y fijos; entonces "**L**" es una función de **a**, que se llama **función de verosimilitud**.

Es decir, que si, en un experimento binomial con **n=5**, obtenemos **y=2**, entonces la verosimilitud es simplemente la probabilidad de dos éxitos en cinco ensayos tomada como una función de la probabilidad de éxito desconocida de la población, **P**.

Por ejemplo[13]:
suponga que independientemente de lo que sucede el resto de los días, el número de trabajos que llegan en un día a un despacho contable tiene una distribución de Poisson con media desconocida μ**.** Suponga además que el primer día de la muestra llega sólo un trabajo y que el segundo (y último) día llegan cuatro. Escriba la función de verosimilitud.

Para resolver este problema, la metodología es la siguiente:

Primer paso: debemos escribir la fórmula básica de la cual estamos partiendo, identificando exhaustivamente todas sus variables; en este caso, la fórmula corresponde a una distribución de Poisson, por lo tanto, recordando que la distribución de Poisson es discreta con:

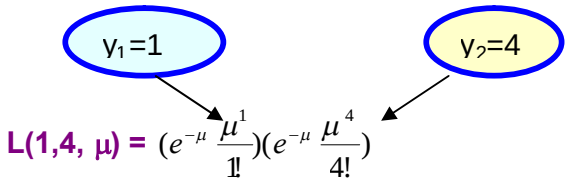
[13] Hildebrand, David, K. & Lyman Ott. R. "Estadística aplicada a la administración y a la economía". Editorial: Addison Wesley Longman. P.p 287

$$P(y) = e^{-\mu}\,\frac{\mu^{y}}{y!}$$

en donde: μ es el número esperado de eventos que suceden en un periodo y

e = 2.71828....

Segundo paso: **sustituir los valores o datos dados por el problema en la fórmula original, teniendo en cuenta la teoría de la función de verosimilitud.** Los valores observados son: **y_1=1** e **y_2=4**. por lo tanto, la función de verosimilitud estará formada por el producto para cada uno de los datos de la fórmula misma. Es decir:



Tercer paso: **realizar las operaciones algebraicas correspondientes a la reducción de la fórmula;** lo cual quiere decir que finalmente la fórmula anterior se puede reducir a:

$$L(1,4,\mu) = e^{-2\mu}\,\frac{\mu^{5}}{(1!)(4!)}$$

Siendo este último resultado la función de verosimilitud solicitada en el problema.
A continuación es necesario entender qué es una **estimación de máxima verosimilitud.**

Estimación máximo verosímil.

Para valores observados en una muestra y_1, y_2,...,y_n, la estimación máximo verosímil de un parámetro e es el valor ê que maximiza la función de verosimilitud L (y_1,,y_2, e).

En el ejemplo anterior podemos encontrar a través de las tablas correspondientes que el valor de μ que maximiza la función de verosimilitud es 2.5, así la estimación máximo verosímil es **μ= 2.5**

En un principio siempre es posible encontrar estimadores de máxima verosimilitud calculando numéricamente la función de verosimilitud. No obstante, el utilizar el cálculo diferencial simplifica el trabajo de encontrar tales estimadores.

La idea básica[14] del método de máxima verosimilitud es muy sencilla y es como sigue:

Se elige aquella aproximación para el valor desconocido de **a** para el cual "L" sea tan grande como sea posible. Si "L" es una función diferenciable de **a,** una condición necesaria para que "**L**" tenga un máximo (no en la frontera) es:

$$\frac{\partial L}{\partial \mathrm{A}} = 0 \text{ ----------6}$$

se escribe una derivada parcial, debido a que "L" también depende de: y_1, y_2,...,y_n y una estimación de (6) que depende de y_1, y_2,...,y_n, se llama estimación de máxima verosimilitud para "a".

Recordemos que para determinar el máximo de una función se iguala a cero la primera derivada y se resuelve la ecuación que de ello resulta.

En los problemas de máxima verosimilitud con frecuencia es más conveniente trabajar con el logaritmo natural de la verosimilitud que con la verosimilitud misma. Por lo tanto, podemos reemplazar (6) por:

$$\frac{\partial \ln(L)}{\partial \mathrm{A}} = 0 \text{ ----------7}$$

debido a que $f \geq 0$, un máximo de "f" en general es positivo y "ln (L) es una función monótona creciente[15] de "L". Esto a menudo simplifica los cálculos.

En principio se debería utilizar el criterio de la segunda derivada para asegurarse que lo que se obtiene es un máximo y no un mínimo. No obstante es muy claro que la solución de la ecuación correspondiente a la primera derivada produce un estimador de máxima verosimilitud y no un mínimo.

Finalmente, si la distribución de "**Y**" contiene "**r**" parámetros: a_1, a_2,...,a_r, entonces en lugar de (6) se tiene las "**r**" condiciones:

$$\frac{\partial L}{\partial \mathrm{A}_1} = 0, \ \frac{\partial L}{\partial \mathrm{A}_2} = 0, \ldots, \frac{\partial L}{\partial \mathrm{A}_r} = 0$$

y en lugar de (7) tenemos:

$$\frac{\partial \ln(L)}{\partial \mathrm{A}_1} = 0, \ \frac{\partial \ln(L)}{\partial \mathrm{A}_2} = 0, \ldots, \ \frac{\partial \ln(L)}{\partial \mathrm{A}_r} = 0$$

[14] Kreyszig, Erwin. "Matemáticas avanzadas para Ingeniería". Editorial: Limusa. Vol. 2. p.p 959

[15] En virtud de que el logaritmo natural es una función creciente, a medida que la verosimilitud se incrementa hacia su máximo, también lo hace su logaritmo.

Por lo tanto, **continuando con el ejemplo anterior** tenemos que:

la función de verosimilitud era:

$$\mathbf{L(1,4,\mu) =}\ e^{-2\mu}\frac{\mu^5}{(1!)(4!)}$$

De modo que continuando con el proceso, **el logaritmo natural de la verosimilitud es:**

$$\mathbf{L(1,4,\mu) =}\ \ln e^{-2\mu} \mathbf{+} \ln\frac{\mu^5}{(1!)(4!)}$$

de donde por leyes de los logaritmos, esta ecuación queda de la siguiente manera:

$$\mathbf{L(1,4,\mu) = -2\mu\ (ln\ e) +} \ln\mu^5 - \ln\left[(1!)(4!)\right]$$

continuando con las leyes de los logaritmos, la expresión toma la forma siguiente:

$$\mathbf{L(1,4,\mu) = -2\mu + 5\ ln\ \mu - ln\ [(1!)(4!)[}$$

y extrayendo **la primera derivada a esta ecuación** tenemos que ésta cobra la siguiente forma:

$$\frac{dL(1,4,\mu)}{d\mu} = \frac{d}{d\mu}(-2\mu) + \frac{d}{d\mu}(5\ln\mu) - \frac{d}{d\mu}\left[\ln(1!)(4!)\right]$$

de donde aplicando leyes de la derivación matemática tenemos que esta expresión se convierte en:

$$\frac{dL(1,4,\mu)}{d\mu} = -2 + \frac{5}{\mu}$$

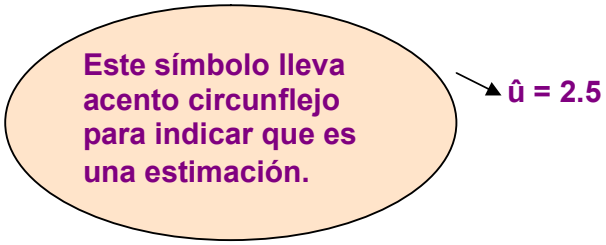
continuando con el proceso, igualamos a "cero" esta primera derivada**, quedando la expresión como se indica a continuación:**

$$\frac{dL(1,4,\mu)}{d\mu} = -2 + \frac{5}{\mu} = 0$$

que es lo mismo que:

$$-2 + \frac{5}{\mu} = 0$$

de donde resolviendo esta ecuación de primer grado **con una incógnita tenemos que:**



de modo que la estimación de máximo verosímil o de máxima verosimilitud de **μ**, es **û = 2.5,** es decir, el promedio de trabajos que llegan al despacho es 2.5 por día, o bien 5 cada dos días.

en resumen, la metodología para encontrar una estimación de máximo verosímil es:

Metodología para encontrar un estimador de máxima verosimilitud:

Primer paso: **identificar la fórmula básica a que se refiere el problema junto con todas sus variables de manera exhaustiva.**

Segundo paso: **encontrar la función de verosimilitud correspondiente (sustituyendo los datos dados en la formula original, teniendo en cuenta la teoría de la función de verosimilitud).**

Tercer paso: **aplicar la función Logaritmo natural a la función de verosimilitud.**

Cuarto paso: realizar las operaciones propias de los logaritmos para desglosar la función en sumas y restas,(dentro de las cuales es común que queden comprendidas: multiplicaciones y divisiones).

Quinto paso: aplicar la primera derivada a la función logaritmo natural.

Sexto paso: realizar operaciones correspondientes a la teoría de derivación.

Séptimo paso: igualar el resultado reducido de la primera derivada a cero.

Octavo paso: resolver la ecuación de primer grado resultante, con lo cual obtenemos el resultado del estimador de máxima verosimilitud.

Ejercicios propuestos:

1. Considere un **experimento binomial** que consiste de saber la aceptación de un producto en el mercado y suponga con fines

ilustrativos que **n = 5** Y **y = 2**. Encuentre el estimador de máximo verosímil correspondiente.

Como **primer paso** para resolver este problema, identificamos que se trata de una distribución de tipo binomial, cuya fórmula es la siguiente:

$$P_Y(y) = C_y^n (p)^y (q)^{n-y}$$

en donde: n = número de experimentos
y = número de evento en el cual queremos tener éxito.
p = probabilidad de éxito
q = probabilidad de fracaso. (q = 1 –p)
C_y^n **= número de combinaciones de "n" elementos tomados de "y" en "y".**

Por lo tanto, la fórmula anterior, también la podemos escribir de la siguiente forma:

$$P_Y(y) = \frac{n!}{y!(n-y)!}(p)^y (1-p)^{n-y}$$

como **segundo paso** tenemos que sustituir los datos dados por el problema en la fórmula para encontrar la función de verosimilitud correspondiente.

Por lo tanto:

$$L(2,p) = \frac{5!}{2!(5-2)!}(p)^2 (1-p)^{5-2}$$

de donde realizando algunas operaciones tenemos que:

$$L(2,p) = \frac{5!}{(2!)(3)!}(p)^2 (1-p)^3$$

siendo esta la verosimilitud del problema (note que debido a la continuación del problema, no es necesario resolver el número de combinaciones presentada)

como **tercer paso**, aplicamos la función logaritmo natural a la función de verosimilitud, quedando esta como se indica a continuación:

$$l(2,p) = \ln(\frac{5!}{(2!)(3)!}) + \ln(p)^2 + \ln(1-p)^3$$

de donde aplicando leyes de los logaritmos **-como cuarto paso-,** se transforma en:

$$l(2,p) = [\ln(5!) - \ln(2!)(3!)[+2\ln(p) + 3\ln(1-p)$$

de donde como **quinto paso** procedemos a calcular la primera derivada de esta expresión matemática:

$$\frac{\partial}{\partial p} l(2, p) = 0 - 0 + 2\frac{1}{p} + 3(\frac{1}{1-p})(-1)$$

es decir, al realizar las operaciones correspondientes como **sexto paso**, queda de la siguiente manera:

$$\frac{\partial}{\partial p} l(2, p) = \frac{2}{p} - \frac{3}{1-p}$$

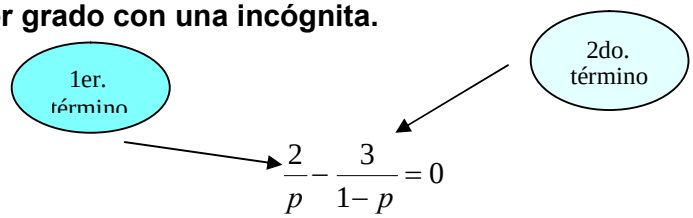
de donde, igualando esta primera derivada a cero –como **séptimo paso**- la expresión se transforma en:

$$0 = \frac{2}{p} - \frac{3}{1-p}$$

o bien:

$$\frac{2}{p} - \frac{3}{1-p} = 0$$

finalmente como **octavo paso**, procedemos a resolver esta **ecuación de primer grado con una incógnita.**



pasando el segundo término al segundo miembro tenemos:

$$\frac{2}{p} = \frac{3}{1-p}$$

pasando el denominador del primer miembro al segundo miembro, y el denominador del segundo miembro al primer miembro tenemos:

$$2(1-p) = 3(p)$$

eliminando paréntesis, la expresión toma la forma:

$$2 - 2p = 3p$$

y pasando el segundo miembro al primer miembro:

$$2 - 2p - 3p = 0$$

efectuando reducción de términos:

$$2 - 5p = 0$$

pasando el segundo término del primer miembro al segundo miembro:

$$2 = 5p$$

y finalmente despejando el valor de "**p**" tenemos que:

$$\frac{2}{5} = p$$

$$p = \frac{2}{5}$$

siendo esta la respuesta al problema propuesto, es decir, que **la probabilidad de éxito en el segundo intento de un experimento binomial con n=5 es de: P=0.40 (40% del mercado aceptan el producto).**

ESTIMACIÓN POR EL MÉTODO DE MOMENTOS.

El caso de la estimación por momentos es otra metodología que estima el parámetro poblacional igualando los momentos muestrales con los momentos poblacionales.

Como se mencionó en la sección III.2. el primer momento poblacional es E(X) (valor esperado de X), el segundo momento poblacional es $E(X^2)$, así sucesivamente. Mientras que el primer momento muestral es $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ (el promedio de la muestra), el segundo momento muestral es $\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$, así sucesivamente.

Considere el caso de una población cuya función densidad de probabilidad es $f_x(x)$ y parámetro desconocido θ, como sigue:

$$fx^{(x)} = \begin{cases} (\theta+1)X^{\theta} & 0 \le \chi \le 1 \\ 0 & 0.C. \end{cases}$$

Entonces si quisiéramos estimar el parámetro θ, entonces debemos calcular el primer momento poblacional e igualarlo con el primer momento muestral, a saber:

Estimar θ por el metodo de momentos.

$$E(x)=\int x f_x(x)dx$$

$$E(x)=\int_0^1(\theta+1)x^{\theta}dx=\int_0^1(\theta+1)x^{\theta+1}dx=\frac{(\theta+1)}{(\theta+2)}x^{\theta+2}\Big|_0^1=\frac{\theta+1}{\theta+2}$$

Igualando el primer momento poblacional con el primer momento muestral, tenemos:

$$\frac{\theta+1}{\theta+2}=\frac{\sum X_1}{n}=\bar{x}$$

Y despejando θ, *tenemos*:

$$\hat{\theta}+1=\bar{x}(\hat{\theta}+2)$$

es decir:

$$\hat{\theta}(1-\bar{x})=2\bar{x}-1$$

$$\hat{\theta}=\frac{2\bar{x}-1}{1-\bar{x}}$$ estimando puntual por momentos.

Así por ejemplo si la variables estudiada X es el porcentaje de agrado de un producto y dicho porcentaje (de 0 a 100) se distribuye de a cuerdo a la función de densidad $f_x(x)$ (que para asumir cierto modelo se puede utilizar una prueba de bondad de ajuste), entonces para estimar θ se determina una muestra aleatoria en la cual consideramos que arroja un promedio $\bar{x}=0.39$ (es decir 39% de satisfacción). Por lo cual en este caso el estimador de θ es: $\hat{\theta}=\frac{2\bar{x}-1}{1-\bar{x}}=\frac{2(0.39)-1}{1-0.39}=-0.36$, valor que no tiene significado práctico pero que a partir del cual se describe le comportamiento de la población y en la cual el promedio es $E(X)=\frac{\theta+1}{\theta+2}=\frac{-0.36+1}{-0.36+2}=0.39$ y así mismo se puede calcular la mediana, moda, varianza, entre otras características.

ESTIMACIÓN DE MEDIAS Y DESVIACIONES ESTÁNDAR[16].

En estadística, numerosos problemas están relacionados con la estimación de la media o la desviación estándar de una población dada, a partir del estudio de una muestra de tamaño "n".

Por ejemplo:

- A una empresa le puede interesar el número promedio de piezas defectuosas producidas por una cierta máquina;
- A un ingeniero especialista en vehículo le puede interesar la **variabilidad** en el funcionamiento de un tipo vehículo.

En las secciones anteriores se vio que si se supone que cada muestra de tamaño "n" tiene la misma probabilidad de ser seleccionada, entonces la media de la distribución de las medias de la muestra es la misma que la de la

[16] Stephen S. Willoughby. "Probabilidad y Estadística". Editorial: Publicaciones cultural, s.a. p.p 138-140

población original, $\mu_{\bar{x}} = \mu$. Aún más, para poblaciones suficientemente grandes, o para muestreos con reemplazo, la desviación estándar de la distribución de las medias de la muestra, $\sigma_{\bar{x}}$, está relacionada con la desviación estándar de la población σ, por la ecuación:

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

si en una aplicación particular fuera práctico seleccionar todas las posibles muestras de tamaño "n", para determinar la media de cada una de ellas y, después, calcular la media y la desviación estándar de la distribución de las medias de las muestras, las fórmulas anteriores permitirían calcular μ y σ directamente. Por lo general, este procedimiento no es práctico. Lo que comúnmente se hace es no estudiar todas las muestras de tamaño "n" sino únicamente "una" de ellas. La media $\bar{x}$ y la desviación estándar "s", de esa muestra únicamente se toman como estimaciones de μ y σ, la media y la desviación estándar que corresponden a la población original. Puesto que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, las estimaciones para $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$, son $\bar{x}$ y $\frac{s}{\sqrt{n}}$ respectivamente. Enseguida se ilustra el procedimiento de estimación con un ejemplo:

Ejemplo[17]:

Se escoge una muestra aleatoria de 36 recién egresados en la carrera de contaduría de cierta universidad y al aplicarles un examen de aptitudes, se obtuvieron las siguientes puntuaciones:

63	64	64	65	65	66
66	66	67	67	67	67
67	68	68	68	69	69
69	69	69	70	70	70
71	72	72	72	72	73
73	74	74	76	76	77

La media de la muestra $\bar{x}$ es de 69, (al punto más próximo), y la desviación estándar "s", es de 3.5. utilizando $\bar{x}$ y "s" como estimaciones de μ y σ, podemos afirmar que la puntuación media de todos los recién egresados de dicha universidad es de **alrededor de** 69 puntos. Aún más, podemos decir que la desviación estándar de las puntuaciones de los recién egresados respecto a la media es, **aproximadamente**, 3.5 puntos.

[17] Stephen S. Willoughby. "Probabilidad y Estadística". Editorial: Publicaciones cultural, s.a. p.p 139-140

el procedimiento anterior es satisfactorio tal como se ha presentado. El problema estriba en el contenido de las palabras **alrededor de** y **aproximadamente**. Cuando decimos que la altura promedio de los niños es de alrededor de 69 cm, ¿queremos significar que esta altura tiene cuando mucho 1 o 10 cm. De diferencia con respecto al verdadero promedio?. Por supuesto, la exactitud de nuestra estimación depende de la muestra escogida. Afortunadamente, en el caso de muestras aleatorias, es posible dar apoyo probabilístico al significado de las palabras **alrededor de** y **aproximadamente**.

Un hecho importante que se debe tener en cuenta en la distribución de las medias de las muestras, cuando ésta es grande y se selecciona aleatoriamente, es que se puede aproximar a una distribución normal que tenga la misma media $\mu_{\bar{x}}$ y la misma desviación estándar $\sigma_{\bar{x}}$. La demostración del hecho anterior –algunas veces se llama teorema central del límite- va más allá del alcance de estas notas.
Puesto que la distribución de las medias de las muestras es aproximadamente normal, se puede utilizar ventajosamente el conocimiento sobre este tipo de distribución,

III.3. ESTIMACIONES POR INTERVALO y FIABILIDAD

Una estimación de un parámetro de la población dada por un solo número se llama una **estimación de punto** del parámetro. No obstante[18], un estimador puntual sólo refiere una parte de la historia. Si bien se espera que el estimador puntual esté próximo al parámetro de la población, se desearía expresar qué tan cerca está. Un intervalo de confianza sirve a este propósito.

Intervalo de confianza:

Un rango de valores que se construye a partir de datos de la muestra de modo que el parámetro ocurre dentro de dicho rango con una probabilidad específica. La probabilidad específica se conoce como: nivel de confianza.

Es decir, Una estimación de un parámetro de la población dada por dos números, entre los cuales se puede considerar encajado al parámetro, se llama una estimación de intervalo del parámetro.

Las estimaciones de intervalo indican la precisión de una estimación y son por tanto preferibles a las estimaciones de punto.

[18] Douglas A. Lind., et al. "Estadística para administración y economía" editorial: Irwin-McGraw-Hill. P.p 242

Por ejemplo:

Si decimos que el porcentaje de productos defectuosos que produce una máquina es del 6%, entonces el nivel se ha medido 0.06 y estamos dando una **estimación de punto**. Por otra parte, si decimos que el porcentaje es 0.05 ± 0.03 m (o sea, que esta entre 2% y 8%), estamos dando una **estimación de intervalo**.

El margen de error (o la precisión) de una estimación nos informa de su fiabilidad.

III.4. INTERVALO PARA ESTIMAR LA MEDIA

De acuerdo a tablas de la distribución normal estándar el área bajo la curva, entre **x´= -1** y **x´= +1**, es 0.6826. por consiguiente, por la definición de la función normal estándar de probabilidad, las desigualdades siguientes se cumplen con probabilidad de 0.6826

$$-1 < x' < 1$$

como la distribución de las medias de las muestras (con media $\mu_{\bar{x}}$ y desviación estándar $\sigma_{\bar{x}}$) es normal, entonces:

si reemplazamos **x´** por $\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$ en las desigualdades anteriores,

se deberá cumplir:

$$-1 < \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} < +1$$

con probabilidad 0.6826. esto es equivalente a que las desigualdades:

$$\bar{X} - \sigma_{\bar{x}} < \mu_{\bar{x}} < \bar{X} + \sigma_{\bar{x}}$$

se cumplan también con probabilidad 0.6826; sustituyendo ahora:

$$\sigma_{\bar{x}} \text{ por } \frac{s}{\sqrt{n}}$$

se tiene que:

$$\bar{X} - \frac{s}{\sqrt{n}} < \mu_x < \bar{X} + \frac{s}{\sqrt{n}}$$

se cumple con la misma probabilidad.

Podemos esperar entonces, con una probabilidad de 0.68 que μ se encuentre dentro del intervalo:

(69 – 0.58, 69 + 0.58)

se dice que éste es un **intervalo de confianza**, de 0.68 o 68%, ya que se tiene una confianza de 68% de que el intervalo contenga la media de la población.

Si una confianza de 68% fuese insuficiente **se puede usar la tabla de la página 116 (del Willougby)** para hallar otros intervalos.

Por ejemplo:

si se deseara encontrar un intervalo e confianza de 0.95 para μ se requeriría determinar "**k**" de tal manera que las desigualdades siguientes se cumplieran con probabilidad de 0.95

$$-k < \mu < +k \quad \text{----------} \quad 1$$

en términos generales, para encontrar un intervalo de cualquier porcentaje de confianza, se hace lo siguiente:

1º. Se divide el porcentaje de confianza requerido entre 100
2º. el resultado del punto anterior se divide entre 2
3º. El valor así obtenido se busca en las tablas de la curva de distribución normal
4º. El valor encontrado junto al anterior en las tablas se sustituye en 1 y comenzamos el proceso nuevamente.

Es decir, en nuestro caso el valor resultante es de 0.475, por lo tanto, el valor en las tablas que se encuentra junto a éste último es "2", es decir, el área bajo la curva normal estándar entre –2 y +2 es 0.9544, o sea, aproximadamente 0.95. así, la probabilidad de que **X´** se encuentre dentro del intervalo:

(-2, +2)

es, aproximadamente 0.95 o, en otra forma, las desigualdades:

$$-2 < X' < +2$$

se cumplen con probabilidad 0.95;

y puesto que se sabe que la distribución de las medias de las muestras es normal,

se puede reemplazar **X´** por $\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$

expresión que aproximada a:

$$\frac{\bar{X}-\mu_x}{\frac{s}{\sqrt{n}}}$$

en las desigualdades anteriores, así se llega a obtener:

$$-2<\frac{\bar{X}-\mu_x}{\frac{s}{\sqrt{n}}}<+2$$

de donde, resolviendo estas desigualdades para μ , se tiene que:

$$\bar{X}-\frac{2s}{\sqrt{n}}<\mu_x<\bar{X}+\frac{2s}{\sqrt{n}} \text{ ----------2}$$

como un intervalo con 0.95 de confianza para μ . por lo tanto, se puede afirmar con 95% de confianza que μ se encuentra dentro del intervalo:

$$\bar{X}-\frac{2s}{\sqrt{n}} \quad \text{y} \quad \bar{X}+\frac{2s}{\sqrt{n}}$$

por lo tanto, sustituyendo los valores de la media y de la desviación estándar, así como del tamaño de la muestra para el ejercicio anterior (media 69, desviación estándar 3.5 y tamaña de muestra 36) en 2 se tiene que el intervalo con 95% de confianza es:

$$69 - \frac{2(3.5)}{\sqrt{36}} < \mu_x < 69 + \frac{2(3.5)}{\sqrt{36}}$$

$$67.8 < \mu_x < 70.2$$

(67.8 , 70.2)

III.5. INTERVALO PARA ESTIMAR LA VARIANZA

De la sección III.2 sabemos que el estimador para varianza poblacional (σ^2) es S^2, sin embargo para estimar un intervalo de confianza para σ^2 es necesario conocer la distribución del estadístico y más aún, la metodología implica que es necesario tener un estadístico que involucre el parámetro desconocido y que además tenga distribución perfectamente conocida. Por lo cual en este caso el estadístico es:

$$\frac{(n-1)S^2}{\sigma^2}$$

Que de acuerdo a lo estudiado en el capítulo II, tiene una distribución Ji-cuadrada con n-1 grados de libertad. Así que, para una muestra particular, dicho estadístico tiene una probabilidad de estar en un rango dado.

Ejemplo:

Considere el caso de estimar si no hay deficiencias en cuanto a una máquina que llena envases a 500 ml., para ello se extrae una muestra periódicamente y si la muestra indica que hay una variación de más menos 5 ml alrededor de los 500 con un nivel de confianza del 95%, entonces se puede decir que el proceso está bajo control.
En este caso lo que importa es la variación en el llenado pues el nivel promedio de llenado se puede controlar programando la máquina. Por lo cual si la muestra arroja una variación arriba de 5 unidades, entonces el proceso no estará bajo control.
Suponga que la muestra de tamaño 41 arroja una varianza de 13 unidades (desviación estándar de 3.60 ml) . Entonces de acuerdo a la estimación por intervalos de confianza se tendrá que:

$$X^2_{0.025} < \frac{(n-1)S^2}{\sigma^2} < X^2_{0.0975}$$

Que de acuerdo a tablas de Ji-cuadrada con 40 grados de libertad $X^2_{0.025}$=24.433 y $X^2_{0.09750}$ = 59.342.

Entonces el intervalo es:

$$24.433 < \frac{(n-1)S^2}{\sigma^2} < 59.342$$

Y sustituyendo los resultados de la muestra se tiene:

$$24.433 < \frac{(40-1)(13)}{\sigma^2} < 59.342$$

Que al obtener inversos multiplicativos tenemos:

$$\frac{1}{24.433} > \frac{\sigma^2}{(40-1)(13)} > \frac{1}{59.342}$$

Y despejando todas las constantes y dejar solo σ^2 se tiene el intervalo:

$$\frac{1}{24.433} > \frac{\sigma^2}{(40-1)(13)} > \frac{1}{59.342}$$

$$20.75 > \sigma^2 > 8.54$$

Y obteniendo raíz cuadrada, se tiene:

$$4.555 > \sigma > 2.92$$

Por lo cual se puede decir que el proceso está bajo control.

III.6. INTERVALO PARA ESTIMAR LA PROPORCIÓN

En el caso de la proporción, el estadístico a utilizar es:

$$\frac{\bar{p}-\mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p}-P}{\sqrt{P(1-P)/n}}$$

Que de acuerdo al teorema del límite central, tendrá distribución normal estándar. En este caso P es la proporción de la población con una característica dada y que se puede estimar por medio de $\bar{p}$, que es la proporción de la muestra con la característica.

Ejemplo:

Considere el caso de la Bolsa Mexicana de Valores y se desea estimar la proporción de las 250 acciones que tendrán una baja en precio al cierre del día. Para ello se observa una muestra de las primeras 4 horas sobre 50 acciones operadas y se observo que el la proporción que bajo de precio son el 0.10 (10%). En el día se estima que no se presente turbulencias por información importante o privilegiada. Se pide determinar el intervalo de confianza para la proporción total de acciones a la baja con un nivel de confianza del 90%.

De acuerdo a la metodología indicada el intervalo estará determinado por:

$$Z_{\alpha/2} < \frac{\bar{p} - P}{\sqrt{\bar{p}(1-\bar{p})/n}} < Z_{1-\alpha/2}$$

Pero de acuerdo a tablas de normal estándar $Z_{\alpha/2} = Z_{0.05} = -1.64$ y $Z_{0.95} = 1.64$ y como $\bar{p} = 0.10$ entonces el intervalo se deduce de:

$$-1.64 < \frac{0.10 - P}{\sqrt{0.10(1-0.10)/50}} < 1.64$$

que equivale a:

$$-1.64(0.0424264) < 0.10 - P < 1.64(0.0424264)$$

y despejando P se tiene:

$$-1.64(0.04242064) - 0.10 < -P < 1.64(0.0424264) - 0.10$$

igual a:

$$1.64(0.0424264) + 0.10 > P > -1.64(0.0424264) + 0.10$$

Por lo cual el int *ervalo es*:

$$0.169 > P > 0.0304$$

Es decir aproximadamente entre el 3% y 17%.

III.6. TAMAÑO DE MUESTRA

Tamaño de muestra para la media

Hemos visto que para estimar por intervalos la media, el ancho del intervalo está dado por:

$$Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Que representa el número de desviaciones estándar alrededor de la media μ dado el nivel de confianza 1-α. Por lo cual si quisiéramos estimar μ con un nivel de confianza dado y obtener un error en la estimación de a lo más B, tenemos que despejar n de la ecuación:

$$B = Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Despejando n, tenemos:

$$B\sqrt{n} = Z_{\alpha/2}S$$

o bien:

$$n = \left(\frac{Z_{\alpha/2}S}{B}\right)^2$$

Observe que la fórmula involucra el valor S de una muestra, por lo cual el muestreo se puede hacer en dos etapas, en una primera prueba piloto se muestrea con un número reducido de elementos y con ello se calcula el tamaño de n, posteriormente se muestrea en una segunda etapa y se completa la muestra dada por el valor de n.

Como ejemplo supongamos que una empresa comercializa soya texturizada (tipo carne) y deseamos estimar el consumo promedio semestral de una población de consumidores potenciales. Suponga que una muestra piloto de 15 personas arroja que S=12.2 kg., así que si deseamos un nivel de confianza del 95% y un error en la estimación de B=2 Kg., entonces el tamaño de muestra en este caso se obtiene como:

$$n = \left(\frac{Z_{\alpha/2}S}{B}\right)^2 = \left(\frac{1.96(12.2)}{2}\right)^2 = 142.9459$$

Es decir se deben muestrear aproximadamente 143 (128 adicionales a los 15 ya muestreados).

Tamaño de muestra para la proporción.

En este caso el error en la estimación está dado por:

$$B = Z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

Que representa el número de desviaciones estándar alrededor de la media P dado el nivel de confianza 1-α. Por lo cual si quisiéramos estimar P con un nivel de confianza dado y obtener un error en la estimación de a lo más B, tenemos que despejar n de la ecuación:

$$B = Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Despejando n, tenemos:

$$B^2 = Z^2_{\alpha/2}\frac{\bar{p}(1-\bar{p})}{n} \quad \text{o bien:}$$

$$n = \left(\frac{Z^2_{\alpha/2}\bar{p}(1-\bar{p})}{B^2}\right)$$

Suponga que se desea estimar la proporción de acciones que tendrán una baja en el día, para lo cual se observa una muestra de 20 acciones, en las cuales el promedio de las que bajaron son $\bar{p} = 0.17$, entonces si se desea tener un nivel del 95% de confianza de cometer un error de cuando mucho B = 0.09 (9%) en la estimación, determinar el tamaño de muestra.

$$n = \left(\frac{Z^2_{\alpha/2}\, \bar{p}(1-\bar{p})}{B^2} \right)^2 = \left(\frac{(1.96)^2 (0.17)(0.83)}{0.09^2} \right)^2 = 66.91$$

Es decir se deben muestrear aproximadamente 67 (47 adicionales a los 20 ya muestreados).

CAPITULO IV PRUEBAS DE HIPÓTESIS

IV.1. INTRODUCCIÓN[1]:

Cuando las personas toman decisiones, inevitablemente lo hacen con base en las creencias que tienen en relación al mundo que les rodea; llevan en la mente una cierta imagen de la realidad, piensan que algunas cosas son verdaderas y otras falsas y actúan en consecuencia. Por lo tanto:

- Una dependencia gubernamental puede prohibir los anuncios de cigarrillos porque sus directores piensen que el tabaco causa enfermedades del corazón y los pulmones;

- otra entidad rechazaría dar licencia para la producción de una nueva droga contra el cáncer porque no se ha presentado ningún caso creíble de su supuesta efectividad;

- una tercera puede requerir que los motociclistas usen cascos porque se piensa que esta precaución reduce los porcentajes de accidentes mortales, y,

- una cuarta hará campañas para detener la destrucción de sembradíos por plagas de polillas, no por medio de dispersión de insecticidas tradicionales sino por la introducción de parásitos intestinales de esas polillas, procedimiento considerado mucho más eficaz en el logro de la meta deseada.

Del mismo modo, los ejecutivos de empresas toman todos los días decisiones de importancia crucial porque tienen ciertas creencias:

- de que un tipo de máquina llenadora pone al menos un kilogramo de detergente en una caja,

- de que cierto cable de acero tiene una resistencia de 200 kg. O más a la rotura,

- de que la duración promedio de una batería es igual a 100 horas,

- de que un proceso de cápsulas que contienen precisamente 100 miligramos de un medicamento,
- que la empresa de transportes *A* tiene tiempos de entrega más rápidos que la *B*,

- de que la producción de la planta oriente contiene menos unidades defectuosas que la de occidente, …y la lista continúa.

Incluso los estadísticos basas su trabajo en creencias tentativas:

---

[1] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 371-384

- que estas dos poblaciones tienen varianzas iguales,

- que esta población está normalmente distribuida,

- que estos datos muestrales se derivan de una población uniformemente distribuida, etc.

en todos estos casos y en un millón más, las personas actúan con base en alguna creencia sobre la realidad, creencia que quizá llegó al mundo como una simple conjetura, como un poco más que una suposición informada; una proposición adelantada tentativamente como una verdad posible es llamada: **hipótesis**.

Sin embargo, tarde o temprano toda hipótesis se enfrenta a la evidencia que la comprueba o la rechaza y, en esta forma, la imagen de la realidad cambia de mucha a poca incertidumbre. A continuación estudiaremos la forma en que las creencias de las personas pueden ser probadas de manera sistemática.

Definición[2]:

Un método sistemático de evaluar creencias tentativas sobre la realidad se llama: prueba de hipótesis; requiere de la confrontación de creencias con evidencia y decidir, en vista de esta evidencia, si dichas creencias se pueden conservar como razonables o deben desecharse por insostenibles.

IV.2. METOLOGÍA PARA UNA PRUEBA DE HIPÓTESIS:

1. formular dos hipótesis opuestas.
2. seleccionar un estadístico de prueba.
3. derivar una regla de decisión.
4. tomar una muestra, calcular el estadístico de prueba y confrontarlo con la regla de decisión.

Paso 1[3]: formulación de dos hipótesis opuestas:

El primer paso para probar una hipótesis es siempre formular dos que sean mutuamente exclusivas, y también colectivamente exhaustivas, de las facetas posibles de la realidad. Cada una de estas hipótesis complementarias es una proposición sobre un parámetro de la población tal que la verdad de una implique la falsedad de la otra. La primera hipótesis del conjunto, simbolizada por: H_0, se denomina **hipótesis nula**; la segunda hipótesis, simbolizada por: H_1 o bien por **Ha**, es la **hipótesis alternativa**.

---

[2] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 372

[3] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 372

Definición:

La **hipótesis nula**, H_0, es la primera de dos opuestas en una prueba de hipótesis. Es una descripción del estado de cosas en un momento dado (status quo) de sabiduría convencional, de lo que las personas han pensado durante mucho tiempo que es cierto. Si Ho se corrobora en una prueba de hipótesis, no es necesario tomar ninguna acción.

La **hipótesis alternativa**, **Ha**, es la segunda de dos opuestas en una prueba de hipótesis. Es un medio para hacer aseveraciones sorprendentes que contradicen la sabiduría convencional. Si H_0, no se puede corroborar en una prueba de hipótesis, **Ha**, se acepta tentativamente y esto requiere iniciar una acción. Por lo tanto, se puede considerar a Ha como la hipótesis de acción.

Por ejemplo:

Establecer[4] las dos hipótesis para cada una de las situaciones siguientes:

1. un fabricante de aviones necesita láminas de aluminio de 0.3 pulgadas de espesor en promedio, ni más ni menos.

   Solución:

   $H_0 : \mu_0 = 0.03$ $\qquad$ $H_1 : \mu_0 \neq 0.03$

2. un fabricante de aviones necesita varillas de acero especial con una resistencia promedio a la tracción de al menos 5000 libras.

   Solución:

   $H_0 : \mu_0 \geq 5000$ $\qquad$ $H_1 : \mu_0 < 5000$

3. un fabricante de computadoras desea probar lo dicho por un supervisor acerca de que el ensamble de una computadora promedia al menos 40 minutos.

   Solución:

   $H_0 : \mu_0 \leq 40$ $\qquad$ $H_1 : \mu_0 > 40$

[4] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 371-374

Tipos de pruebas de hipótesis[5]:

Las pruebas de hipótesis se clasifican como direccionales o no direccionales dependiendo de cuando la hipótesis nula Ho involucra el signo de igualdad (=).

Si la afirmación de H_0 contiene el signo de igualdad entonces la prueba se llama Prueba no direccional, mientras que **si tal afirmación no contiene = (esto es, si involucra < o >), entonces la prueba se llama prueba direccional**.

Las pruebas no direccionales se llaman también **pruebas de dos colas** y **las direccionales se nombran pruebas de una cola**.

Para pruebas referentes a una muestra de datos, **si la afirmación de "H_0" contiene el símbolo ">", entonces la prueba se llama prueba de cola izquierda**, y

si la afirmación de H_0 tiene el símbolo "<", entonces la prueba se denomina: prueba de cola derecha.

paso 2: selección de un estadístico de prueba.

El segundo paso para probar hipótesis es la selección de un estadístico de prueba:

Un estadístico de prueba es aquel calculado de una sola muestra aleatoria simple tomada de la población de interés, en una prueba de hipótesis para establecer la verdad o falsedad de la hipótesis nula.

Paso 3[6]: derivación de una regla de decisión:

Una vez formuladas dos hipótesis opuestas y seleccionado el tipo de estadístico con qué probarlas, el paso siguiente en la prueba de hipótesis es la derivación de una regla de decisión:

Una regla de decisión es una regla para prueba de hipótesis que especifica por adelantado –para todos los valores posibles de un estadístico de prueba que pueda calcularse de una muestra- **si la hipótesis nula debe ser aceptada o si debe ser rechazada a favor de la alternativa**.

Los valores numéricos del estadístico de prueba para lo que **H_0** es aceptada se dice que están en la **región de aceptación** y son considerados **no significativos estadísticamente**.

Los valores numéricos del estadístico de prueba para lo que **H_0** es rechazada se dice que están en la **región de rechazo** y son considerados **significativos estadísticamente**, porque aconsejan que la hipótesis alternativa sustituya a la entonces desacreditada hipótesis nula.

Es importante notar que: la aceptación o rechazo se refiere a la Hipótesis nula **H_0**.

---

[5] Weimer, Richard C. "Estadística". Editorial: CECSA. P.p 462

[6] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 378

Al principio, la selección de una regla de decisión puede parecer superflua. ¿No es obvio que el valor calculado del estadístico de prueba concuerda con la hipótesis nula o bien la contradice? Pero, si se piensa mejor, queda claro que el asunto es más complicado de lo que pudiera parecer a primera vista. Si bien es cierto que cualquiera puede decir de inmediato si el valor calculado del estadístico de prueba concuerda o no con la hipótesis nula, no es obvio que la divergencia dada entre el valor observado y el hipotético demuestre en forma automática que la hipótesis nula es falsa. Lo demostrado por dicha divergencia es cuestionable porque cualquier estadístico muestral es una **variable aleatoria**; su valor depende en mucho de la muestra particular que sea seleccionada de la población en cuestión.

Paso 4[7]: toma de una muestra, cálculo del estadístico de prueba y confrontación con la regla de decisión.

El paso final en la prueba de hipótesis requiere:

a) seleccionar una muestra aleatoria simple de tamaño **n,** de la población de interés,

b) calcular el valor real (opuesto al crítico) del estadístico de prueba (seleccionado en el paso 2), y

c) su confrontación con la regla de decisión (derivada en el paso 3).

Ejemplo # 1 consideremos el siguiente problema:

Con el fin de determinar la efectividad de una nueva vacuna para prevenir el resfriado común[8], diez personas a las cuales se les inyecto la vacuna se mantuvieron en observación durante un año. De las diez personas, ocho pasaron el invierno sin enfermarse de resfriado. Suponga que se sabe que cuando no se usa la vacuna, la probabilidad de pasar el invierno sin resfriarse es de 0.5 y que el hecho de que una persona pase el invierno sin resfriarse es independiente del estado de salud de cualquier otra persona. ¿Cuál es la probabilidad de observar 8 o más personas que no se resfriaron durante el invierno dado que la vacuna no tiene efecto alguno?

SOLUCIÓN: Suponiendo que la vacuna no es efectiva, la probabilidad de pasar el invierno sin resfriarse es de **p = 0.5**. si "y" representa el número de personas que pasan el invierno sin resfriarse, la distribución de probabilidad de "**y**" esta dada por:

$$P_{(y)} = C_y^{10}(0.5)^y(0.5)^{10-y}$$

[7] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 384

[8] [8] Estadística para administración y economía. Mendenhall/Reinmuth. Grupo editorial Iberoameérica. P.p 131-132

de donde aplicando las leyes de los exponentes tenemos que:

$$P_{(y)} = C_y^{10}(0.5)^{10}$$

por lo tanto:

$$P_{(8omás)} = P_{(8)} + P_{(9)} + P_{(10)}$$

$$P_{(8omás)} = C_8^{10}(0.5)^{10} + C_9^{10}(0.5)^{10} + C_{10}^{10}(0.5)^{10}$$

$$P_{(8omás)} = 0.0439 + 0.0098 + 0.0010$$

$$P_{(8omás)} = 0.055$$

prueba de una hipótesis:

El problema anterior de la vacuna contra el resfriado, ilustra la **prueba estadística de una hipótesis**. El problema práctico esta relacionado con la determinación de la efectividad de la vacuna, es decir: ¿presentan los datos de la muestra una evidencia suficiente que indique que la vacuna es efectiva?

El razonamiento que se emplea en la prueba de una hipótesis es muy semejante al que se emplea en un proceso de tipo judicial.

Al juzgar a un individuo por algún delito, la corte supone que el acusado es inocente mientras no se pruebe su culpabilidad. Y el fiscal debe obtener y presentar todas las evidencias disponibles en un intento por contradecir la hipótesis de “no-culpabilidad.”

En el problema estadístico, la vacuna se presenta como el acusado. La hipótesis a probar, llamada **hipótesis nula,** corresponde a la no-efectividad de la vacuna. La evidencia en este caso está contenida en la muestra extraída de la población de posibles usuarios de la vacuna. El experimentador, representando el papel del fiscal, cree que la **hipótesis alternativa** es verdadera, es decir, que la vacuna es efectiva. Por lo tanto, **el experimentador usará la evidencia contenida en la muestra en un intento por rechazar la hipótesis nula (vacuna no efectiva) apoyando así la hipótesis alternativa de que la vacuna es realmente efectiva.**

Este procedimiento es un ingrediente fundamental del método científico donde todas las teorías propuestas deben compararse con la realidad.

Intuitivamente, se seleccionará el número "**y**" de personas que no contrajeron la enfermedad como una medida de la cantidad de evidencia que presenta la muestra. Si "**y**" fuera muy grande, se rechazaría la hipótesis nula y se concluiría que la vacuna es efectiva. Por otra parte, un valor pequeño de "**y**" constituiría muy poca evidencia para apoyar el rechazo de la hipótesis nula. De hecho, si la hipótesis nula fuera verdadera, la probabilidad de pasar el invierno sin enfermarse de resfriado sería:

$$P = 0.5$$

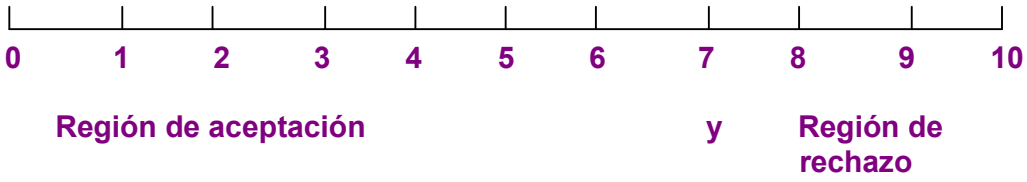
Y el valor promedio de "**Y**" sería:

$$E(y) = np$$
$$= 10\,(0.5)$$
$$= 5$$

Luego entonces, la mayoría de las personas tendrías poca dificultad para tomar una decisión en el caso en que **y = 10**. que parece evidencia suficiente para rechazar la hipótesis nula, así como en los casos en que **y = 5, 4, 3, 2, o 1** en los que lo razonable parece ser aceptarla.

Pero, ¿qué podría decirse en los casos en que **y = 7, 8 o 9**? Es claro que, independientemente del procedimiento que se siga para la toma de decisiones, Se escogería aquel para el que la probabilidad de tomar una decisión incorrecta sea mínima.

La prueba estadística[9] se lleva a cabo de manera objetiva y semejante al procedimiento intuitivo descrito anteriormente. La información contenida en la muestra se obtiene a través de la llamada **estadística de prueba**. En nuestro ejemplo, el número "**y**" de personas que no enfermaron de resfriado constituye la estadística de prueba. Una vez determinado el conjunto de valores que puede tomar la estadística de prueba, éste se divide en dos regiones como se muestra en la siguiente figura:



En la figura anterior, una de las regiones se denomina **región de aceptación** y la otra **región de rechazo.** Si como resultado de la realización de un experimento se obtiene un valor de "**Y**" contenido en la región de rechazo, se rechaza la hipótesis. En caso contrario se acepta. (Precaución: La hipótesis nula será rechazada o aceptada solamente si los riesgos de una decisión equivocada son pequeños para estas dos acciones.)

[9] [9] Estadística para administración y economía. Mendenhall/Reinmuth. Grupo editorial Iberoameérica. P.p 148

Por ejemplo, en el experimento de la vacuna se podría escoger **y = 8, 9 y 10** como los elementos de la región de rechazo y los otros posibles valores de "**y**" como los elementos de la región de aceptación. Como el valor observado de "**y**" fue de 8, se rechaza la hipótesis nula de que la vacuna no es efectiva y se concluye que la probabilidad de pasar el invierno sin resfriarse es mayor que:
P = 0.5 si se usa la vacuna.

¿Cuál es la probabilidad de rechazar la hipótesis nula cuando ésta es en realidad verdadera? Esta es la probabilidad de que "**y**" sea igual a 8, 9 o 10 dado que
p = 0.5. su valor fue calculado en el ejemplo y corresponde a 0.055. como el valor de esta probabilidad es pequeño podemos estar razonablemente tranquilos respecto a la decisión tomada.

Es fácil notar que el fabricante de la vacuna contra el resfriado se enfrenta a dos tipos posibles de error. Por una parte, se podría rechazar la hipótesis nula y concluir erróneamente que la vacuna es efectiva, lo que podría producir pérdidas financieras al implementar algún programa de producción. Por otra parte, se podría no rechazar la hipótesis nula y concluir erróneamente que la vacuna no es efectiva. Esto último podría redundar en una pérdida de ganancias potenciales que podrían derivarse de la venta de una vacuna efectiva.

Definición[10]: error tipo I

Rechazar la hipótesis nula cuando ésta es verdadera se denomina error tipo I para una prueba estadística.

A la probabilidad de cometer un error tipo I se le asigna el símbolo α (letra griega alfa)

La probabilidad de α aumenta o disminuye a medida que aumenta o disminuye el tamaño de la región de rechazo. Entonces, ¿por qué no se disminuye el tamaño de la región de rechazo para hacer α tan pequeña como sea posible? Desgraciadamente, al disminuir el valor de α aumenta la probabilidad de no rechazar la hipótesis nula cuando ésta es falsa y alguna hipótesis alternativa es verdadera. Aumenta entonces la probabilidad de cometer el llamado error de tipo II para una prueba estadística.

Nivel de significancia[11]:

Cualquier resultado muestral que lleve al rechazo de H_0 se denomina: **resultado estadísticamente significativo.**

---

[10] Estadística para administración y economía. Mendenhall/Reinmuth. Grupo editorial Iberoameérica. P.p 149
[11] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 371-380

Problema ejemplo[12]: Incurrir en un riesgo **α**

El estadístico de una compañía está a punto de probar si las varillas de acero especial tienen un promedio de resistencia a la tensión de al menos 5000 libras, como la tenían. ¿Cuáles son las implicaciones si el nivel de significancia de la prueba de hipótesis se fija en: α = 0.08?

Solución:
Dadas las hipótesis:

$H_0 : \mu_0 \geq 5000$ y $H_1 : \mu_0 < 5000$

el procedimiento asegura lo siguiente:

aún cuando las varillas tengan de hecho un promedio de resistencia a la tensión de 5000 libras o más, en el 8% de todas las pruebas la conclusión será lo contrario.

Esto equivale a decirles a las personas que no tienen SIDA, que lo tienen; dichos "positivos falsos" constituyen el **error tipo I** de rechazo. Sin embargo, en el 92% de dichas pruebas este tipo de error se evita, lo que indica su **nivel de confianza**.

Definición[13]: error tipo II

Aceptar la hipótesis nula cuando ésta es falsa se denomina error tipo II para una prueba estadística.
A la probabilidad de cometer un error de tipo II se le asigna el símbolo β (letra griega beta)

Para un tamaño de muestra fijo, α y β están inversamente relacionados; al aumentar uno el otro disminuye. El aumento del tamaño de muestra produce mayor información sobre la cual puede basarse la decisión y por lo tanto reduce tanto α como β. En una situación experimental las probabilidades de los errores de tipo I y II para una prueba miden el riesgo de tomar una decisión incorrecta. El experimentador selecciona los valores de estas probabilidades y la región de rechazo y el tamaño de muestra se escogen de acuerdo a ellas.

Problema ejemplo[14]: incurrir en un riesgo **β**:

---

[12] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 371-385
[13] Estadística para administración y economía. Mendenhall/Reinmuth. Grupo editorial Iberoameérica. P.p 149
[14] Kohler, Heinz. Estadística para negocios y economía. Editorial: CECSA. Primera reimpresión, México, 1998. 1053 páginas. P.p 371-386

El estadístico de una compañía está por probar si el ensamble de una computadora toma un promedio de 40 minutos, como era. ¿Cuáles son las implicaciones si el riesgo **β** de la prueba es igual a 0.2?

Solución:
Dadas las hipótesis:

$$H_0 : \mu_0 \leq 40 \qquad \text{y} \qquad H_1 : \mu_0 > 40$$

el procedimiento asegura lo siguiente: incluso si el tiempo de ensamble en efecto promedia más de 40 minutos, en el 20% de todas las pruebas la conclusión será lo contrario. Esto es equivalente a decir a las personas con SIDA, que no lo tienen; dichas “negativas falsas” constituyen el **error tipo II** de aceptación.

Sin embargo, en el 80% de dichas pruebas este tipo de error se evita, lo que indica la potencia de la prueba.

Es posible[15] determinar la probabilidad asociada con tomar una decisión correcta –no rechazar H_0 cuando es verdadera o rechazarla cuando es falsa. La probabilidad de no rechazar H_0 cuando es verdadera es igual a $1 - \alpha$.

Esto se puede demostrar notando que:

$$P_{(\text{rechazar Ho cuando es verdadera})} + P_{(\text{no rechazar Ho cuando es verdadera})} = 1$$

Como $P_{(\text{rechazar Ho cuando es verdadera})} = \beta$, tenemos:

$$P_{(\text{no rechazar Ho cuando es verdadera})} = 1 - \beta$$

Note que la probabilidad de no rechazar H_0 cuando es verdadera es el nivel de confianza $1 - \alpha$[16]

La probabilidad de rechazar H_0 cuando es falsa es igual a $1 - \beta$. esto se puede demostrar notando que:

$$P_{(\text{rechazar Ho cuando es falsa})} + P_{(\text{no rechazar Ho cuando es falsa})} = 1$$

Pero como: $P_{(\text{no rechazar Ho cuando es falsa})} = \beta$, tenemos:

$$P_{(\text{rechazar Ho cuando es falsa})} = 1 - \beta.$$

La probabilidad de rechazar la hipótesis nula H_0 cuando es falsa se llama **Potencia de la prueba**. Las probabilidades asociadas con los cuatro resultados posibles de un prueba de hipótesis[17] se resumen en la siguiente tabla:

Símbolo de la probabilidad	*Definición*
α	**Nivel de significancia: Probabilidad de un error tipo I**
β	**Probabilidad de un error tipo II**
$1 - \alpha$	**Nivel de confianza: Probabilidad de no rechazar H_0 cuando es verdadera**
$1 - \beta$.	**Potencia de la prueba: Probabilidad de rechazar H_0 cuando es falsa.**

[15] Weimer, Richard, C. "Estadística" Editorial: Cecsa. P.p 461

[16] estudiado en el capítulo 9 del libro: "Estadística" de Richard C. Weimer. Editorial CECSA.

[17] La prueba de hipótesis nunca se puede usar para "establecer" verdades absolutas, ya que se tiene la posibilidad de error con cualquier decisión. Cuando rechazamos la hipótesis nula, tenemos evidencia que indica que la hipótesis alternativa es plausible pero no necesariamente cierta; además, dejar de rechazar la hipótesis Ho no debe implicar que uno deba aceptar Ho, más bien, este juicio debe reservarse a menos de que se conozca la probabilidad de cometer un error tipo II. Si β es pequeña, uno puede concluir que Ho es plausible, aunque no necesariamente cierta. (Weimer, Richard, C. "Estadística" Editorial: Cecsa. P.p 461- 462)

A menudo[18] los datos muestrales sugieren que algo relevante está sucediendo en la población o proceso subyacente. Una muestra de clientes potenciales puede poner de manifiesto que una mayor proporción prefiere una nueva marca sobre la ya existente. Una muestra del tiempo que tardan los empleados de la oficina de reservaciones en atender las llamadas telefónicas puede mostrar que hay un incremento en el tiempo medio de espera por parte del cliente. Una muestra de los cigüeñales elaborados con una nueva aleación puede mostrar una disminución en la desviación estándar de la dureza del metal. En cada caso, los datos provienen de una muestra limitada y por lo mismo están sujetos a cierto grado de variación aleatoria.

La pregunta es si el resultado o el efecto aparente en la muestra es una indicación de que algo está sucediendo en la población (o proceso) subyacente o si el resultado observado es posiblemente una casualidad, un fruto de la variación aleatoria. Probar hipótesis estadísticas es una manera de estimar si los resultados aparentes en una muestra indican concluyentemente que en realidad algo está pasando.

Quienes investigan el mercado tienen una **hipótesis alternativa o de investigación**: que el nuevo producto es superior al anterior. Formalmente, una hipótesis alternativa, denotada con **H_1**, es un enunciado acerca de la población. La **hipótesis nula**, denotada con **Ho**, es la negación de la hipótesis alternativa **H_1**. Como el nombre sugiere, la hipótesis nula con frecuencia tiene una calidad negativa. En el ejemplo de la investigación del mercado, si ***p < 0.5**, el nuevo producto no se refiere a la versión anterior. Llamamos a **H_0: $p \leq 0.50$** la hipótesis nula porque niega o contradice nuestra hipótesis alternativa.

La hipótesis alternativa puede ser **unilateral o de una sola cola** (dirigida) o **bilateral, de dos colas** (no dirigida).

La estrategia básica en las pruebas de hipótesis es tratar de apoyar la hipótesis alternativa "contradiciendo" la hipótesis nula. Se "contradice" a la hipótesis nula si los datos de la muestra son poco creíbles dada H_0 y sumamente verosímiles dada H_1. Así, para apoyar **Ha: $p > 0.50$.**

Los datos se deben sintetizar en un **estadístico de prueba** (E:P). dicho estadístico se calcula para ver si es razonablemente compatible con la hipótesis nula. Cuando se prueba una proporción el estadístico de la prueba:
E:P: Y = Numero de éxitos.

En el ejemplo del nuevo producto contra el anterior, suponemos que éste es al menos tan bueno como aquél. Dada la hipótesis, es muy poco probable que **Y**, el número de consumidores en la muestra que prefieren al nuevo producto, sea muy grande. Así, si **Y** resulta ser muy grande, rechazamos la hipótesis nula y apoyamos la hipótesis alternativa de que el nuevo producto es mejor. Para ser más precisos, digamos que la lógica básica es la siguiente:

[18] Estadística aplicada a la administración y a la economía. David K. Hildebrand y R. Lyman Ott. Addison Wesley Longman.

1. suponga que **Ho** es cierta (**$p \leq 0.50$**).
2. calcule el valor del **E.P : Y = número de clientes en la muestra que prefieren el nuevo producto;**
3. si este valor es inverosímil (lo que, en este caso, significa muy grande), rechace **H_0** y acepte **H_1**.

El número de pasos que integran una prueba de hipótesis llega a variar dependiendo del autor del libro en el cual se encuentran. Así por ejemplo, los siguientes se consideran **los seis pasos de una prueba estadística**[19]:

1. Hipótesis nula **H_0: $p \leq 0.50$**
2. Hipótesis alternativa o de investigación **H_1: $p > 0.50$**
3. estadístico de la prueba **E.P ; Y = número de cliente que prefieren el nuevo producto.**
4. Región de rechazo
5. región de aceptación
6. conclusión.

Procedimientos de prueba equivalentes[20]:

Hay tres procedimientos equivalentes para decidir si la hipótesis nula **H_0** es verdadera o plausible:

Procedimiento # 1. compare el valor del estadístico de prueba calculado de la muestra con el valor crítico obtenido de la distribución muestral de la media.

Procedimiento # 2. encuentre el puntaje ***z*** para el valor del estadístico de prueba y compárelo con el valor ***z*** correspondiente a un área igual a el valor del nivel de significancia en la cola que corresponda de la distribución normal estándar.

Procedimiento # 3. encuentre la probabilidad de obtener un valor tan grande como el estadístico de prueba y compárelo con el nivel de significancia correspondiente. Si esta probabilidad es menor o igual que el nivel de significancia, rechace Ho, en otro caso, no lo haga.

Ejemplo de pruebas de hipótesis[21] utilizando el primer procedimiento:

Suponga que cierta máquina de refrescos operada con monedas fue diseñada para servir, en promedio 8 onzas de bebida por vaso; después de un largo periodo de uso sospechamos que está expidiendo en promedio, menos de 8 onzas por vaso y resolvimos probar las hipótesis siguientes:

$$\text{Ho: } \mu \geq 8$$
$$H_1: \ \mu < 8$$

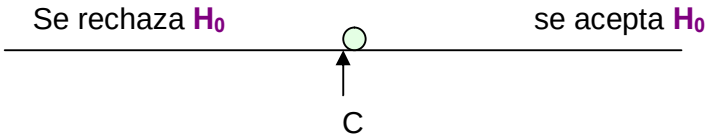
[19] Estadística aplicada a la administración y a la economía. David K. Hildebrand y R. Lyman Ott. Addison Wesley Longman.
[20] Weimer, Richard C. "Estadística". Editorial: CECSA. P.p 466-472
[21] Weimer, Richard C. "Estadística". Editorial: CECSA. P.p 466-468

En este caso en particular, el problema de exponer dos hipótesis opuestas esta resuelto. Y si podemos rechazar **Ho**, concluimos que nuestras sospechas son correctas.

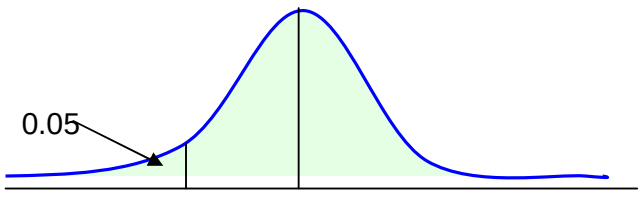
Como siguiente paso, podemos obtener una muestra aleatoria de 30 vasos y calcular la media muestral. Supongamos que la media muestral es de 7.6 onzas y que la desviación estándar muestral es de 0.75. y escogemos la media muestral como el estadístico de prueba.

A continuación, estipulamos el nivel de significancia α**,** eso nos ayudará a formular la regla de decisión; este nivel puede ser cualquier valor entre 0 y 1, aunque suele ser de 0.05 (es decir de 5%) o 0.01. usemos $\alpha = 0.05$**,** esto significa que la probabilidad de rechazar H_0 cuando es cierta es 0.05; como estamos interesados en producir una evidencia que apoye la veracidad de H_1, suponemos que H_0 es verdadera esperando rechazarla. La afirmación de que H_0 es verdadera es sólo una hipótesis que debe ser probada en cuanto a su veracidad o falsedad. Resultado, procedemos bajo la hipótesis de que H_0 es verdadera.

Como el nivel de significancia es $\alpha = 0.05$ y la prueba es de cola izquierda (porque: $H_0: \mu \geq 8$, se puede establecer una región de rechazo, localizando el valor en la distribución muestral de la media que tiene 5% de las medias muestrales por debajo de él; en otras palabras, se puede determinar un valor que determine un área de 0.05 en la cola izquierda de la distribución muestral, este valor se llama **valor crítico**, y separa la distribución muestral en dos regiones: la región de rechazo y la de aceptación; denotaremos un valor crítico por el símbolo **C** y todos los valores a su izquierda forman la región de rechazo. Así, si la media muestral cae en la región de rechazo, rechazamos H_0; de otra manera, dejamos de rechazar H_0, como se indica en la siguiente figura:



Como la distribución muestral de la media es aproximadamente normal, para encontrar el valor crítico **C**, localizamos su valor **z** y nos referimos a la distribución normal estándar



y como la desviación estándar de la población es desconocida y **n ≥ 30**, entonces la desviación estándar de la muestra es una estimación puntual adecuada para la desviación estándar de la población. Por lo tanto:

$$z = \frac{C - \mu}{\frac{S}{\sqrt{n}}}$$

de donde sustituyendo valores tenemos que:

$$-1.65 = \frac{C - 8}{\frac{0.75}{30}}$$

el valor 1.65 se encuentra en la tabla **z** y corresponde a un área de

0.5-0.05 = 0.45;

y si despejamos **C** de la última ecuación, el valor crítico es: **C = 7.77**

como siguiente paso, es necesario localizar el valor de la media muestral de 7.6 en su distribución muestral. Como una consecuencia del teorema del límite central y el hecho de que **n = 30**, la distribución muestral de la media es aproximadamente normal; su media es **μ,** la media de la población de la que procede la muestra. Como suponemos cierta **H_0**, la media de la distribución muestral es mayor o igual que 8 onzas. ¿Qué valor debemos usar para **μ**? Usaremos siempre el valor de la igualdad expresada en Ho, llamado valor nulo, entonces, **μ = 8** es el valor nulo. Si se usa un valor mayor que 8, se demostraría que la probabilidad de rechazar **H_0** cuando es cierta es menor que el nivel estipulado de significancia **α = 0.05;** esto es, el valor nulo presenta el peor caso en términos de proporcionar la mayor probabilidad posible de rechazar **H_0** cuando se supone que es cierta; es por eso que en lo que resta del presente texto, siempre emplearemos el valor nulo como la media de la distribución muestral del estadístico de prueba.

Como $\bar{X}$ **= 7.6 < 7.77** rechazamos **H_0** a favor de **H_1**, con esta decisión, nos arriesgamos a cometer un error del tipo I, rechazar **H_0** cuando es cierta; la probabilidad de este tipo de error es **α = 0.05.** Cuando una prueba da lugar a rechazar una hipótesis nula bajo un nivel de significancia específico, se dice que la prueba es significativa y el resultado se denomina resultado significativo. Para nuestro ejemplo, el resultado **μ < 8** se clasifica como significativo. Luego entonces, hemos encontrado evidencia estadística significativa que indica que la máquina está despachando en promedio, menos de 8 onzas de bebida por vaso.

Un procedimiento de prueba se considera como bueno cuando tanto las probabilidades de suceso del error tipo I como del II, son pequeñas.

Tenemos control sobre la probabilidad del error tipo I α porque se ha estipulado antes de obtener los datos. En general, tenemos poco control

sobre β, la probabilidad del error tipo II, esta probabilidad varía dependiendo del verdadero valor de parámetro poblacional.

Por ejemplo, si queremos poner a prueba la hipótesis nula **Ho: μ = 25** y se asegura que el verdadero valor de μ es 26, entonces se puede calcular un valor para β, dado un tamaño de muestra fijo, depende de la diferencia entre el valor que se asegura y el supuesto.

Es deseable minimizar las probabilidades de ambos tipos de error, para un nivel de significancia fijo α, se puede, en general, mantener β en un mínimo escogiendo el tamaño de muestra tan grande como sea posible. **Recuerde que la probabilidad de rechazar la hipótesis nula cuando es falsa se llama la potencia de la prueba y se denota por 1 - β pudiéndolo incrementar aumentando el tamaño *n* de la muestra.**

IV.3. PRUEBAS DE HIPOTEISIS SOBRE VARIANZAS.

Siguiendo la metodología 1, considere que se está estudiando que tan estable es el proceso de fabricación de una batería recargable. Para lo cual los ingenieros tienen la conjetura que en promedio la duración de la batería es de 100 horas pero con una variabilidad de ½ hora. Considere que en este caso se desea probar la aseveración de que la batería tiene una variabilidad de menos de ½ hora (es decir $\sigma \leq 0.5$ ó bien $\sigma^2 \leq 0.25$) contra el hecho que $\sigma \geq 0.5$, para lo cual se extrae una muestra de 40 baterías y se les mide el tiempo de funcionamiento, obteniendo S=0.52.

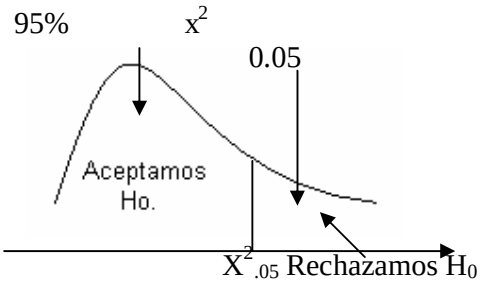
En este caso, la prueba de hipótesis a realizar es:

Ho: $\sigma^2 \leq 0.25$
H_1: $\sigma^2 > 0.25$

El estadístico de prueba en este caso es:

$$\frac{(n-1)S^2}{\sigma^2} \; que\ se\ distribuye\ como\ X^2\ con\ n-1\ g.l.$$

La región de rechazo y aceptación de la hipótesis nula de acuerdo a un nivel de significancia del 95% es:



$$\alpha = 0.05$$

Y en este caso, buscando en tablas tenemos que el punto crítico $X^2_{0.05}$ es igual a 54.572. Pero como el estadístico calculado, suponiendo H_0 cierta es:

$$\frac{(n-1)S^2}{\sigma^2}=\frac{(39)(0.52)^2}{(0.5)^2}=42.1824$$

Valor que está dentro de la región de aceptación de H_0 y por lo tanto no hay evidencia en la muestra para rechazar la conjetura que el proceso de fabricación es estable (variabilidad menor a menos de ½ hora).

IV.3. PRUEBAS DE HIPOTEISIS SOBRE MEDIAS DE DOS POBLACIONES.

Considere el ejemplo del gerente de una refinería que produce gasolina, quién piensa modificar el proceso para producir gasolina a partir de petróleo crudo. El gerente hará la modificación sólo si la gasolina promedio que se obtiene en este nuevo proceso (expresada como un porcentaje del crudo) aumenta su valor con respecto al proceso en uso con base en un experimento de laboratorio y mediante el empleo de 2 muestras aleatorias de tamaño 12, una para cada proceso, la cantidad de gasolina promedio del proceso en uso es de $\overline{X_1}=24.6$ y $S_1=2.3$, y para en proceso propuesto $\overline{X}_2=28.2$ y $S_2=2.7$.

Suponiendo que los dos procesos son independientes, normalmente distribuidos y con varianzas iguales, ¿Debe adoptarse el nuevo proceso?

Solución:

En este caso para hacer la prueba se supondrá $\sigma_1=\sigma_2$, de otro modo varía la metodología (este supuesto puede verificarse también realizando una prueba sobre varianzas).

Ho: $\mu_1=\mu_2$ *ó bien* $\mu_1-\mu_2=0$

Ha: $\mu_2\geq\mu_1$ *ó bien* $\mu_1-\mu_2\leq 0$

$$\alpha=0.05$$

En este caso, el estadístico de prueba debe ser tal que involucre la diferencia de medias poblacionales (μ_1-μ_2) y tenga distribución totalmente conocida, para lo cual se puede tener el estadístico:

$$\frac{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\approx t(n_1+n_2-2)$$

Que como se indica se distribuye como t-student con n_1+n_2-2 grados de libertad. La prueba de esto es la siguiente:

$\overline{X_1} \cap \cup N(\mu_1, \sigma_1^2)$

$\overline{X}_2 \cap \cup N(\mu_2, \sigma_2^2)$

$E(\overline{X}_1) = \mu_1$

$E(\overline{X}_2) = \mu_2$

y si las muestras son independie ntes :

$VAR\ (\overline{x_1} - \overline{x_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$= \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$ *ya que se* sup *one* $\sigma_1^2 = \sigma_2^2$

Así que :

$\frac{(\overline{X_1} - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ *se distribuye normal estándar* $N(0,1)$

pero sabemos que una $t -$ *student se obtiene como* :

$$\frac{X}{\sqrt{(X_1^2 + X_2^2 + ... + X_n^2)/n}} \approx t(n);$$

es decir es una $N(0,1)$ *entre* $\sqrt{Ji - cuadrada\ /n}$

pero sabemos que :

$\frac{(n-1)S^2}{\sigma^2}$ *se distribuye como* $X^2(n-1)$

Así que :

$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}$ *se distribuye como* $X^2(n_1 + n_2 - 2)$

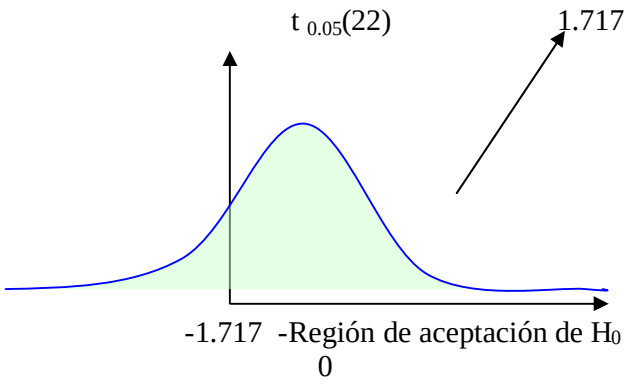
y entonces :

$$\frac{\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}} \Big/ \sqrt{n_1 + n_2 - 2}} \approx t(n_1 + n_2 - 2)$$

Que simplificando queda:

$$\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t(n_1 + n_2 - 2)$$

La para el ejemplo es de una sola cola:



Y al evaluar el estadístico, suponiendo cierta H_0, se tiene:

$$\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{24.6 - 28.2}{\sqrt{\frac{11(2.3)^2 + 11(2.7)^2}{22}}\sqrt{\frac{1}{12} + \frac{1}{12}}} = -3.51$$

Por lo cual como -3.51 < -1.717, entonces el estadístico cae en la región de rechazo de H_0, concluyéndose que la evidencia de la muestra indica que el nuevo proceso aumenta el nivel de gasolina obtenida.

IV.4. PRUEBAS DE HIPOTEISIS SOBRE VARIANZAS DE DOS POBLACIONES.

Considere el ejemplo de dos acciones de la Bolsa Mexicana de Valores y se desea saber que acción es la más inestable, en el sentido que tiene una varianza mayor.

En los últimos meses se ha considerado que la acción de Walmex V (con varianza poblacional igual a σ_1) es en promedio igual de inestable que la de Cemex CPO (con varianza σ_2), pero se desea analizar si está hipótesis se mantiene. Para lo cual se extrae una muestra de 51 rendimientos diarios de Cemex y 41 de Walmex, obteniéndose que $S_{1=}15\%$ y $S_{2=}10\%$.

Así, en este caso se desea realizar la prueba siguiente:

$$Ho: \sigma^2_1 = \sigma^2_2 \quad vs. \quad Ha: \sigma^2_1 > \sigma^2_2$$

En este caso necesitamos un estadístico de prueba que involucre a σ^2_1 y σ^2_2 y cuyo modelo de distribución sea totalmente conocida. A saber el estadístico es:

$$\frac{S^2_1}{S^2_2}$$

El cual tiene distribución F con n-1 grados de libertad en el numerador y m-1 grados de libertad en el denominador.

Lo anterior se demuestra como sigue:

$\frac{(n-1)S^2_1}{\sigma^2_1}$ se distribuye como Ji - cuadrada con n - 1 g.l.

y

$\frac{(m-1)S^2_2}{\sigma^2_2}$ se distribuye como Ji - cuadrada con m - 1 g.l.

Así que el cociente de las dos anteriores entre los grados de libertad, se distribuye como una F con (n -1) y (m - 1) g.l.

$$\frac{\left(\frac{(n-1)S^2_1}{\sigma^2_1}\right)/(n-1)}{\left(\frac{(m-1)S^2_2}{\sigma^2_2}\right)/(m-1)} \quad ...(1)$$

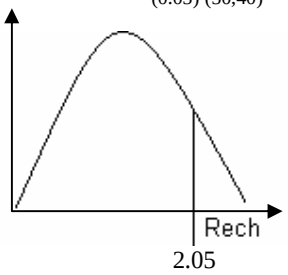
pero si Ho es cierta entonces $\sigma^2_1 = \sigma^2_2$ y (1) se reduce a :

$\frac{S^2_1}{S^2_2}$ que se comporta como F con (n -1) g.l. en el número ador y (m - 1) g.l. en el denominador

Por lo anterior, al calcular el estadístico bajo Ho, se tiene:

$$\frac{S^2_1}{S^2_2} = \frac{(15)^2}{(10)^2} = 2.25$$

Y revisando en tablas de F se tiene que una el cuantil de 0.05 de una F con 50 g.l. en el numerador y 40 g.l. en el denominador es $F_{(0.05)\,(50,40)} = 2.05$, así que como la región de rechazo es:



1-α = 95%

Y como el estadístico es igual a 2.25 que es mayor que 2.05 (valor de la F), concluimos que la acción de Cemex es más riesgosa que la de Walmex.

IV.5. COMPARACIÓN DE LAS TÉCNICAS DE INTERVALOS DE CONFIANZA Y DE LA PRUEBA DE HIPÓTESIS DE DOS COLAS:

Hay dos técnicas clásicas para hacer inferencias sobre el valor de un parámetro desconocido:

1. **la estimación[22] y**
2. **la prueba de hipótesis.**

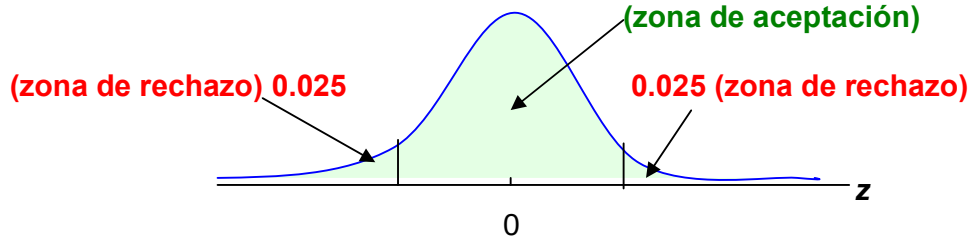
Una comparación de un parámetro desconocido con una constante conocida que utiliza una prueba de dos colas con un nivel de significancia igual a α, se puede hacer construyendo un intervalo del **$(1 - \alpha)100\%$** de confianza para el parámetro. Si el valor supuesto del parámetro está contenido en el intervalo de confianza, entonces no podemos concluir que ese parámetro sea distinto de la constante conocida.

Vemos el siguiente ejemplo:
Se supone que una tableta para bajar la temperatura contiene 10 gramos (0.648 g) de aspirina. Una muestra aleatoria de 100 tabletas produjo una media de 10.2 gramos y una desviación estándar de 1.4. ¿Podemos concluir que μ es diferente de 10 con un nivel de significancia del 5%?

Resolvamos este ejemplo, utilizando la prueba de hipótesis:
Paso # 1. primero establecemos las dos hipótesis opuestas y dado que se supone que la tableta contiene10 gramos de aspirina, entonces:

$$H_o: \mu = 10$$
$$H_1: \mu \neq 10$$

Observe que dado que aparece el signo de igualdad en la hipótesis nula, entonces **la prueba es de dos colas** (no direccional) y la región de rechazo consiste de los valores en las colas izquierda y derecha de la distribución. **Como la probabilidad de cometer un error tipo I, (rechazar H_0 cuando es cierta) es 0.05 y la región de rechazo se ubica en ambas colas, colocamos $\frac{\alpha}{2}$ = 0.025 de la distribución en cada una de las regiones de las colas, tal y como se indica en la siguiente figura:**

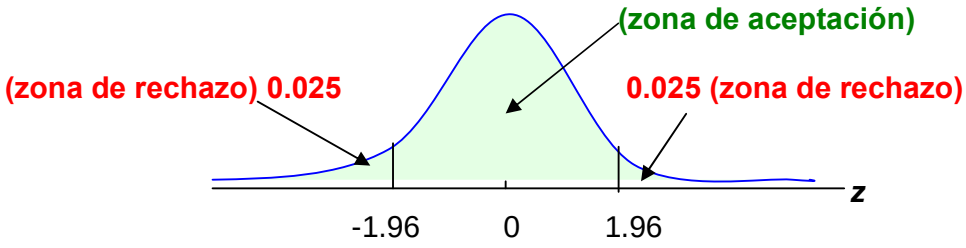


[22] Weimer, Richard C. "Estadística". Editorial: CECSA. Cap. 9

paso # 2. Selección del estadístico de prueba: El estadístico de prueba es el valor de **z** para $\bar{X}$. como se desconoce σ, **n = 100**, la desviación estándar muestral **s** proporciona un buen estimado para σ. Por lo tanto:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

paso # 3. derivar una regla de decisión: **rechazar H_0 si $z < -z_{0.025}$ o $z > z_{0.025}$** resulta claro al utilizar una tabla de distribución normal que los valores crítico son: $\pm z_{0.025} = \pm 1.96$, tal y como se muestra en la siguiente figura:



paso # 4. toma de la muestra, calculo del estadístico de prueba y confrontación del mismo con la regla de decisión:

para este caso, tenemos que los datos son:

n = 100
$\bar{X}$ = 10.2
μ = 10
σ = 1.4

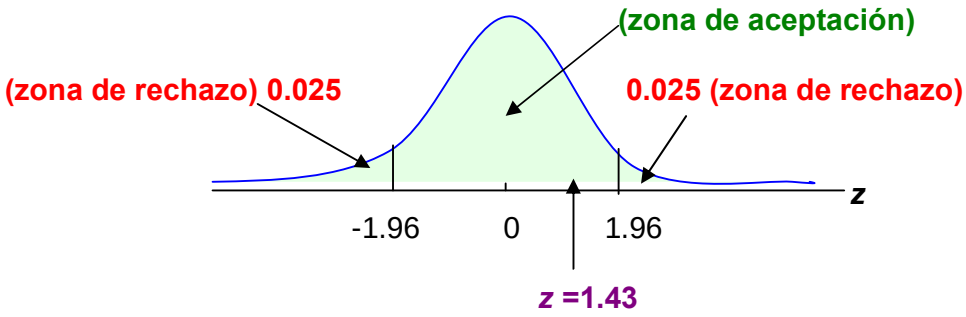
y teniendo en cuenta que el estadístico de prueba es:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

entonces, al sustituir datos en el estadístico de prueba tenemos que:

$$z = \frac{10.2 - 10}{\frac{1.4}{\sqrt{100}}}$$

para finalmente al realizar operaciones obtenemos el valor: ***z* = 1.43** y al confrontarlo con la regla de decisión finalmente vemos que:



el valor de ***z*** cae dentro de la zona de aceptación, por lo tanto, aceptamos la hipótesis nula H_0, con lo cual concluimos que no hay evidencia estadística de que μ sea diferente de 10. aceptar H_0 se interpreta como que nuestra evidencia es estadísticamente significativa con α = 5%.

Nota: existe la posibilidad de cometer un error tipo II, pues H_0 puede ser falsa y no la rechazamos; la probabilidad β en este caso es desconocida; en consecuencia, el experimentador debe reservarse el juicio sobre H_0 hasta obtener más datos, en este caso, la decisión es no rechazar H_0. Como lo dijimos antes, esta decisión no implica que H_0 se acepta como verdadera o plausible.

Solución utilizando intervalos de confianza:
Si ahora construimos un intervalo de confianza del 95% de confianza para el promedio del contenido de aspirina, tenemos que recordando que los límites del intervalo de confianza se encuentran usando:

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

y teniendo en cuenta que el valor crítico es: $z_{0.025}=1.96$, que $n = 100$ y que σ es desconocida, ***s*** proporciona un buen estimado de σ. En consecuencia los límites son:

$$10.2 \pm 1.96 \frac{1.4}{\sqrt{100}} = 10.2 \pm 0.27$$

es decir, que un intervalo del 95% de confianza para μ es (9.93, 10.47). por lo tanto, como el valor supuesto 10 está contenido en el intervalo, no podemos concluir que $\mu \neq 10$. Nota: este resultado da la misma conclusión a la que llegamos usando el procedimiento de prueba de hipótesis.

Conclusiones:
Un intervalo de confianza proporciona más información que una prueba de hipótesis; con base en los datos, pudimos rechazar la hipótesis nula y encontrar que el resultado no tenía importancia práctica, pero si usamos el intervalo de confianza correspondiente y un poco de sentido común podemos determinar si los resultados de la prueba de hipótesis son de importancia práctica.

Para ilustrar esto, consideremos una situación hipotética en la cual se está probando una nueva guía de estudio para mejorar las calificaciones de Matemáticas en cierta universidad. Se sabe que la media y la desviación estándar de dichas calificaciones son 500 y 100 respectivamente. Se usa una muestra de 1 millón de estudiantes para determinar si la nueva guía de estudios produce una media de calificaciones de Matemáticas diferente de 500; la hipótesis nula es que la media de las calificaciones de Matemáticas en dicha universidad para estudiantes que usan la guía nueva es de 500, la misma que cuando el grupo no la usaba. Así, la hipótesis nula es:

$$\mathbf{Ho : \mu = 500}$$

Supongamos que el grupo que usa la nueva guía tiene una media de 500 en sus calificaciones. Si usamos la fórmula:

$$\bar{X} \pm Z_{0.05} \frac{s}{\sqrt{n}}$$

encontramos que el intervalo del 95% de confianza para μ es:

$$500.4 \pm 1.96 \frac{100}{\sqrt{1000000}}$$

$$500.4 \pm 1.96(0.1)$$

es decir, que el intervalo tendría los límites siguientes:

(500.2 , 500.6)

como 500 no está contenido en el intervalo, debemos rechazar la hipótesis nula y concluir que la nueva guía de estudio tiene un efecto estadísticamente significativo en las calificaciones de los exámenes de Matemáticas de la mencionada Universidad. La calificación promedio para el grupo en el examen, $\bar{X} = 500.4$ está cuatro desviaciones estándar arriba de la media de su distribución muestral. Pero ¿Tiene alguna importancia práctica este resultado significativo? Seguramente pocas personas, si hay alguna, usaría la nueva guía de estudio para elevar su calificación un 0.4 de punto.

El ejemplo anterior, aunque hipotético, sirve para ilustrar los siguientes puntos:

1. **Una prueba de hipótesis puede producir resultados significativos que no tengan importancia práctica.**
2. **Un tamaño de muestra grande aumenta la posibilidad de rechazar la hipótesis nula.**

En el capítulo anterior aprendimos que cuando el tamaño de la muestra crece, el ancho del intervalo de confianza tiende a cero. Así, cuando la muestra es la población completa, $\bar{X} = \mu$ y el ancho del intervalo de confianza es cero, cualquier hipótesis nula: **Ho : $\mu = \mu_o$** se rechazaría salvo para el caso en que μ_o sea el verdadero valor de μ.

Desde un punto de vista teórico, cualquier hipótesis nula se puede rechazar si escogemos una muestra suficientemente grande, uno puede concluir entonces que dejar de rechazar una hipótesis nula es el resultado de que la muestra no sea suficientemente grande. Desde luego, en muchas aplicaciones prácticas que involucran la prueba de hipótesis, la cantidad de datos está basada en consideraciones económicas, así como en la naturaleza del experimento. Para algunos experimentos, como el estudio de enfermedades raras, no es posible obtener una gran cantidad de datos.

Pruebas de hipótesis (muestras pequeñas)[23]

Introducción:

En capítulos anteriores se estudió un tipo de prueba de hipótesis estadística. Se utilizó la distribución normal estándar, que es la distribución "z", como estadístico de prueba. Para emplear la distribución "z" es necesario conocer la desviación estándar (sigma) de la población o tener una muestra grande (de 30 observaciones por lo menos).

Sin embargo, en muchas situaciones no se conoce sigma y el número de observaciones en la muestra es menor de 30. en estos casos, se puede utilizar la desviación estándar de la muestra "s" como una estimación de σ "sigma", pero no es posible usar la distribución "z" como estadístico de prueba. El estadístico de prueba adecuado es la **t de student**, o simplemente **distribución t**. Cuando se utiliza la **t de student**, se supone que la población tiene una distribución normal.

Para iniciar este capítulo, se describen las características de la **distribución t**. Luego se estudian tres situaciones de prueba de hipótesis.

Características de la distribución t de student.

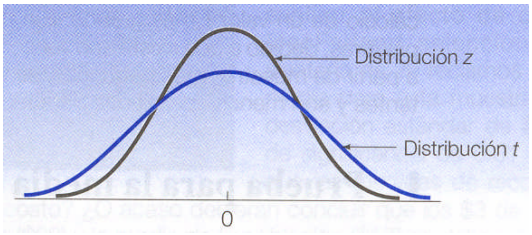
William S. Gossett[24] desarrolló la distribución t de student. Él se interesó por la distribución exacta de:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

[23] Mason, R. D. et al. (1994) "***Estadística para administración y economía***"(3ª. Edición) México. Irwin-McGraw-Hill. Pp 307

[24] William Gosset nació en Inglaterra en 1876 y murió allí mismo en 1937. trabajó durante muchos años en la cervecera Arthur Guinnes e hijos. De hecho, en sus últimos años estuvo a cargo de la cervecera Guinnes en Londres. Guinnes prefería que para publicar documentos, sus empleados usaran seudónimos, por lo que en 1908, cuando Gossett escribió "El error probable de una media", empleó el nombre de "Student". En este documento describió las propiedades de la distribución t.

donde se utiliza a “s” como estimador de σ, lo que es preocupante en especial por la discrepancia entre “s” y s” cuando σ se calcula con base en una muestra pequeña. En el siguiente diagrama se muestran la distribución t y la distribución normal estándar.



Observe en particular que la distribución t es más plana y amplia que la distribución “z”.

Si la población de interés es normal, la distribución “t” tendrá las siguientes características[25]:

1. al igual que la distribución “z”, es una distribución continua.
2. al igual que la distribución “z”, tiene forma acampanada y simétrica.
3. no hay una distribución “t”, sino una “familia” de distribuciones “t”. Todas con la misma media cero, pero con su respectiva desviación estándar diferente de acuerdo con el tamaño de la muestra “n”. Existe una distribución “t” para una muestra de 20, otra para una muestra de 22, y así sucesivamente.
4. la distribución “t” es más ancha y más plana en el centro que la distribución normal estándar. Sin embargo, a medida que aumenta el tamaño de la muestra, la distribución “t” se aproxima a la distribución normal estándar.

[25] Mason, R. D. et al. (1994) “***Estadística para administración y economía***”(3ª. Edición) México. Irwin-McGraw-Hill. Pp 307

CAPITULO IV ESTADÍSTICA NO PARAMÉTRICA

V.1. INTRODUCCIÓN.

Las pruebas no paramétricas son útiles sobre todo cuando no se conoce la distribución del cual provienen los datos y por tanto no se conoce la distribución del estadístico para hacer una estimación por intervalos de confianza o una prueba de hipótesis. Estas pruebas son útiles por ejemplo cuando el tipo de datos es nominal u ordinal. En este capítulo se considerarán cuatro tipos de pruebas libres de distribución que necesitan una ordenación por rango: prueba de signo, prueba U de Mann-Withney, prueba de rangos de Kruskal-Wallis y prueba de rangos con signo por pares ajustados de Wilcoxon.

V.2. PRUEBA DE SIGNO.

Esta prueba se refiere a un cambio de signo en el comportamiento de una variable, por ejemplo si las ventas de enero fueron de 950,000 y para febrero fueron de 870,000 entonces hubo una variación negativa (signo negativo), o bien si en un programa de capacitación antes había deficiencias en las operaciones y después del programa mejora el desempeño, entonces hay una variación positiva (signo positivo).

MUESTRAS PEQUEÑAS.

Ejemplo.

Considere el caso de una compañía que ha implementado un programa de capacitación en cómputo para cubrir las deficiencias en el sistema administrativo. El programa se implementará en todas las dependencias de la empresa, y una vez llevado a cabo el programa, se desea saber si éste fue capaz de mejorar la capacidad de los empleados para el manejo de la computadora.

Del total de empleados participantes en el programa se seleccionó una muestra de tamaño 10, y los resultados antes y después del programa se indican a continuación:

Habilidad en el manejo de la computadora antes y después de programa			
Nombre	Antes	Después	Signo de Diferencia
JUAN HERNANDÉZ	BUENO	SORESALIENTE	+
PEDRO GÓMEZ	ACEPTABLE	EXCELENTE	+
PABLO PÉREZ	EXCELENTE	BUENO	-
JULIO ACOSTA	DEFICIENTE	BUENO	+
JANETT ORDAZ	BUENO	SOBRESALIENTE	+
ALMA VALDEZ	DEFICIENTE	ACEPTABLE	+
EMMA PINEDA	EXCELENTE	SOBRESALIENTE	+
LUIS TENA	BUENO	DEFICIENTE	-
ARLETT CRUZ	DEFICIENTE	BUENO	+
ALFONSO SANCHEZ	BUENO	SOBRESALIENTE	+
MARIA BADILLO	ACEPTABLE	EXCELENTE	+
HILDA LÓPEZ	BUENO	ACEPTABLE	-
KARLA MARTINEZ	BUENO	SOBRESALIENTE	+

LUZ ESTRADA	DEFICIENTE	BUENO	+

Las calificaciones están en escala ordinal como sigue: Sobresaliente, Excelentes, Buenas, Aceptables y Deficientes.

Entonces en este caso, la hipótesis a probar es:

Ho: p = 0.5 (el programa no ha tenido repercusiones)
Vs.
Ha: p > 0.5 (el programa mejoró la capacitación)

El estadístico de prueba en este caso es la distribución binomial, ya que:

1.- Sólo hay dos posibles resultados o se mejoró (éxito) o no se mejoró (fracaso).
2.- La probabilidad de éxito bajo Ho es p=0.5.
3.- El número de ensayos es fijo (el tamaño de muestra).
4.- Cada ensayo es independiente (el resultado en un empleado es independiente de otro).

Para realizar la prueba calculamos los valores acumulados de una binomial con parámetros n=14 y p=0.5, y si utilizamos un nivel de significancia de 0.10, entonces el cuantil que deja 0.10 de área a la derecha de una binomial es de acuerdo a la tabla siguiente es para x = 10 éxitos (que es el más cercano)

La probabilidad de éxito se calculó utilizando la función DISTR.BINOM(número de éxitos;14;0.50;0) de Excel y truncado a cinco dígitos decimales.

Número de Éxitos	Probabilidad de Éxito	Probabilidad Acumulada
0	0,00006	0,9999
1	0,00085	0,9999
2	0,00555	0,9990
3	0,02221	0,9935
4	0,06109	0,9713
5	0,12219	0,9102
6	0,18328	0,7880
7	0,20947	0,6047
8	0,18328	0,3952
9	0,12219	0,2120
10	0,06109	0,0898
11	0,02221	0,0287
12	0,00555	0,0065
13	0,00085	0,0009
14	0,00006	0,0001

Así que la regla de decisión es rechazar Ho si el número de signos positivos es de 10 ó más. Y como en el ejemplo el número de signos positivos es 11 entonces se puede concluir que el programa de capacitación si mejoró la capacidad en el manejo de la computadora.

MUESTRAS GRANDES.

Si el número de pares utilizados es mayor a 20, se considera en este caso muestra grande y en vez de aplicar una distribución binomial, se puede utilizar una distribución normal y en este caso deberá cumplirse que tanto np como n(1-p) sean mayores a 5.

Debido a que la media de una binomial es np y la varianza es np(1-p), entonces para utilizar la normal se considera como media $\mu = np$ y $\sigma = \sqrt{np(1-p)}$.

El estadístico de prueba para una prueba de dos colas es:

$$Z = \frac{(x \pm p) - \mu}{\sigma}$$

Y si el número de signos + o – es mayor que n/2, se usa el siguiente estadístico de prueba:

$$Z = \frac{(x - p) - \mu}{\sigma} = \frac{(x - p) - p(n)}{\sqrt{p(1-p)n}}$$

Mientras que si el número de signos es menor que n/2, el estadístico de prueba es:

$$Z = \frac{(x + p) - \mu}{\sigma} = \frac{(x + p) - p(n)}{\sqrt{p(1-p)n}}$$

Ejemplo.

Considere el caso de un estudio de mercado respecto de un nuevo producto de bebidas embotelladas en dos versiones la dulce (A) y la amarga (B). Se desea saber cual de las dos se lanzará al mercado y para ello se realiza una encuesta a 64 personas (clientes reales) y se les pregunta cuál bebida prefieren, obteniendo como resultado que 46 prefieren la bebida A (dulce), el signo "+" se asignó al tipo A y el signo "–" al tipo B.

En este caso la hipótesis a probar con un nivel de significancia de 0.05 es:

Ho: p = 0.5 (no hay preferencia)
Vs.
Ha: p > 0.5 (hay preferencia de la A sobre la B)

Puesto que 46 es mayor que n/2 = 32, entonces el estadístico de prueba es:

$$Z = \frac{(x - p) - \mu}{\sigma} = \frac{(x - p) - p(n)}{\sqrt{p(1-p)n}} = \frac{(46 - 0.50) - 0.50(64)}{0.50\sqrt{64}} = 3.375$$

Por su parte, el cuantil de una normal que deja un área de 0.05 a su derecha es 1.64.

Y como 3.375 > 1.64 concluimos que la bebida A se prefiere sobre la bebida B.

V.3. PRUEBA DE HIPÓTESIS SOBRE LA MEDIANA.

Ejemplo.

Suponga que una gran cadena de tiendas desea probar la hipótesis de que la mediana de las ventas de abarrotes es $ 1,300 diarios. Para ello se extrae una muestra aleatoria de 102 notas la cual reveló que 60 de ellas eran superiores a $ 1,300 y 40 eran menores a $ 1,300.

En este caso es conveniente formular la hipótesis mediante:

Ho: mediana = $ 1,300
Vs.
Ha: mediana ≠ $ 1,300 Con un nivel de significancia de $\alpha = 0.10$

Se puede aplicar la prueba de manera similar a la prueba de signos, donde en este caso como 60 elementos de la muestra son mayores a $ 1,300, entonces hay 60 elementos con signo "+" y a su vez hay 40 con signo "-", descartándose los elementos iguales a $1,300 y como 60 es mayor que n/2 = 100/2 = 50, entonces el estadístico de prueba es:

$$Z = \frac{(x-p)-\mu}{\sigma} = \frac{(x-p)-p(n)}{\sqrt{p(1-p)n}} = \frac{(60-0.50)-0.50(100)}{0.50\sqrt{100}} = 1.90$$

Y por su parte, como es una prueba de dos colas, entonces el cuantil $\alpha/2 = 0.05$ de una normal es 1.64. Por lo cual si el estadístico queda a la izquierda de –1.64 ó a la derecha de 1.64 rechazamos Ho. Pero como el estadístico es 1.90 que está a la derecha de 1.64, entonces se puede concluir que el nivel de ventas no tiene como mediana a $ 1,300.

V.4. PRUEBA U DE MANN-WHITNEY.

Esta prueba es útil cuando se seleccionan dos conjuntos aleatorios independientes y su escala es de tipo ordinal al menos. La prueba consiste en determinar si las dos muestras presentan los mismos promedios poblacionales o no (prueba de medias). Para está prueba se considerará que el estadístico de prueba se comportará como una distribución de Mann-Whitney, y en ocasiones se prefiere a diferencia de la t-student, debido a que en ocasiones la varianza de las dos poblaciones son independientes o más aún los datos son de tipo ordinal.

MUESTRAS PEQUEÑAS

Ejemplo.

En una empresa se está haciendo una prueba de aptitud mecánica en la línea de producción, y se desea saber si la aptitud mecánica de los hombres es la misma que la

de las mujeres o son distintas (prueba de dos colas). Para ello se extrae una muestra de 9 hombres y 5 mujeres y se les calificó en puntos el nivel de aptitud, variando este último en un rango de 600 a 1600 puntos y a cada puntuación se le asigna un rango del 1 al 14 (rango 1 = mayor puntuación y rango 14 = menor puntuación todo sobre los 14 elementos de toda la muestra), obteniéndose los siguientes resultados:

Puntuaciones y Rangos de Hombres y Mujeres en la Prueba de aptitudes mecánicas				
HOMBRES			**MUJERES**	
Puntuación	**Rango**		**Puntuación**	**Rango**
1 500	2		1400	3
1 600	1		1200	6
670	13		780	12
800 *	10.5		1350	4
1 100	8		890	9
800 *	10.5			
1 320	5		TOTAL	34
1 150	7			
600	14			
TOTAL	71			

Nota: * El caso de empate se resolvió asignando el promedio de los rangos que le corresponderían y que serían el rango 10 y rango 11.

Así que se calculan los estadísticos U y U', de la siguiente manera:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum Rangos_1 \quad y \quad U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum Rangos_2$$

que en este caso es igual a:

$$U = (9)(5) + \frac{9(10)}{2} - 71 = 19 \quad y \quad U' = (9)(5) + \frac{5(6)}{2} - 34 = 26$$

La regla de decisión es rechazar Ho si el estadístico U es menor al cuantil $\alpha/2$, donde:

Ho: $\mu_1 = \mu_2$
Vs.
Ha: $\mu_1 \neq \mu_2$ Con un nivel de significancia de $\alpha = 0.05$

Buscando en tablas de Mann-Whitney[1], se tiene que el cuantil es 7, por lo que como U > 7 entonces se acepta la hipótesis nula de que no hay diferencia en las aptitudes.

[1] Libro: Estadística para Administración y Economía, Levine

MUESTRAS GRANDES

En el caso de que una de las dos muestras exceda las 20 observaciones, se aplica la prueba Z, en donde el estadístico de prueba es:

$$Z = \frac{\sum R_1 - \sum R_2 - \left[(n_1 - n_2) \frac{n_1 + n_2 + 1}{2} \right]}{\sqrt{n_1 n_2 \left[\frac{n_1 + n_2 + 1}{3} \right]}}$$

Ejemplo.

Suponga que en el problema de aptitud mecánica se tiene una muestra de 20 mujeres y 15 hombres, y que las calificaciones de ambos grupos, por rango, son:

Rangos para los hombres				**Rango para mujeres**				
26	6	38		7	33	39	16	19
10	14	24		20	1	2	37	29
30	17	40		27	9	5	23	31
3	22	25		28	13	36	15	18
32	34	12		4	21	8	11	35
Total=	ΣRangos$_1$	= 487		Total=	ΣRangos$_2$	= 333		

Se desea probar:
Ho: $\mu_1 = \mu_2$
Vs.
Ha: $\mu_1 \neq \mu_2$ Con un nivel de significancia de $\alpha = 0.05$.

En este caso $\Sigma R_1 = 487$ y $\Sigma R_2 = 333$, $n_1=25$, $n_2=15$, por lo que el estadístico Z es igual a: Z = -0.71.

Buscando en tablas de normal, el cuantil que deja $\alpha/2 = 0.025$ a la izquierda y $\alpha/2 = 0.025$ a la derecha es: -1.96 y 1.96 y como el estadístico cae entre estos dos valores, por lo tanto se acepta la hipótesis nula de que no existe diferencia entre las aptitudes de los hombres con el de las mujeres a un nivel de significancia $\alpha = 0.05$.

V.5. PRUEBA DE VARIAS MEDIAS DE KRUSKAL-WALLIS.

En este caso se tarta de realizar la prueba:

Ho: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$
Vs.
Ha: al menos $\mu_i \neq \mu_j$

En este caso, el procedimiento de prueba es combinar todas las muestras de las poblaciones, se ordenan de menor a mayor y a cada valor se le asigna un rango comenzando con 1 al de puntuación más baja.

Esta prueba se utiliza porque nos se puede suponer normalidad en las puntuaciones y además las varianzas poblacionales no son las mismas.

Ejemplo.

Se va a realizar un estudio en las diferentes unidades operativas de una empresa para saber si los departamentos de contabilidad, mercadotecnia y finanzas para conocer las habilidades en el manejo de la computadora, ya que se tiene planeado implementar un programa de capitación en cómputo y se desea saber si es necesario implementarlo para diferentes niveles para cada área.

Para ello se toma una muestra de 7 empleados de contabilidad, 8 de mercadotecnia y 6 de finanzas, obteniéndose los siguientes puntajes:

Puntuaciones y rangos en una prueba de aptitudes computacionales							
Dpto. Contabilidad			**Depto. Mercadotecnia**			**Depto. Finanzas**	
Puntaje	**Rango**		**Puntaje**	**Rango**		**Puntaje**	**Rango**
51	9		14	1		89	19
32	8		31	7		20	3.5
17	2		58	13		60	11
69	14		87	18		72	15
86	17		20	3.5		56	10
62	12		28	6		22	5
96	20		77	16			
			97	21			
Totales =	82			85.5			63.5

En este caso el estadístico de prueba es:

$$H = \frac{12}{N(N+1)}\left[\frac{\left(\sum R_1\right)^2}{n_1} + \frac{\left(\sum R_2\right)^2}{n_2} + \ldots + \frac{\left(\sum R_k\right)^2}{n_k}\right] - 3(n+1)$$

y

$$N = n_1 + n_2 + \ldots + n_k$$

El cual esta distribuido aproximadamente como una Ji-cuadrada con k-1 (k = # de poblaciones).

Como ΣR_1=82, ΣR_2=85.5 y ΣR_3=63.5, entonces el estadístico H para la muestra es igual a H = 0.1401, y como el cuantil de una Ji-cuadrada con dos grados de libertad que deja 0.05 a su derecha es 5.991, el estadístico H es entonces menor que el cuantil y por lo

tanto no se rechaza Ho a un nivel 0.05 de significancia. Así que los tres departamentos tienen el mismo nivel de aptitudes en el manejo de la computadora.

V.6. PRUEBA WILCOXON PARA PROBAR DIFERENCIAS POR PARES.

La prueba de hipótesis de comparar una misma población antes y después de un tratamiento, se puede hacer mediante una distribución t-student, pero en caso de que no se pueda suponer normalidad o bien cuando los datos se encuentren en escala ordinal, entonces se debe aplicar la prueba no paramétrica Wilcoxon.

Ejemplo.

Suponga que se desea saber si el programa de capacitación en cómputo mejoró las habilidades de los empleados en dicha materia. Para los cual se observa el nivel de habilidades antes del programa y después del programa en una muestra de 22 empleados y se les calificó sus habilidades antes y después del programa, obteniéndose los siguientes resultados:

Número Empleado	Puntaje Antes (a)	Puntaje Después (b)	Diferencia b-a	Diferencias Absolutas Ordenadas	Rango	Rangos con signos correctos
1	18	15	-3	2	1	1
2	60	70	10	3	2	-2
3	81	75	-6	4	3	-3
4	15	20	5	5	4	4.5
5	20	50	30	5	5	4.5
6	17	40	23	6	6	-6
7	26	50	24	8	7	-7.5
8	11	30	19	8	8	7.5
9	20	40	20	9	9	-9
10	38	30	-8	10	10	10.5
11	80	85	5	10	11	10.5
12	59	86	27	11	12	12
13	12	72	60	19	13	13
14	87	98	11	20	14	15
15	88	79	-9	20	15	15
16	64	88	24	20	16	15
17	88	90	2	23	17	17
18	76	96	20	24	18	17.5
19	43	39	-4	24	19	17.5
20	90	98	8	27	20	20
21	40	60	20	30	21	21
22	50	60	10	60	22	22

Suma de Rangos positivos =	-27.5
Suma de Rangos negativos =	223.5

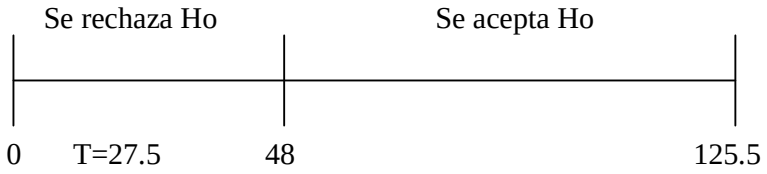
La hipótesis a probar es:

Ho: No hay diferencia significativa debido al tratamiento
Vs.

Ha: Hay diferencia significativa por el tratamiento

La columna de rangos con signos correctos, se determinó mediante el promedio de rangos, si la diferencia absoluta se repite, y los rangos con signos correctos preserva el signo de la diferencia que le dio origen. Por ejemplo para el rango 4 y 5 se promedio (4+5)/2 = 4.5 y como el rango 4 corresponde a una diferencia 5 positiva entonces se le asigna 4.5 positivo, lo mismo para el rango 5. En el caso de los rangos 7 y 8 (correspondientes a una diferencia de 8), el promedio es 7.5, y como la diferencia de 8 corresponde a un valor negativo y otro positivo, entonces se le asigna un rango con signo correcto de -7.5 y 7.5.

El estadístico de prueba en este caso es T=27.5, y el cuantil que deja un área de 0.01 para 22 grados de libertad para una prueba de dos colas es igual a 48. Además si se aceptará Ho entonces ambas categorías (antes y después) deberán tener una suma de rangos igual a (27.5+223.5)/2 = 125.5.



Por lo que en este caso se rechaza Ho y por lo tanto podemos concluir que, a un nivel de significancia de 0.01, el programa de capacitación en cómputo mejoró las habilidades del personal.

CAPITULO VI ANÁLISIS DE REGRESIÓN SIMPLE Y CORRELACIÓN

VI.1. INTRODUCCIÓN

El **análisis de regresión lineal o bivariada**[1] es un procedimiento estadístico que sirve para estudiar la relación entre dos variables cuando una se considera como variable dependiente y la otra como variable independiente. Por ejemplo, podría ser de interés analizar la relación entre las ventas (variable dependiente) y la publicidad (variable independiente). Si el investigador estima la relación entre los gastos publicitarios y las ventas mediante el análisis de regresión, podrá predecir las ventas para diferentes niveles publicitarios[2]. **Cuando se emplean dos o más variables independientes en el problema (tales como la publicidad y el precio del producto) para pronosticar la variable dependiente de interés, se aplica el análisis de regresión múltiple.**

Una de las decisiones de mercadotecnia más difíciles que enfrenta cualquier empresa es: cuánto destinar a las promociones. **John Wanamaker**, el magnate de las tiendas departamentales, dijo en cierta ocasión: "se que la mitad de mi publicidad se desperdicia, pero no sé cuál mitad. Dedique 2 millones de dólares a publicidad pero no sé si eso es la mitad de lo que debería gastar o el doble de lo necesario". En general, las empresas gastan cantidades muy diferentes en las promociones, por ejemplo, en la industria de los cosméticos es común que se gaste entre el 20 y el 30% de las ventas en promoción y sólo del 5 al 10% en el caso de la maquinaria industrial. Sin embargo existen cuatro métodos usados con frecuencia para establecer el presupuesto para la publicidad:

1. método de lo factible;
2. método del porcentaje de ventas;
3. método de la paridad competitiva y
4. el método de objetivo y tarea[1].

Muchas empresas aplican **el método de lo factible**, es decir, establecen el presupuesto para promociones en un nivel al cual la empresa puede tener acceso. Un ejecutivo explica este método así: "es muy simple. Primero subo a la oficina del contralor y le pregunto cuánto nos puede proporcionar este año. Contesta que un millón y medio. A continuación, el jefe me pregunta cuánto deberíamos gastar y yo respondo: bueno, alrededor de millón y medio. Algunas otras empresas aplican **el método del porcentaje de ventas**, es decir, establecen su presupuesto para promociones de acuerdo con cierto porcentaje de las ventas, presentes o pronosticadas. Por ejemplo: las empresas automovilísticas sueles presupuestar un porcentaje fijo para su promoción, con base en el precio proyectado para el auto. Otras empresas aplican **el método de la paridad competitiva** y establecen su presupuesto para promociones a semejanza de las partidas de la competencia. Observan la publicidad de la competencia o consiguen estimaciones del gasto para promociones de la industria, de publicaciones o asociaciones del gremio, y después establecen sus presupuestos con base en el promedio de la industria. El método más lógico para establecer presupuesto es **el método de objetivos y tarea,** con el cual la empresa establece su presupuesto para promociones con base en lo que quiere lograr con sus promociones. Los mercadólogos preparan sus presupuestos para promociones:

- definiendo los objetivos específicos;
- determinando las tareas que se deben realizar para alcanzar estos objetivos y
- estimando los costos para realizar estas tareas.

La suma de estos costos se convierte en el presupuesto de promoción que se propone. Este método de objetivo y tarea obliga a la gerencia a detallar sus hipótesis en cuanto a la relación entre el dinero gastado y los resultados de las promociones. Sin embargo, también es el método más difícil de usar.

(Mercadotecnia. Philip Kotler & Gary Armstrong. Prentice Hall 6a. Edición, 1996. p.p 563-565)

---

[1] McDaniel, Carl & Gates, Roger. Investigación de mercados contemporánea. Cuarta edición. International Thomson editores. P.p 558

[2] La publicidad es una forma de comunicación masiva, unilateral, impersonal y se puede difundir por muchos medios distintos como: la televisión, la radio, periódicos, diarios, revistas, libros, correo directo, correo electrónico, carteleras, etc. la publicidad tiene dos áreas principales de decisión:

1. determinar el mensaje que se va a transmitir al mercado objetivo y
2. la selección de los medios[2].

Naturaleza de la relación[3]:

Para estudiar la naturaleza de la relación entre la variable dependiente y la independiente, se construye un diagrama de dispersión. La variable dependiente "**y**" se grafica en el eje vertical y la variable independiente "**x**" en el eje horizontal. Al examinar el diagrama de dispersión, se ve si la relación entre las dos variables, en caso de que existe, es lineal o curva. Si la relación parece lineal o está cerca de ella, puede aplicarse la regresión lineal. Cuando se observa una relación no lineal en el diagrama de dispersión, se emplean técnicas de regresión no lineal para la adaptación a una curva, en cuyo caso se utilizan técnicas que se encuentran más allá del alcance de este análisis.

Cuando en el congreso de los Estados Unidos[4] se debatía sobre la ley de educación superior en 1992, sus partidarios observaron que los ingresos de trabajadores egresados de universidades, ajustados a la inflación, se habían elevado durante la década de los 80´s en relación a los de trabajadores que tenían sólo preparatoria o menos. Los economistas tienen una explicación fácil para este hecho: **en años recientes, la demanda de trabajadores capacitados se ha elevado más que la de los no capacitados**, y la oferta no se ha modificado de manera correspondiente.

El único factor más importante en el lado de la demanda ha sido el cambio tecnológico con sesgo en la capacitación, la rápida automatización del trabajo es un ejemplo. La introducción del correo electrónico, por citar un caso, reduce la demanda de recepcionistas no calificadas que contestan teléfonos y llenan cajas de mensajes, pero al mismo tiempo crea demanda para especialistas calificados en computación que instalan y dan servicio al sistema. Cambios similares ocurren en las líneas de producción, en donde los robots accionados por computadoras toman el lugar de los anticuados trabajadores de líneas de ensamble; se necesitan menos trabajadores no calificados pero más trabajadores calificados.

Cabe mencionar aquí que la demanda de trabajadores calificados no es proporcional al despido de trabajadores no capacitados, y Bill Gates pronostica en su libro: Camino al futuro, que precisamente en el futuro, sólo habrá trabajo para el 20% de la población mundial y que el otro 80% en consecuencia vivirá a expensas de ese 20%

Raíces históricas y conceptos básicos:

Cualquier método estadístico que busque establecer una ecuación que permita estimar el valor desconocido de una variable, a partir del valor conocido de una o más variables, se denomina **análisis de regresión**.

---

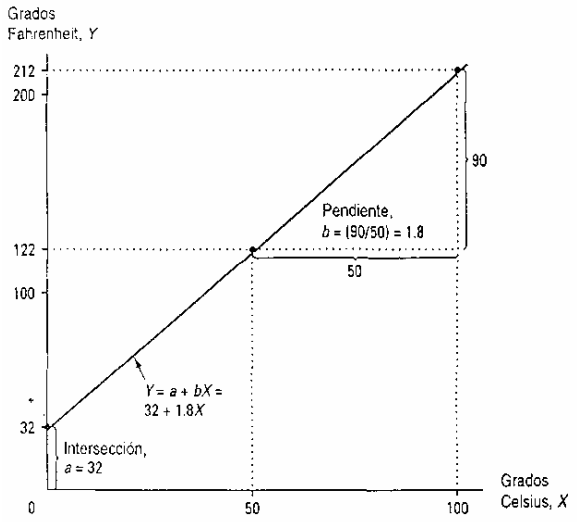
[3] McDaniel, Carl & Gates, Roger. Investigación de mercados contemporánea. Cuarta edición. International Thomson editores. P.p
[4] Kohler, Heinz. Estadística para negocios y economía. Editorial CECSA. Primera reimpresión, 1998. p.p 524

El origen de la regresión simple está cercanamente unido al genetista y estadístico inglés Francis Galton (1822-1911) quien experimentó con chicharos para determinar la ley de la herencia en el tamaño.

En el análisis de regresión[5], una variable cuyo valor se suponga conocido y que se utilice para explicar o predecir el valor de otra variable de interés se llama **variable independiente** y se simboliza por: **X**.

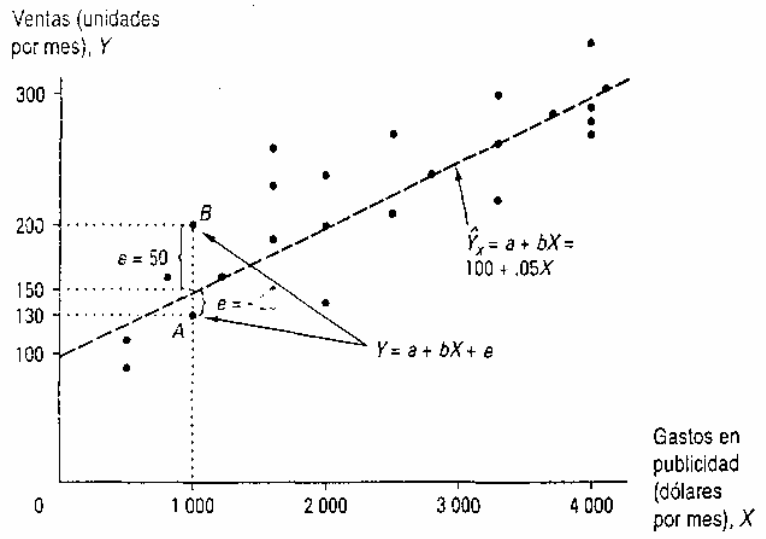
En el análisis de regresión, una variable cuyo valor se suponga desconocido y que se explique o prediga con ayuda de otra se llama **variable dependiente**; se simboliza por **Y**.

Una **relación determinística**[6] entre dos variables cualesquiera, ***x*** y ***y***, se caracteriza por el hecho de que el valor de y está determinado de manera única siempre que el valor de ***x*** se especifique.



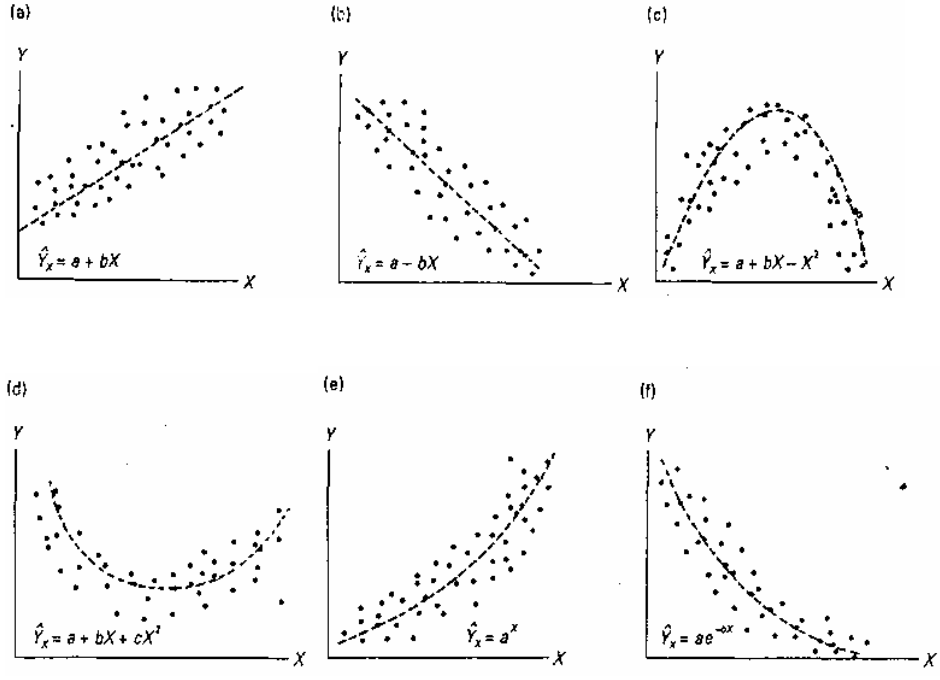
Una **relación estocástica**[7] entre dos variables cualesquiera, ***x*** y ***y*** es imprecisa en el sentido de que muchos valores posibles de "**y**" se pueden asociar con cualquier valor de "***x***". sin embargo, un resumen gráfico de la relación estocástica entre la variable independiente "***x***" y la variable dependiente "***y***" estará dado por una línea de regresión, misma que reduce al mínimo los errores cometido cuando la ecuación de esa línea se utilice para estimar **y** a partir de ***x***.

[5] Kohler, Heinz. Estadística para negocios y economía. Editorial CECSA. Primera reimpresión, 1998. p.p 528-529
[6] Kohler, Heinz. Estadística para negocios y economía. Editorial CECSA. Primera reimpresión, 1998. p.p 530
[7] Kohler, Heinz. Estadística para negocios y economía. Editorial CECSA. Primera reimpresión, 1998. p.p 530

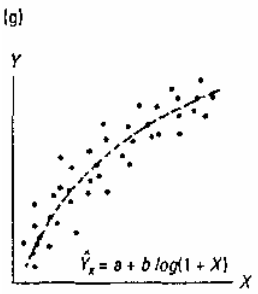


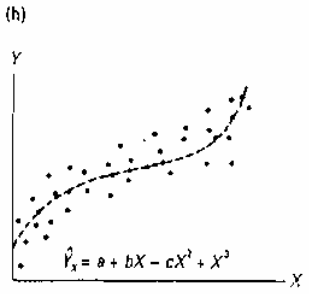
Relaciones alternativas[8] entre *x* y *y*.

Cada punto en las gráficas siguientes representa un par hipotético de observaciones alrededor de una variable independiente, ***x***, y una variable dependiente, ***y***. las líneas discontinuas resumen la naturaleza de su relación:



[8] Kohler, Heinz. Estadística para negocios y economía. Editorial CECSA. Primera reimpresión, 1998. p.p 531





Método de los Mínimos cuadrados[9]:

Antes de iniciar con el ejemplo, es necesario advertir que **el análisis de regresión no se puede interpretar como un procedimiento para establecer una relación de causa a efecto entre variables**. Sólo puede indicar cómo, o hasta qué grado las variables están **asociadas** entre sí. Cualquier conclusión acerca de causa y efecto se debe basar en el **juicio** del o los individuos con más conocimientos sobre la aplicación.

Este método, al que también se le llama con frecuencia: de los mínimos cuadrados, es un procedimiento para encontrar la ecuación de regresión.

Karl Friedrich Gauss (1777-1855) propuso el método de los cuadrados mínimos. Fue el primero en demostrar que la ecuación estimada de regresión minimiza la suma de cuadrados de errores.

Anderson, Sweeney & Williams, 1999. Estadística para administración y economía, international thomson editores, México, p.p 549

Para ilustrarlo consideremos el siguiente **Ejemplo:**

Domino's Pizza es una cadena de restaurantes dedicados exclusivamente a la distribución de Pizzas. Los lugares donde sus establecimientos han tenido más éxito están cercanos a establecimientos de educación superior. Los administradores creen que las ventas trimestrales en esos restaurantes (representadas por "**y**"), se relacionan en forma positiva con la población estudiantil (representada por "**x**"). Esto es, que los restaurantes cercanos a centros escolares con gran población tienden a generar más ventas que los que están cerca de centros con población pequeña. Aplicando el análisis de regresión podremos plantear una ecuación que muestre cómo se relaciona la variable dependiente "**y**" con la variable independiente "**x**".

En este ejemplo, cada restaurante está asociado con un valor de "**x**" (población estudiantil) y un valor correspondiente de "**y**" (ventas trimestrales). La ecuación

[9] Anderson, Sweeney & Williams, 1999. Estadística para administración y economía, international thomson editores, México, p.p`547

que describe cómo se relaciona "**y**" con "**x**" se llama: **ecuación de regresión lineal simple** y tiene la forma:

$$\hat{Y} = b_0 + b_1 x$$

En la regresión lineal simple, la gráfica de la ecuación de regresión se llama: línea de regresión estimada; b_0 es la ordenada al origen, b_1 es la pendiente y $\hat{y}$ es el valor estimado de "**y**" para determinado valor de "**x**".

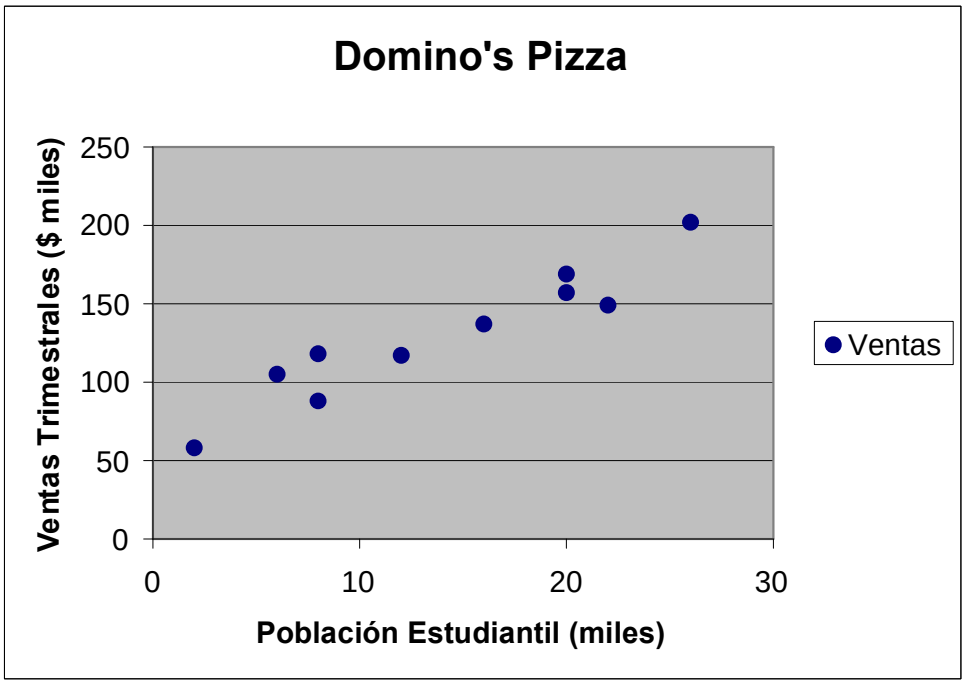
> En la regresión lineal simple, el análisis de datos de dos variables (datos bivariados) implica medir dos variables para cada elemento de una muestra.
>
> Anderson, Sweeney & Williams, 1999. Estadística para administración y economía, international thomson editores, México, p.p 545

Para ilustrarlo, supongamos que en el caso de Domino's Pizza se reunieron datos de una muestra de 10 restaurantes ubicados cerca de centros educativos. Para el i-ésimo restaurante de la muestra, x_i es el tamaño de la población estudiantil, en miles, y y_i son las ventas trimestrales (en miles de pesos). Los valores de x_i y y_i para los 10 restaurantes de la muestra se resume en la siguiente tabla:

Datos de población estudiantil y ventas trimestrales para 10 restaurantes de Domino's Pizza.

Restaurante i	Población de estudiantes x_i (miles)	Ventas trimestrales y_i ($ miles)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

La siguiente gráfica corresponde al diagrama de dispersión de los datos de la tabla anterior:


Domino's Pizza
Ventas Trimestrales ($ miles)
250
200
150
100
50
0
0
10
20
30
Población Estudiantil (miles)
Ventas

¿Cuáles son las conclusiones que podemos sacar de la gráfica anterior?

- ✓ Parece que las ventas son mayores en los centros con más población de estudiantes.
- ✓ Para esos datos, la relación entre el tamaño de la población de estudiantes y las ventas parece poderse aproximar con una línea recta.
- ✓ Parece haber una relación lineal positiva entre "**x**" y "**y**".

En consecuencia, elegimos el modelo de regresión lineal simple para esta opción, nuestra siguiente tarea será emplear los datos de la muestra de la tabla para determinar los valores de b_0 y b_1 en la ecuación de regresión lineal simple. Por lo tanto, para el i.ésimo restaurante, la ecuación de regresión es:

$$\hat{Y}_i = b_0 + b_1 X_i$$

En la que:

x_i = tamaño de la población estudiantil (miles) para el i-ésimo restaurante,
b_0 = ordenada al origen de la línea estimada de regresión,
b_1 = pendiente de la línea estimada de regresión,
$\hat{Y}_i$ = valor estimado de las ventas trimestrales, en miles, para el i-ésimo restaurante.

En estas condiciones, los que se pretende es que los errores de la regresión sean los mínimos posibles, donde los errores son:

$$\sum_{i=1}^{10} e_i = \Sigma\,(\hat{Y}_i - Y_i)$$

donde $\hat{Y}_i$=El valor estimado de ventas
Y_i=El valor observado de ventas

Pero para minimizar los errores se requiere considerar el valor absoluto $|\hat{Y}_i - Y_i|$, pues puede en caso contrario puede suceder que los errores sean muy grandes (positivos y negativos) pero al sumar todos se reduzcan o eliminen, o bien considerar los cuadrados $(\hat{Y}_i - Y_i)^2$. Se utilizará esto último.

Entonces necesitamos minimizar la suma:

$$\text{Minimizar } \Sigma\,(\hat{Y}_i - Y_i)^2 = \text{Min } \Sigma(\, b_0 + b_1 X_i - Y_i)^2$$

Problema que es resuelto por el cálculo para varias variables (b_0 y $b1$), y que en nuestro caso no detallaremos, pero cuyo resultado permite estimar los términos b_0 y b_1, a saber:

$$b_0 = \frac{SS_{xy}}{SS_x}$$

$$b_1 = \overline{Y} - b_1 \overline{X}$$

Donde:

$$SS_{xy} = \sum_{i=1}^{n} XY - \frac{(\sum X)(\sum Y)}{n}$$

y

$$SS_x = \sum_{i=1}^{n} X^2 - \frac{(\sum X)^2}{n}$$

En nuestro caso es al realizar los cálculos necesarios tenemos que:

Restaurante	Población de estudiantes	Ventas trimestrales			
I	x_i (miles)	y_i ($ miles)	X^2	Y^2	XY
1	2	58	4	3364	116
2	6	105	36	11025	630
3	8	88	64	7744	704
4	8	118	64	13924	944
5	12	117	144	13689	1404
6	16	137	256	18769	2192
7	20	157	400	24649	3140
8	20	169	400	28561	3380
9	22	149	484	22201	3278
10	26	202	676	40804	5252
TOTALES	140	1300	2528	184730	21040

SSxy =	2840
SSx =	568

b1 =	**5**
bo =	**60**

Y por ejemplo si se planea construir un nuevo centro en el cual la población estudiantil es de aproximadamente 30 mil, entonces el nivel de ventas estimado sería igual a 60 + 5(30) = 210 mil trimestrales. Se puede también hacer una estimación por intervalos de confianza pero antes es necesario verificar la validez del modelo como sigue:

Coeficiente de determinación r^2.

Se utiliza para evaluar la bondad de ajuste para la ecuación de regresión y se define como:

$$r^2 = \frac{\text{Suma de Cuadrados de la regresión}}{\text{Suma de cuadrados Totales}} = \frac{SSR}{SST}$$

$$= \frac{\Sigma\,(\hat{Y}_i - \bar{Y})^2}{\Sigma\,(Y_i - \bar{Y})^2}$$

El cual se puede interpretar como el porcentaje de variación de la variable Y que se puede explicar con el modelo de regresión, en nuestro ejemplo, r^2=0.9027 (cálculo hecho con excel con la función coeficiente.R2(rango de x;rango de y)), así que el 90.27% de la variación de las ventas se puede explicar por el modelo de regresión, lo cual hace que sea un buen modelo.

Prueba de hipótesis F de significancia del modelo.

La hipótesis a probar en este caso es:

Ho: $\beta = 0$
Vs.
Ha: $\beta \neq 0$

Y en caso de rechazar Ho, nos indicará que si hay una relación de dependencia entre la variable Y y la variable X, pero la desventaja es que no nos dice si esta dependencia es lineal (puede ser necesario ajustar otros modelos como el cuadrático, polinomial, logarítmico o exponencial).

El estadístico de prueba es:

$$F = \frac{MSR}{MSE} = \frac{SSR\,/\#\,\text{variables independientes}}{SSE\,/(n - 1 - \#\,\text{variables independientes})}$$

Donde SSE = $\Sigma\,(Y_i - \hat{Y}_i)^2$, observándose que SST = SSR + SSE, es decir la suma de cuadrados totales son iguales a los de la regresión más la de los errores del modelo, entre menor sea SSE es mejor el modelo.

La regla de decisión para la prueba de hipótesis es, rechazar Ho si el estadístico F es mayor que el cuantil que deja α a la derecha de una distribución F-Fisher.

Para el ejemplo, el cálculo se puede resumir en la siguiente tabla:

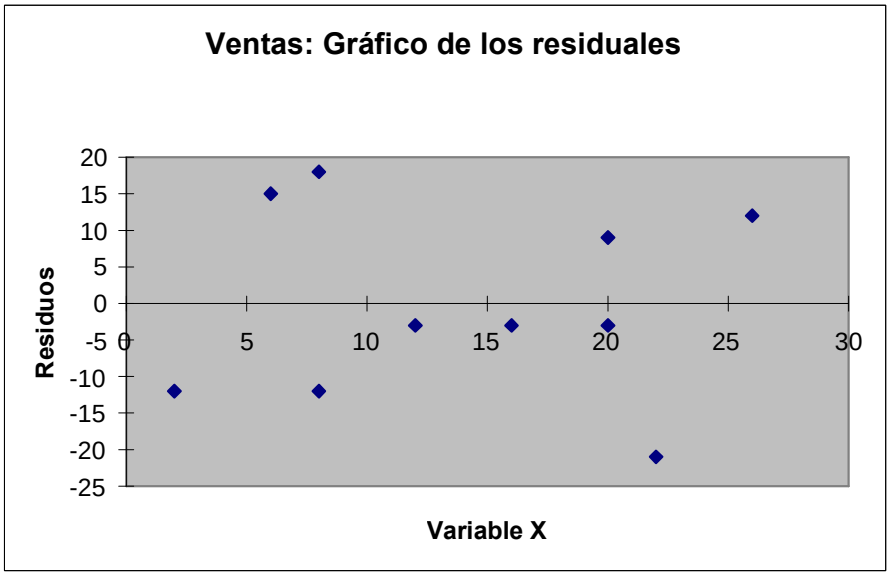
ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	#var.indep =1	SSR = 14200	MSR = 14200	MSR/MSE=74.248366	2.5489E-05
Residuos	8	SSE = 1530	MSE = 191.25		
Total	n -1 =9	SST = 15730			

Y como el valor crítico de la F es menor a α = 0.05, rechazamos Ho y se puede decir que si hay una dependencia significativa entre X y Y.

Para saber si la dependencia es lineal es necesario graficar los residuales y ver si hay un comportamiento cuadrático o de otro tipo.

Es necesario mencionar que tanto la tabla de análisis de varianza como la gráfica de residuales se obtuvieron mediante excel, en el menú herramientas, análisis de datos, regresión.



Como los residuales son una nube de puntos alrededor del origen y sin patrón alguno, entonces podemos decir que nuestro modelo es bueno.

Cabe mencionar que en dado caso que se quisieran hacer proyecciones, estas deben hacerse con precaución, pues el modelo sólo es valido para un rango de X entre 2,000 y 26,000. Pueden hacerse proyecciones dentro de este rango o en un intervalo fuera de este pero no muy distante.

Pronóstico para $E(Y|X_0)$.

En el caso de tratar de estimar por ejemplo el nivel de ventas promedio para una población de estudiantes de 25 000, es decir las ventas para una población de 25,000 pueden ser diferentes en varios casos, pero estamos interesados en el promedio de ventas.

El intervalo de confianza al (1-α)% en este caso es:

$$\hat{Y} \pm t_{\alpha/2} \sqrt{\sum (Y_i - \hat{Y}_i)^2 / (n-2)} \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{SSx}} = \hat{Y} \pm t_{\alpha/2} \sqrt{SSE/(n-2)} \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{SSx}}$$

donde los grados de libertad de la t es n-2.

Que para el caso del ejemplo de Pizzas es igual a:

$\overline{X}$ = 14

$\hat{Y}$ = 185

$t_{0.05}$ = 1.86

$\sqrt{SSE/(n-2)}$ = 0.5594892

SS_x = 568

Por lo cual el intervalo es:
(170.6085,199.3914)

Pronóstico para Y en un valor particular X_0.

En este caso no es un estimado del promedio, es más bien un valor particular de Y para un valor particular X_0, y dependerá del tipo de estimación requerida para saber cual estimación considerar, pues si se quiere el promedio de las ventas para una población estudiantil por ejemplo de 25,000 o bien una estimación para una tienda en particular en un nivel de 25,000 en la población.

La fórmula para el intervalo al (1-α)% en este caso es:

$$\hat{Y} \pm t_{\alpha/2} \sqrt{\sum (Y_i - \hat{Y}_i)^2 / (n-2)} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{SSx}}$$

También la t es de n-2 grados de libertad.

Y en este caso el intervalo al 90% de confianza es:

(155.5252, 214.4747)

que es de mayor amplitud comparado con el del promedio de Y (en general siempre sucederá esto).

Ajuste de un modelo no-lineal.

Suponga que de acuerdo al análisis de residuales se requiere que ajustemos un modelo exponencial de la forma $Y = ab^x$, entonces se puede hacer un ajuste lineal transformando los datos, al considerar logaritmos como sigue:

$Log(Y) = Log(ab^x)$ ➔ $Log(Y) = Log(a) + XLog(b)$ que es de la forma $Y' = a' + b'X$

Es decir se tiene que realizar una regresión lineal con la variable dependiente Log(Y) y con la variable independiente X, y el resultado será **a'=Log(a)**, es decir $\mathbf{a=10^{a'}}$ y **b'=Log(b)**, que equivale a $b=10^{b'}$.

Si el modelo fuera Y = a + Log(1+X) entonces tenemos que realizar un regresión lineal entre Y y transformando los valores de X por el de Log(1+X).

Por último, si la regresión requiere que se ajuste un modelo $Y= a + bX+cX^2+dX^3$, es necesario realizar una regresión lineal múltiple como se verá en la siguiente sección.

Aplicación Práctica.

Considere los rendimientos históricos entre el precio de la acción de Kimberly serie A y el índice de precios y cotizaciones.

El rendimiento del mercado se determina mediante $(IPyC_t - IPyC_{t-1})/ IPyC_{t-1}$, por ejemplo para el 3/01/2002, se tiene que le rendimiento es (6603.75-6410.05)/6410.05. Lo mismo para el precio de las acciones de Kimberly, que para el día 4/01/2002, se calcula el rendimiento mediante (27.15-26.99)/26.99.

Comportamiento de los precios de Kimber A

Número	Fecha	Precio de Cierre	Rendimiento diario de la acción	Precio de Cierre	Rendimiento diario del mercado
				IPyC	
0	02/01/2002	26.99		6410.05	
1	03/01/2002	26.99	0.00	6603.75	3.02
2	04/01/2002	27.15	0.59	6612.08	0.13
3	07/01/2002	26.65	-1.84	6565.44	-0.71
4	08/01/2002	25.79	-3.23	6641.14	1.15
5	09/01/2002	25.40	-1.51	6560.58	-1.21
6	10/01/2002	25.20	-0.79	6453.01	-1.64
7	11/01/2002	25.48	1.11	6420.15	-0.51
8	14/01/2002	25.40	-0.31	6388.27	-0.50
9	15/01/2002	25.62	0.87	6573.19	2.89
10	16/01/2002	25.57	-0.20	6579.35	0.09
11	17/01/2002	25.53	-0.16	6604.66	0.38
12	18/01/2002	25.96	1.68	6600.73	-0.06
13	21/01/2002	25.96	0.00	6607.80	0.11
14	22/01/2002	25.70	-1.00	6590.04	-0.27
15	23/01/2002	25.98	1.09	6782.78	2.92
16	24/01/2002	25.38	-2.31	6768.30	-0.21
17	25/01/2002	25.06	-1.26	6831.43	0.93
18	28/01/2002	24.95	-0.44	6872.44	0.60
19	29/01/2002	24.86	-0.36	6793.38	-1.15
20	30/01/2002	24.85	-0.04	6750.80	-0.63
21	31/01/2002	25.01	0.64	6927.87	2.62
22	01/02/2002	25.52	2.04	6901.81	-0.38
23	04/02/2002	25.68	0.63	6865.13	-0.53
24	05/02/2002	25.68	0.00	6865.13	0.00
25	06/02/2002	26.20	2.02	6786.74	-1.14
26	07/02/2002	25.97	-0.88	6780.94	-0.09
27	08/02/2002	25.82	-0.58	6681.45	-1.47
28	11/02/2002	25.40	-1.63	6589.84	-1.37
29	12/02/2002	25.68	1.10	6673.36	1.27
30	13/02/2002	25.66	-0.08	6715.55	0.63
31	14/02/2002	25.88	0.86	6717.61	0.03

32	15/02/2002	25.95	0.27	6697.67	-0.30
33	18/02/2002	25.95	0.00	6679.71	-0.27
34	19/02/2002	25.95	0.00	6609.30	-1.05
35	20/02/2002	26.00	0.19	6583.99	-0.38
36	21/02/2002	26.00	0.00	6533.64	-0.76
37	22/02/2002	26.00	0.00	6473.12	-0.93
38	25/02/2002	26.23	0.88	6562.94	1.39
39	26/02/2002	26.39	0.61	6637.96	1.14
40	27/02/2002	27.54	4.36	6795.90	2.38
41	28/02/2002	28.42	3.20	6734.44	-0.90
42	01/03/2002	28.70	0.99	6898.00	2.43
43	04/03/2002	29.66	3.34	7029.60	1.91
44	05/03/2002	29.81	0.51	7053.54	0.34
45	06/03/2002	29.97	0.54	7154.14	1.43
46	07/03/2002	29.60	-1.23	7061.00	-1.30
47	08/03/2002	29.93	1.11	7192.22	1.86
48	11/03/2002	29.87	-0.20	7161.40	-0.43
49	12/03/2002	29.88	0.03	7278.06	1.63
50	13/03/2002	30.06	0.60	7218.51	-0.82
51	14/03/2002	29.94	-0.40	7188.77	-0.41
52	15/03/2002	29.99	0.17	7273.08	1.17
53	18/03/2002	31.00	3.37	7344.57	0.98
54	19/03/2002	32.10	3.55	7427.92	1.13
55	20/03/2002	32.68	1.81	7384.13	-0.59
56	22/03/2002	32.19	-1.50	7439.50	0.75
57	25/03/2002	31.39	-2.49	7381.57	-0.78
58	26/03/2002	30.12	-4.05	7351.19	-0.41
59	27/03/2002	30.29	0.56	7361.86	0.15
60	28/03/2002	30.29	0.00	7361.86	0.00
61	29/03/2002	30.29	0.00	7361.86	0.00
62	01/04/2002	30.66	1.22	7371.89	0.14
63	02/04/2002	30.79	0.42	7316.69	-0.75
64	03/04/2002	30.38	-1.33	7191.94	-1.71
65	04/04/2002	30.19	-0.63	7253.50	0.86
66	05/04/2002	30.61	1.39	7335.76	1.13
67	08/04/2002	30.85	0.78	7271.77	-0.87
68	09/04/2002	30.70	-0.49	7271.22	-0.01
69	10/04/2002	31.14	1.43	7517.68	3.39
70	11/04/2002	31.26	0.39	7441.52	-1.01
71	12/04/2002	31.82	1.79	7391.25	-0.68
72	15/04/2002	31.50	-1.01	7396.72	0.07
73	16/04/2002	31.78	0.89	7535.07	1.87
74	17/04/2002	31.36	-1.32	7574.35	0.52
75	18/04/2002	30.80	-1.79	7532.14	-0.56
76	19/04/2002	30.00	-2.60	7509.22	-0.30
77	22/04/2002	29.10	-3.00	7509.22	0.00
78	23/04/2002	29.68	1.99	7403.38	-1.41
79	24/04/2002	29.50	-0.61	7504.73	1.37
80	25/04/2002	29.30	-0.68	7476.37	-0.38
81	26/04/2002	29.57	0.92	7491.78	0.21
82	29/04/2002	29.98	1.39	7434.19	-0.77
83	30/04/2002	30.74	2.54	7480.74	0.63
84	02/05/2002	30.90	0.52	7486.86	0.08
85	03/05/2002	31.70	2.59	7521.95	0.47

86	06/05/2002	31.75	0.16	7542.48	0.27
87	07/05/2002	31.71	-0.13	7431.89	-1.47
88	08/05/2002	31.16	-1.73	7517.84	1.16
89	09/05/2002	30.70	-1.48	7351.08	-2.22
90	10/05/2002	30.57	-0.42	7303.57	-0.65
91	13/05/2002	30.57	0.00	7307.16	0.05
92	14/05/2002	30.35	-0.72	7361.94	0.75
93	15/05/2002	30.62	0.89	7402.80	0.56
94	16/05/2002	30.57	-0.16	7514.90	1.51
95	17/05/2002	30.29	-0.92	7537.82	0.30
96	20/05/2002	30.45	0.53	7469.68	-0.90
97	21/05/2002	31.12	2.20	7385.92	-1.12
98	22/05/2002	31.08	-0.13	7385.15	-0.01
99	23/05/2002	30.99	-0.29	7398.50	0.18
100	24/05/2002	30.90	-0.29	7366.43	-0.43
101	27/05/2002	29.20	-5.50	7357.24	-0.12
102	28/05/2002	29.03	-0.58	7303.57	-0.73
103	29/05/2002	29.03	0.00	7130.71	-2.37
104	30/05/2002	28.98	-0.17	7061.40	-0.97
105	31/05/2002	28.98	0.00	7031.64	-0.42
106	03/06/2002	28.85	-0.45	6997.05	-0.49
107	04/06/2002	28.50	-1.21	7016.14	0.27
108	05/06/2002	27.47	-3.61	6974.92	-0.59
109	06/06/2002	27.47	0.00	6791.91	-2.62
110	07/06/2002	27.44	-0.11	6857.42	0.96
111	10/06/2002	27.39	-0.18	6835.00	-0.33
112	11/06/2002	27.24	-0.55	6862.89	0.41
113	12/06/2002	27.09	-0.55	6801.65	-0.89
114	13/06/2002	26.60	-1.81	6760.68	-0.60
115	14/06/2002	26.53	-0.26	6720.37	-0.60
116	17/06/2002	26.41	-0.45	6788.94	1.02
117	18/06/2002	26.20	-0.80	6765.19	-0.35
118	19/06/2002	25.81	-1.49	6691.04	-1.10
119	20/06/2002	25.62	-0.74	6580.81	-1.65
120	21/06/2002	25.23	-1.52	6502.97	-1.18
121	24/06/2002	25.54	1.23	6331.32	-2.64
122	25/06/2002	25.25	-1.14	6354.26	0.36
123	26/06/2002	25.70	1.78	6171.63	-2.87
124	27/06/2002	26.67	3.77	6400.89	3.71
125	01/07/2002	27.03	1.35	6363.05	-0.59
126	02/07/2002	28.00	3.59	6306.52	-0.89
127	03/07/2002	28.00	0.00	6326.49	0.32
128	04/07/2002	27.73	-0.96	6352.89	0.42
129	05/07/2002	27.46	-0.97	6462.83	1.73
130	08/07/2002	26.97	-1.78	6488.92	0.40
131	09/07/2002	27.10	0.48	6460.95	-0.43
132	10/07/2002	27.20	0.37	6371.27	-1.39
133	11/07/2002	26.70	-1.84	6390.17	0.30
134	12/07/2002	27.00	1.12	6400.42	0.16
135	15/07/2002	26.50	-1.85	6372.08	-0.44
136	16/07/2002	26.30	-0.75	6316.16	-0.88
137	17/07/2002	26.52	0.84	6403.28	1.38
138	18/07/2002	26.47	-0.19	6433.83	0.48
139	19/07/2002	25.58	-3.36	6336.95	-1.51

140	22/07/2002	24.96	-2.42	6113.83	-3.52
141	23/07/2002	24.39	-2.28	5892.41	-3.62
142	24/07/2002	24.00	-1.60	6010.42	2.00
143	25/07/2002	23.91	-0.38	5922.34	-1.47
144	26/07/2002	23.92	0.04	5900.44	-0.37
145	29/07/2002	23.76	-0.67	6103.88	3.45
146	30/07/2002	23.75	-0.04	6014.68	-1.46
147	31/07/2002	23.53	-0.93	6021.84	0.12
148	01/08/2002	23.01	-2.21	5755.99	-4.41
149	02/08/2002	22.86	-0.65	5644.70	-1.93
150	05/08/2002	22.55	-1.36	5534.47	-1.95
151	06/08/2002	22.94	1.73	5747.44	3.85
152	07/08/2002	23.00	0.26	5855.90	1.89
153	08/08/2002	23.11	0.48	6029.77	2.97
154	09/08/2002	23.49	1.64	5913.21	-1.93
155	12/08/2002	23.81	1.36	5901.83	-0.19
156	13/08/2002	23.71	-0.42	5841.90	-1.02
157	14/08/2002	24.00	1.22	6053.60	3.62
158	15/08/2002	24.01	0.04	6146.11	1.53
159	16/08/2002	24.11	0.42	6190.60	0.72
160	19/08/2002	24.20	0.37	6200.87	0.17
161	20/08/2002	24.45	1.03	6192.91	-0.13
162	21/08/2002	24.30	-0.61	6267.88	1.21
163	22/08/2002	24.40	0.41	6239.49	-0.45
164	23/08/2002	24.00	-1.64	6148.89	-1.45
165	26/08/2002	24.00	0.00	6262.44	1.85
166	27/08/2002	23.89	-0.46	6157.42	-1.68
167	28/08/2002	23.73	-0.67	6115.56	-0.68
168	29/08/2002	23.75	0.08	6181.67	1.08
169	30/08/2002	23.80	0.21	6216.43	0.56
170	02/09/2002	23.99	0.80	6165.93	-0.81
171	03/09/2002	23.80	-0.79	6094.29	-1.16
172	04/09/2002	24.00	0.84	6114.35	0.33
173	05/09/2002	23.80	-0.83	6067.25	-0.77
174	06/09/2002	24.01	0.88	6113.32	0.76
175	09/09/2002	24.00	-0.04	6159.90	0.76
176	10/09/2002	24.45	1.88	6225.16	1.06
177	11/09/2002	24.81	1.47	6260.61	0.57
178	12/09/2002	24.94	0.52	6219.93	-0.65
179	13/09/2002	24.80	-0.56	6190.52	-0.47
180	16/09/2002	24.80	0.00	6190.52	0.00
181	17/09/2002	25.00	0.81	6079.95	-1.79
182	18/09/2002	24.75	-1.00	5960.38	-1.97
183	19/09/2002	23.73	-4.12	5645.00	-5.29
184	20/09/2002	23.89	0.67	5788.78	2.55
185	23/09/2002	23.65	-1.00	5741.73	-0.81
186	24/09/2002	23.55	-0.42	5705.67	-0.63
187	25/09/2002	22.80	-3.18	5808.44	1.80
188	26/09/2002	22.93	0.57	5956.93	2.56
189	27/09/2002	22.70	-1.00	5801.12	-2.62
190	30/09/2002	22.83	0.57	5728.46	-1.25
191	01/10/2002	23.19	1.58	5926.66	3.46
192	02/10/2002	22.64	-2.37	5827.71	-1.67
193	03/10/2002	22.48	-0.71	5898.38	1.21

194	04/10/2002	22.00	-2.14	5869.22	-0.49
195	07/10/2002	22.01	0.05	5853.55	-0.27
196	08/10/2002	22.63	2.82	5849.42	-0.07
197	09/10/2002	22.53	-0.44	5762.40	-1.49
198	10/10/2002	22.86	1.46	5762.15	0.00
199	11/10/2002	23.00	0.61	5845.33	1.44
200	14/10/2002	23.24	1.04	5865.18	0.34
201	15/10/2002	24.96	7.40	6040.32	2.99
202	16/10/2002	24.99	0.12	5924.94	-1.91
203	17/10/2002	24.65	-1.36	5985.77	1.03
204	18/10/2002	24.44	-0.85	5973.21	-0.21
205	21/10/2002	24.50	0.25	6017.37	0.74
206	22/10/2002	24.06	-1.80	5979.57	-0.63
207	23/10/2002	23.74	-1.33	6000.62	0.35
208	24/10/2002	23.10	-2.70	5908.07	-1.54
209	25/10/2002	23.43	1.43	5905.58	-0.04
210	28/10/2002	23.31	-0.51	5887.29	-0.31
211	29/10/2002	23.01	-1.29	5893.76	0.11
212	30/10/2002	23.70	3.00	5963.83	1.19
213	31/10/2002	24.41	3.00	5967.73	0.07
214	01/11/2002	24.56	0.61	6045.16	1.30
215	04/11/2002	25.00	1.79	6058.90	0.23
216	05/11/2002	24.99	-0.04	6040.17	-0.31
217	06/11/2002	24.98	-0.04	6064.00	0.39
218	07/11/2002	23.72	-5.04	6009.93	-0.89
219	08/11/2002	23.46	-1.10	5988.53	-0.36
220	11/11/2002	23.47	0.04	5891.79	-1.62
221	12/11/2002	23.49	0.09	5865.11	-0.45
222	13/11/2002	23.65	0.68	5813.36	-0.88
223	14/11/2002	23.43	-0.93	5898.06	1.46
224	15/11/2002	23.25	-0.77	5819.09	-1.34
225	18/11/2002	23.22	-0.13	5726.00	-1.60
226	19/11/2002	23.01	-0.90	5641.74	-1.47
227	21/11/2002	23.69	2.96	5859.05	3.85
228	22/11/2002	23.35	-1.44	5818.43	-0.69
229	25/11/2002	23.36	0.04	5861.82	0.75
230	26/11/2002	23.56	0.86	5922.41	1.03
231	27/11/2002	24.16	2.55	6129.25	3.49
232	28/11/2002	24.61	1.86	6158.49	0.48
233	29/11/2002	24.72	0.45	6156.83	-0.03
234	02/12/2002	24.87	0.61	6223.59	1.08
235	03/12/2002	24.88	0.04	6221.68	-0.03
236	04/12/2002	25.00	0.48	6187.67	-0.55
237	05/12/2002	25.10	0.40	6152.01	-0.58
238	06/12/2002	25.20	0.40	6126.23	-0.42
239	09/12/2002	25.00	-0.79	6053.75	-1.18
240	10/12/2002	24.49	-2.04	6081.52	0.46
241	11/12/2002	23.74	-3.06	6135.39	0.89
242	13/12/2002	23.48	-1.10	6114.20	-0.35
243	16/12/2002	24.02	2.30	6185.95	1.17
244	17/12/2002	23.87	-0.62	6168.46	-0.28
245	18/12/2002	23.94	0.29	6089.66	-1.28
246	19/12/2002	24.11	0.71	6120.46	0.51
247	20/12/2002	23.49	-2.57	6130.83	0.17

248	23/12/2002	23.70	0.89	6153.22	0.37
249	24/12/2002	23.75	0.21	6151.48	-0.03
250	26/12/2002	23.84	0.38	6182.91	0.51
251	27/12/2002	23.80	-0.17	6126.24	-0.92
252	30/12/2002	23.83	0.13	6124.51	-0.03
253	31/12/2002	24.00	0.71	6127.09	0.04

Utilice excel para realizar una regresión entre la variable independiente X=rendimiento del mercado y la variable dependiente Y=rendimiento de la acción.

En estas condiciones al obtener el modelo Y = a + bX, al término b se le llamará la beta de la acción que en finanzas mide el riesgo sistemático que no es diversificable por una cartera de valores.

Estadística Inferencial. Examen de Regresión Lineal

Nombre del alumno:_____

Una aplicación[10] importante del análisis de regresión en contabilidad es para estimar costos. Al reunir datos sobre volumen y costo, y aplicar el método de cuadrados mínimos para formar una ecuación de regresión donde se relaciona el volumen y el costo, un contador puede estimar el costo asociado con determinada operación de manufactura. Se obtuvo la siguiente muestra de volúmenes de producción y costo total para una operación de manufactura.

Volumen de producción (unidades)	Costo total ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

a) use los datos para deducir una ecuación de regresión con la que se pueda predecir el costo total para determinado volumen de producción.
b) Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación en el costo total puede explicar el volumen de producción?
c) Calcule el coeficiente de correlación.

[10] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 565

d) El programa de producción de la empresa indica que el mes próximo se deben producir 500 unidades. ¿Cuál será el costo total estimado para esta operación?

No. De Lista:_____

Estadística Inferencial. Examen de Regresión Lineal

Nombre del alumno:_____

¿A los directores[11] y principales ejecutivos se les paga de acuerdo con las ganancias obtenidas por la empresa? La siguiente tabla es una lista de datos corporativos sobre el cambio porcentual en el rendimiento de las acciones durante un periodo de dos años, y el cambio porcentual en la paga a los directores y principales ejecutivos, inmediatamente después de dos años.

Empresa	Cambio bianual en el Rendimiento (%)	Cambio en el pago Al ejecutivo (%)
Walt mart	201.3	18
Bodega Aurrera	146.5	28
Grupo Gigante	76.7	10
Comercial Mexicana	158.2	28
Home mart	-34.9	15
Price club	73.2	-9
Sams club	-7.9	-20

a) forme la ecuación de regresión con el cambio porcentual bianual de rendimiento de las acciones como variable independiente.
b) Calcule el coeficiente de correlación. ¿se sentiría cómo al usar el cambio porcentual bianual de rendimiento de las acciones para predecir el

[11] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 566

cambio porcentual en la paga de los principales ejecutivos? Comente sus razones.

c) ¿Refleja el coeficiente de correlación una relación intensa o débil entre el rendimiento y la compensación a ejecutivos?

No. De Lista:_____

Estadística Inferencial. Examen de Regresión Lineal

Nombre del alumno:_____

Suponga que con fines de instalar un restaurante más de la cadena “El Portón” se reunieron datos de una muestra de 10 restaurantes ubicados cerca de áreas industriales. Para el i-ésimo restaurante de la muestra, x_i es el tamaño de la población estudiantil, en miles, y y_i son las ventas trimestrales (en miles de pesos). Los valores de x_i y y_i para los 10 restaurantes de la muestra se resume en la siguiente tabla:

Restaurante i	Población de oficinistas x_i (miles)	Ventas trimestrales y_i ($ miles)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

e) Defina sus variables.
f) realice el diagrama de dispersión correspondiente.
g) Calcule e interprete el modelo de regresión.
h) Calcule el coeficiente de determinación.
i) Calcule el coeficiente de correlación.
j) ¿Cuáles serán las ventas estimadas para un restaurante situado cerca de una zona industrial con una población de 30 mil personas.

No. De Lista:_____

Nombre del alumno:_____

Una aplicación[12] importante del análisis de regresión en contabilidad es para estimar costos. Al reunir datos sobre volumen y costo, y aplicar el método de cuadrados mínimos para formar una ecuación de regresión donde se relaciona el volumen y el costo, un contador puede estimar el costo asociado con determinada operación de manufactura. Se obtuvo la siguiente muestra de volúmenes de producción y costo total para una operación de manufactura.

Volumen de producción (unidades)	Costo total ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

k) Defina sus variables.
l) realice el diagrama de dispersión correspondiente.
m) Calcule e interprete el modelo de regresión.
n) Calcule el coeficiente de determinación.
o) Calcule el coeficiente de correlación.
p) El programa de producción de la empresa indica que el mes próximo se deben producir 500 unidades. ¿Cuál será el costo total estimado para esta operación?

[12] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 565

No. De Lista:_____

Estadística Inferencial. Examen de Regresión Lineal .

Nombre del alumno:_____

¿A los directores[13] y principales ejecutivos se les paga de acuerdo con las ganancias obtenidas por la empresa? La siguiente tabla es una lista de datos corporativos sobre el cambio porcentual en el rendimiento de las acciones durante un periodo de dos años, y el cambio porcentual en la paga a los directores y principales ejecutivos, inmediatamente después de dos años.

Empresa	Cambio bianual en el Rendimiento (%)	Cambio en el pago Al ejecutivo (%)
Walt mart	201.3	18
Bodega Aurrera	146.5	28
Grupo Gigante	76.7	10
Comercial Mexicana	158.2	28
Home mart	-34.9	15
Price club	73.2	-9
Sams club	-7.9	-20

d) Defina sus variables.
e) Elabore el diagrama de dispersión correspondiente.
f) Haga e interprete el modelo de regresión correspondiente.
g) Calcule el coeficiente de determinación. ¿se sentiría cómo al usar el cambio porcentual bianual de rendimiento de las acciones para predecir el cambio porcentual en la paga de los principales ejecutivos? Comente sus razones.
h) ¿Refleja el coeficiente de correlación una relación intensa o débil entre el rendimiento y la compensación a ejecutivos?

[13] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 566

No. De Lista:_____

Estadística Inferencial. Examen de Regresión Lineal

Nombre del alumno:_____

Un economista[14] del DDF está preparando un estudio sobre el comportamiento del consumidor. Los datos que obtuvo los plasmo en la siguiente tabla:

Consumidor	1	2	3	4	5	6	7	8	9	10	11	12
Ingreso	24.3	12.5	31.2	28	35.1	10.5	23.2	10	8.5	15.9	14.7	15
Consumo	16.2	8.5	15	17	24.2	11.2	15	7.1	3.5	11.5	10.7	9.2

Para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo. Si los valores de ingreso como los de consumo están dados en miles de pesos:

a) defina las variables.
b) Haga el diagrama de dispersión correspondiente.
c) Calcule e interprete el modelo de regresión.
d) Calcule los coeficientes de determinación y de correlación e interprételos.
e) ¿Qué le dice este modelo sobre la relación entre el consumo y el ingreso?.
f) ¿Qué consumo pronosticaría el modelo para alguien que gana $27,500.00

[14] Allen L. Webster. Estadística aplicada a los negocios y la economía. Editorial: Irwin-McGrawHill. P.p 335. problema ·10.

No. De Lista:

Estadística Inferencial. Examen final

Nombre del alumno:_____

1. Walt Mart[15] llevó a cabo recientemente una investigación con el fin de medir el efecto del tráfico vehicular en las cercanías de ciertas tiendas sobre sus ventas anuales.
Para realizar esto de manera adecuada, se identificaron 20 tiendas prácticamente idénticas en cuanto a las demás variables con efecto significativo sobre las ventas (como superficie, disponibilidad de estacionamiento, datos demográficos de la colonia en que se ubican, entre otros). Este análisis específico forma parte del esfuerzo general que realiza Walt Mart para identificar y cuantificar los efectos de los diversos factores que ejercen impacto sobre las ventas de sus tiendas. Su meta final es desarrollar un modelo para evaluar sitios potenciales a fin de ubicar tiendas, con el fin de analizarlos y elegir los más convenientes y que produzcan mayores niveles de ventas, comprar el terreno y construir la tienda.
Tras identificar 20 sitios, la empresa realizó recuentos diarios del tráfico en cada punto durante 30 días. Además obtuvo de sus registros internos los datos de ventas totales de cada una de las 20 tiendas de prueba en los 12 meses anteriores. Tales datos se encuentran registrados en la siguiente tabla:

Número de tienda	*Conteo vehicular diario Promedio en miles*	*Ventas anuales en miles de pesos*
1	62	1,121
2	35	766
3	36	701
4	72	1304
5	41	832
6	39	782
7	49	977
8	25	503
9	41	733
10	39	839
11	35	893
12	27	588
13	55	957
14	38	703
15	24	497
16	28	657
17	53	1,209
18	55	997
19	33	844
20	29	883

a) ¿Cuáles serán las ventas anuales estimadas para una tienda que tenga un conteo vehicular promedio de 30,000. Encuentre los coeficientes de determinación y de correlación además de dar sus conclusiones.

2. considere[16] una agencia para renta de autos donde por experiencia se sabe que la desviación estándar de la población de millas por galón normalmente distribuida de sus carros es de cuatro millas por galón. Construya un intervalo de confianza del 99% para la media de millas por galón de la flota de la agencia de 100 carros, si para tal efecto se elige una muestra aleatoria simple, sin reemplazo, de 36 de sus carros, y la media resulta ser de 29. (ojo si n es mayor que el 5% de N, se utiliza el caso de población pequeña).

3. considere un experimento Binomial con N= 5 y Y= 2; Encuentre el estimador de máxima verosimilitud.

No. De Lista:_____

[15] McDa..., ..., Roger, 1999. Investigación de Mercados contemporánea. International thomson editores cuarta edición. P.p 560
[16] Kohler, Heinz, 1998. estadística para negocios y economía. Editorial CECSA. P.p 341

Estadística Inferencial. Examen final

Nombre del alumno:_____

1.Walt[17] Mart llevó a cabo recientemente una investigación con el fin de medir el efecto del tráfico vehicular en las cercanías de ciertas tiendas sobre sus ventas anuales.
Para realizar esto de manera adecuada, se identificaron 20 tiendas prácticamente idénticas en cuanto a las demás variables con efecto significativo sobre las ventas (como superficie, disponibilidad de estacionamiento, datos demográficos de la colonia en que se ubican, entre otros). Este análisis específico forma parte del esfuerzo general que realiza Walt Mart para identificar y cuantificar los efectos de los diversos factores que ejercen impacto sobre las ventas de sus tiendas. Su meta final es desarrollar un modelo para evaluar sitios potenciales a fin de ubicar tiendas, con el fin de analizarlos y elegir los más convenientes y que produzcan mayores niveles de ventas, comprar el terreno y construir la tienda.
Tras identificar 20 sitios, la empresa realizó recuentos diarios del tráfico en cada punto durante 30 días. Además obtuvo de sus registros internos los datos de ventas totales de cada una de las 20 tiendas de prueba en los 12 meses anteriores. Tales datos se encuentran registrados en la siguiente tabla:

Número de tienda	Conteo vehicular diario Promedio en miles	Ventas anuales en miles de pesos
1	62	1,121
2	35	766
3	36	701
4	72	1304
5	41	832
6	39	782
7	49	977
8	25	503
9	41	733
10	39	839
11	35	893
12	27	588
13	55	957
14	38	703
15	24	497
16	28	657
17	53	1,209
18	55	997
19	33	844
20	29	883

a) identifique las variables que intervienen en el problema.
b) elabore el diagrama de dispersión correspondiente
c) elabore el modelo de regresión lineal correspondiente
d) calcule el coeficiente de correlación e indique que tipo de relación existe entre las variables (fuerte, débil, etc.)
e) de sus conclusiones.

2. considere[18] un lote de 800 cabezas de ganado de engorda que está por enviarse por tres. Con ayuda de una muestra aleatoria simple de 30 animales, sin reemplazo, construya un intervalo de confianza del 90% del peso medio por animal de todo el embarque. El peso medio muestral por cabeza resulta ser de 1,301 libras, con una desviación estándar muestral de 290 libras. (ojo, si n es menor que el 5% de N, entonces se aplica el caso de población grande)

3. considere un experimento Binomial con N= 5 y Y= 2; Encuentre el estimador de máxima verosimilitud.

No. De Lista:_____

[17] M... Gates, Roger, 1999. Investigación de Mercados contemporánea. International thomson editores cuarta edición. P.p
[18] Kohler, Heinz, 1998. estadística para negocios y economía. Editorial CECSA. P.p 342

Nombre del alumno:_____

1. suponga[19] que usted forma parte de un grupo de protección al consumidor, y esta interesado en determinar si el peso promedio de cierta marca de detergente, empacado en paquetes de 15 onzas, es menor que el peso anunciado, para lo cual, usted elige una muestra aleatoria de 50 bolsas, de las cuales obtiene una media de 14.4 onzas y una desviación estándar de 1.2, si el nivel de significancia es del 5%, concluya usted si la marca de detergente cumple con las especificaciones indicadas en la bolsa.

2. Un fabricante[20] de autos, molesto por las frases publicitarias de un rival, desea estimar la diferencia entre la media de kilómetros por litro de dos modelos de autos, A y B; se desea un intervalo de confianza del 98% para la diferencia entre las medias de kilómetros por litro. Se observan diez pares de pilotos, apareados según su habilidad para conducir autos; la media y la desviación estándar de las diferencias entre kilómetros por litro alcanzadas por el modelo A y las alcanzadas por el modelo B se encuentran como:

$$\varpi = 5 Km/litro \qquad \text{y} \qquad S_{\omega} = 2 Kilómetros/litro$$

3. Una aplicación[21] importante del análisis de regresión en contabilidad es para estimar costos. Al reunir datos sobre volumen y costo, y aplicar el método de cuadrados mínimos para formar una ecuación de regresión donde se relaciona el volumen y el costo, un contador puede estimar el costo asociado con determinada operación de manufactura. Se obtuvo la siguiente muestra de volúmenes de producción y costo total para una operación de manufactura.

Volumen de producción (unidades)	Costo total ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

q) Identifique sus variables.
r) Realice el diagrama de dispersión correspondiente.
s) Realice el modelo de regresión lineal
t) Calcule el coeficiente de correlación.
u) El programa de producción de la empresa indica que el mes próximo se deben producir 500 unidades. ¿Cuál será el costo total estimado para esta operación?
v) De sus conclusiones.

4. suponga que independientemente de lo que sucede el resto de los días, el número de trabajos que llegan en un día a un taller mecánico tiene una distribución de Poisson con media desconocida (mu). Suponga además que el primer día de la muestra llega sólo un trabajo y que el segundo (y último) día llegan cuatro. Escriba la función de verosimilitud.

No. De Lista:_____

[19] Weimer, Richard C. 1999. **Estadìstica**. Editorial: cecsa. México. P.p 470
[20] Pag. 784 del Kohler
[21] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 565

Nombre del alumno:_____

1. se supone que una tableta[22] para bajar la temperatura contiene 10 gramos de aspirina. Una muestra aleatoria de 100
tabletas produjo una media de 10.2 gramos y una desviación estándar de 1.4. ¿podemos concluir que la media es diferente de 10 con un nivel de significancia del 5%?

2. ¿A los directores[23] y principales ejecutivos se les paga de acuerdo con las ganancias obtenidas por la empresa? La siguiente tabla es una lista de datos corporativos sobre el cambio porcentual en el rendimiento de las acciones durante un periodo de dos años, y el cambio porcentual en la paga a los directores y principales ejecutivos, inmediatamente después de dos años.

Empresa	Cambio bianual en el Rendimiento (%)	Cambio en el pago Al ejecutivo (%)
Walt mart	201.3	18
Bodega Aurrera	146.5	28
Grupo Gigante	76.7	10
Comercial Mexicana	158.2	28
Home mart	-34.9	15
Price club	73.2	-9
Sams club	-7.9	-20

i) forme la ecuación de regresión con el cambio porcentual bianual de rendimiento de las acciones como variable independiente.
j) Calcule el coeficiente de correlación. ¿se sentiría cómo al usar el cambio porcentual bianual de rendimiento de las acciones para predecir el cambio porcentual en la paga de los principales ejecutivos? Comente sus razones.
k) ¿Refleja el coeficiente de correlación una relación intensa o débil entre el rendimiento y la compensación a ejecutivos?

3. Un fabricante[24] de autos, molesto por las frases publicitarias de un rival, desea estimar la diferencia entre la media de kilómetros por litro de dos modelos de autos, A y B; se desea un intervalo de confianza del 98% para la diferencia entre las medias de kilómetros por litro. Se observan diez pares de pilotos, apareados según su habilidad para conducir autos; la media y la desviación estándar de las diferencias entre kilómetros por litro alcanzadas por el modelo A y las alcanzadas por el modelo B se encuentran como:

$$\varpi = 5 Km/litro \qquad \text{y} \qquad S_{\omega} = 2 Kilómetros/litro$$

4. considere un experimento Binomial con N= 5 y Y= 2; Encuentre el estimador de máxima verosimilitud.

[22] Weimer, Richard C. 1999. **Estadìstica**. Editorial: cecsa. México. P.p 473
[23] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 566
[24] Pag. 784 del Kohler

No. De Lista:_____

Estadística Inferencial. Examen Final

Nombre del alumno:_____

1. una escuela comercial anuncia[25] que sus alumnos pueden llegar a escribir un promedio de 80 palabras por minuto (ppm) cuando se gradúan. Se examinó una muestra de 60 graduados recientes y los resultados mostraron una media de 78 ppm y una desviación estándar de 6.2 ppm. A un nivel de significancia de 0.05 ¿tiene razón la escuela en su anuncio?

2. Un fabricante[26] de autos, molesto por las frases publicitarias de un rival, desea estimar la diferencia entre la media de kilómetros por litro de dos modelos de autos, A y B; se desea un intervalo de confianza del 98% para la diferencia entre las medias de kilómetros por litro. Se observan diez pares de pilotos, apareados según su habilidad para conducir autos; la media y la desviación estándar de las diferencias entre kilómetros por litro alcanzadas por el modelo A y las alcanzadas por el modelo B se encuentran como:

$$\varpi = 5 Km/litro \quad \text{y} \quad S_{\omega} = 2 Kilómetros/litro$$

3. ¿A los directores[27] y principales ejecutivos se les paga de acuerdo con las ganancias obtenidas por la empresa? La siguiente tabla es una lista de datos corporativos sobre el cambio porcentual en el rendimiento de las acciones durante un periodo de dos años, y el cambio porcentual en la paga a los directores y principales ejecutivos, inmediatamente después de dos años.

Empresa	Cambio bianual en el Rendimiento (%)	Cambio en el pago Al ejecutivo (%)
Walt mart	201.3	18
Bodega Aurrera	146.5	28
Grupo Gigante	76.7	10
Comercial Mexicana	158.2	28
Home mart	-34.9	15
Price club	73.2	-9
Sams club	-7.9	-20

l) forme la ecuación de regresión con el cambio porcentual bianual de rendimiento de las acciones como variable independiente.
m) Calcule el coeficiente de correlación. ¿se sentiría cómo al usar el cambio porcentual bianual de rendimiento de las acciones para predecir el cambio porcentual en la paga de los principales ejecutivos? Comente sus razones.
n) ¿Refleja el coeficiente de correlación una relación intensa o débil entre el rendimiento y la compensación a ejecutivos?

4. suponga que independientemente de lo que sucede el resto de los días, el número de trabajos que llegan en un día a un taller mecánico tiene una distribución de Poisson con media desconocida (mu). Suponga además que el primer día de la muestra llega sólo un trabajo y que el segundo (y último) día llegan cuatro. Escriba la función de verosimilitud.

[25] Weimer, Richard C. 1999. **Estadística**. Editorial: cecsa. México. P.p 475
[26] Pag. 784 del Kohler
[27] Anderson, Sweeney & Williams. 1999. estadística para administración y economía. International Thomson editores. P.p 566

CAPITULO VII. SERIES DE TIEMPO

Objetivos del capitulo

Después de estudiar este capítulo, se deberá estar en condiciones de:

1. Explicar el término “serie de tiempo”.

2. Describir el enfoque clásico al análisis de las series de tiempo, mediante la identificación de los cuatro componentes de una serie y la explicación de cada uno de ellos

3. Identificar los modelos de series de tiempo “aditivo” y “multiplicativo”.

4. Resolver problemas sencillos de series de tiempo.

Una serie de tiempo es un conjunto de datos que se recopilan, registran u observan en incrementos sucesivos de tiempo. Una serie de tiempo se integra por distintos componentes los cuales resultan difíciles de distinguir a simple vista, siendo necesario proceder a la descomposición de la serie para tener una mejor idea de las causas de la variabilidad; al aislar cada uno de ellos, se facilita el proceso de análisis y en su caso también el de pronóstico.

Componentes de una serie de tiempo

Existen cuatro componentes de una serie de tiempo: la tendencia, la variación cíclica, la variación estacional y la variación irregular o errática.

Tendencia secular

Las tendencias a largo plazo en ventas, empleo, precios de valores o acciones, y otras series de negocios y económicas, se ajustan a diversos esquemas. Algunas se mueven continuamente hacia arriba, otras declinan y otras más permanecen igual en un cierto periodo o intervalo de tiempo

Variación cíclica

La segunda componente de una serie de tiempo es la variación cíclica. El ciclo normal en un negocio consiste en un periodo de prosperidad seguido de periodos de recesión, depresión, y luego, recuperación. Se observan fluctuaciones considerables que representan más de un año, arriba y debajo de la tendencia secular.

Variación estacional

La tercera componente de una serie de tiempo es la variación estacional. Muchas series como ventas, producción y otras, fluctúan según las estaciones del año. La unidad de tiempo indicada es el lapso trimestral o mensual.

Variación irregular

Muchos analistas prefieren subdividir la variación irregular en variaciones episódicas y residuales. Las episódicas no son predecibles, pero pueden identificarse. El impacto inicial en la economía de una huelga importante o una guerra, puede identificarse, pero no es posible predecir un paro laboral o un conflicto bélico. Después de que las fluctuaciones episódicas se han eliminado, a la variación restante se le llama variación residual. Los cambios residuales , comúnmente conocidos como fluctuaciones aleatorias, sea episódica o residual, puede proyectarse a futuro.

Para aislar y comprender los elementos de una serie de tiempo, debe tenerse en cuenta las relaciones matemáticas que los unen.

Existen dos modelos para realizar el análisis de series de tiempo; uno recibe el nombre de “multiplicativo” y el otro “aditivo”. El primero de ellos considera una serie cronológica como si fuera la resultante del producto de los componentes individuales, en tanto que la última la considera como si fuera la resultante de la suma de los componentes individuales. De este modo el modelo multiplicativo tiene la forma

$$Y = T \times C \times E \times I$$

Donde:

Tendencia secular (**T**). La tendencia secular es el componente que representa el comportamiento (crecimiento o decrecimiento), en un periodo largo de tiempo.

Variación cíclica (**C**). El componente cíclico es la fluctuación que puede observase ocurre alrededor de la tendencia, Cualquier patrón regular de variaciones arriba o debajo de la recta que representa a la tendencia puede atribuirse a la componente cíclica.

Variación estacional (**E**). El componente estacional muestra un comportamiento regular en los mismos periodos de tiempo, reflejando costumbres o modas que se repiten regularmente dentro del periodo de observación. En la gráfica la estacionalidad quedaría representada por ejemplo por las variaciones semanales en los rendimientos, no visibles por el periodo de información que se está manejando.

Variación irregular (**I**). Es el componente que queda después de separar a otras componentes, es el resultado de factores no explicables que siguen un comportamiento aleatorio, siendo por ello una parte no previsible de la serie.

Y el modelo aditivo tiene la forma:

$$Y = T + C + E + I$$

En ambos modelos el resultado de la tendencia es una cantidad real (por ejemplo 50 000 toneladas). Aunque parece mas sencillo trabajar con el modelo aditivo, el modelo multiplicativo se utiliza mas, debido principalmente a que representa de manera mas adecuada la experiencia real, al expresar frecuentemente las variaciones c´cilicas, estacionales e irregulares como porcentajes de la tendencia secular

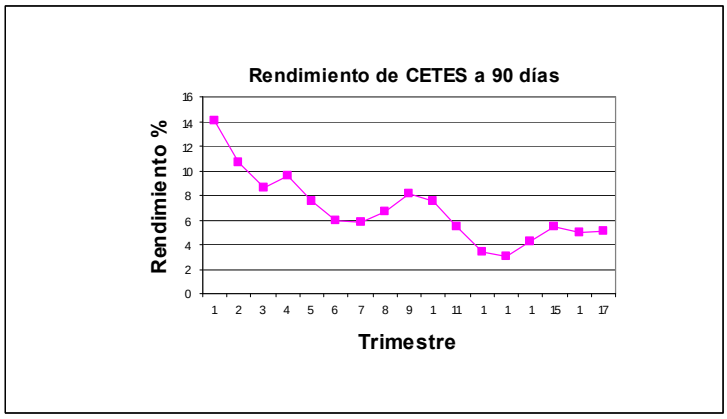
Ejemplo:
Supongamos que tenemos la información siguiente, correspondiente al comportamiento del rendimiento de los CETES a 90 días, estos valores representan una serie de tiempo

Rendimiento de CETES a 90 días

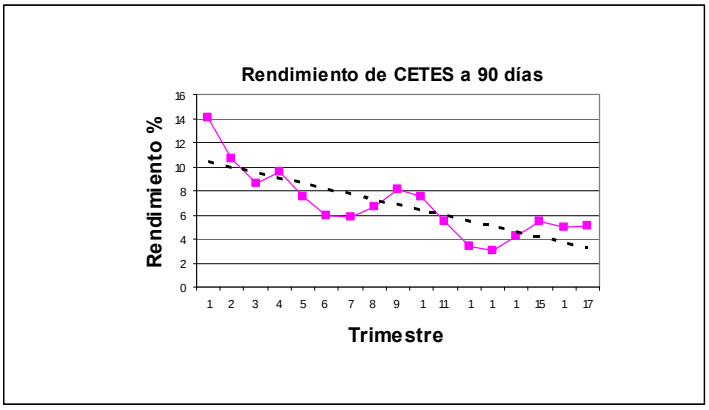
Trimestre	%
1	14.03
2	10.69
3	8.63
4	9.58
5	7.48
6	5.98
7	5.82
8	6.69
9	8.12
10	7.51
11	5.42
12	3.45
13	3.02
14	4.29
15	5.51
16	5.02
17	5.07

El registro de rendimientos trimestrales de los CETES representan una serie de tiempo, ya que se han obtenido en periodos sucesivos.

Si se analiza el registro podemos observar que han una disminución en los valores de rendimiento, de mayor a menor, pero nos resulta difícil afirmar en que proporción ha ocurrido y de cuánto han sido las variaciones. Si este registro lo analizamos como una serie tendremos la gráfica siguiente:



Ahora separando el componente de tendencia secular, tendremos que queda representado por la línea que atraviesa la serie de un extremo a otro, indicándonos que ha existido un comportamiento con tendencia a la baja.



La separación de la tendencia secular es posible de realizar, al calcular la recta de regresión a la serie de datos, utilizando el método de mínimos cuadrados analizado en el capítulo anterior. De esta manera podemos conocer la ecuación matemática que representa la tendencia secular, su pendiente y su ordenada al origen, con lo cual estaríamos en condiciones de conocer en cada punto la tendencia del rendimiento.

Supongamos ahora que nos interesa conocer la variación que han tenido los rendimientos respecto de la tendencia, es decir el componente cíclico, el cual queda representado en la gráfica por los valores mayores y menores respecto de la tendencia secular. Si deseamos conocer el valor numérico de este comportamiento podemos proceder como sigue:

Calcular para cada fecha de interés el valor del rendimiento de acuerdo con la ecuación de la tendencia (Yt) y compararlo con el correspondiente del registro, estableciendo una proporción entre estos dos valores de la manera siguiente:

$$C = \frac{Y}{Yt} (100)$$

En donde:

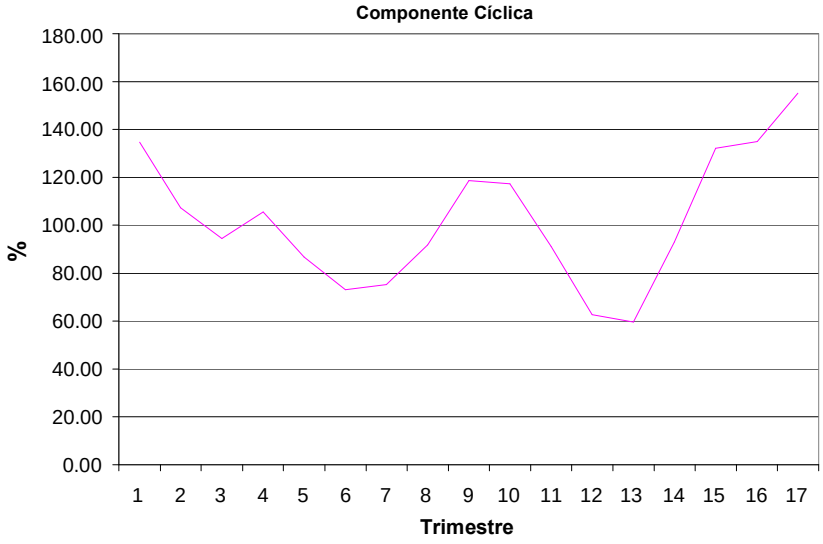
Y representa el rendimiento registrado.
Yt representa el rendimiento calculado con la ecuación de tendencia.

De esta manera es posible calcular el valor del componente cíclico, en porcentaje, respecto del valor de la tendencia secular: Los valores que estén por encima de la ecuación de la tendencia alcanzarán un porcentaje superior a cien, mientras que los que se encuentren por debajo de ella tendrán valores inferiores a cien.

Cálculo del componente cíclico

Trimestre	Y	Yt	C
1	14.03	10.41	134.77
2	10.69	9.96	107.33
3	8.63	9.15	94.32
4	9.58	9.07	105.62
5	7.48	8.62	86.77
6	5.98	8.18	73.11
7	5.82	7.73	75.29
8	6.69	7.29	91.77
9	8.12	6.84	118.71
10	7.51	6.4	117.34
11	5.42	5.95	91.09
12	3.45	5.5	62.73
13	3.02	5.07	59.57
14	4.29	4.61	93.06
15	5.51	4.17	132.13
16	5.02	3.72	134.95
17	5.07	3.27	155.05

Puede observarse en la columna correspondiente al componente cíclico las variaciones por arriba y por abajo al 100%, reflejando el comportamiento respecto de los valores de la tendencia secular, es posible elaborar una gráfica con estos valores, observemos que se representan respecto de una recta horizontal ubicada en el valor 100%, ahora es posible ver con mucha claridad cual ha sido el comportamiento de los rendimientos respecto de la tendencia secular. Podemos observar que las fluctuaciones a la baja han sido más importantes que las correspondientes a la alza. El componente cíclico se separa siempre de manera posterior a la tendencia secular.



El análisis del componente estacional requiere disponer de un número importante de datos. Por el método que se utiliza para descomponerlo, cuando la serie de tiempo contiene datos diarios, semanales o mensuales, el primer componente que debe ser aislado es el estacional.

Finalmente se puede mencionar que en el ejemplo anterior, no se identifica alguna variación irregular.

Bibliografía:

Estadística aplicada a la administración y a la economía
Autores: David K. Hildebrand y
R. Lyman Ott
Editorial: Addison Wesley Longman.

Matemáticas avanzadas para Ingeniería
Autor: Erwin Kreyszig
Editorial: Limusa. Tercera edición. Vol. 2

Rescate de empresas en crisis
Autores[1]: E. González; L. Leyva; y C. Ruiz
International thomson editores

Calculo con geometría analítica
Autor: Earl W. Swokowski
Editorial: grupo editorial iberoamérica.

[1] Los tres autores tienen entre otros grados académicos, el grado de Master en dirección de empresas por el IPADE (instituto panamericano de alta dirección de empresa)