UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

AUTOR: ELISEO FLORES ALAMILLA

Estadística II		Clave: 1353
Plan: 2005		Créditos: 8
Licenciatura: Administración y Contaduría		Semestre: 3
Área: Matemáticas		Hrs. Asesoría: 4
Requisitos: Seriación antecedente obligatoria: Estadística I		Hrs. por semana: 4
Tipo de asignatura:	Obligatoria (x)	Optativa ( )

Objetivo general de la asignatura

El alumno inferirá las características de una población, con base en la información contenida y contrastará diversas pruebas para la toma de decisiones.

Temario oficial (horas sugeridas 64 hrs.)

1. Teoría del muestreo (8 hrs.)
2. Distribuciones muestrales y el teorema central del límite (10 hrs.)
3. Estimación de parámetros e intervalos de confianza (10 hrs.)
4. Pruebas de hipótesis (14 hrs.)
5. Estadística no paramétrica (10 hrs.)
6. Análisis de regresión y correlación lineal (8hrs.)
7. Series de tiempo (4 hrs.)

Introducción

En esta asignatura el estudiante dará continuación al curso previo de Estadística I. Observando la importancia que tiene el aprenderla, así:

En el **tema 1** investigará y aplicará la teoría del muestreo a diferentes tipos de problemas y, en consecuencia, diferentes tipos de muestras. Observará los retos que implica la correcta selección de una muestra con el objetivo de que su estudio tenga la validez científica y la exactitud de la matemática.

En el **tema 2** estudiará las distribuciones muestrales y el teorema central del límite, los cuales pueden ayudar para la posterior elaboración de los intervalos de confianza.

En el **tema 3** estimará los parámetros principales con el fin de tomar decisiones en un entorno de incertidumbre.

En el **tema 4** aplicará las pruebas de hipótesis en el ambiente administrativo y contable para poder decidir continuar o desechar alguna forma de actuar de la compañía donde se encuentre laborando, basado en hechos científicos.

En el **tema 5** analizará la estadística no paramétrica para poder racionalizar fenómenos que no son cuantificables, pero que por su importancia merecen ser estudiados.

En el **tema 6** investigará el análisis de regresión lineal para averiguar el comportamiento de las variables y sus diferentes relaciones de un estudio de tipo "pronóstico".

En el **tema 7** analizaremos las series de tiempo para observar su aplicación a diferentes problemas de la vida diaria de las empresas.

Tema 1. Teoría del muestreo

Objetivo particular

En esta unidad, el alumno investigará y analizará tanto el concepto como las bases matemáticas de la teoría del muestreo, tomándolo como el punto de partida para el estudio completo de la estadística inferencial.

Aplicará también los conceptos aprendidos a situaciones reales, observando de esta manera la importancia del muestreo en el ámbito profesional.

Temario detallado

1. Teoría del muestreo

1.1 Introducción al muestreo

1.2 Diferentes tipos de muestreo

1.3 Estimación de parámetros

1.1 Introducción al muestreo

La teoría del muestreo estudia la relación entre una población y las muestras tomadas de ella; es decir, se utiliza para **estimar** magnitudes desconocidas de una población —tales como valores promedio y de dispersión, 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 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 del análisis de diferentes muestras obtenidas de ésta, con indicación de la precisión de tales inferencias, se llama **inferencia estadística**.

La teoría de las probabilidades es el fundamento de los métodos de muestreo; para usarla hay que poseer un buen nivel de conocimiento, desde el punto de vista de la matemática, de álgebra, cálculo y probabilidades, así como de los **métodos generales de estadística** y de la teoría básica de las estimaciones, desde el punto de vista estadístico; todo ello es esencial para un entendimiento adecuado del desarrollo riguroso de la teoría del muestreo.

Así pues, "**muestreo**" es el proceso para obtener información acerca del conjunto de una población o universo examinando sólo una parte del mismo.

1.2 Diferentes tipos de muestreo

Existen básicamente dos métodos para seleccionar una muestra. Si cada elemento de una población tiene la misma posibilidad de ser seleccionado para integrar la muestra, el método se denomina **muestro aleatorio**; por el contrario, si los elementos tienen diferentes posibilidades de ser elegidos, el método se denomina **muestreo no aleatorio**.

Cuando un muestreo se realiza devolviendo al conjunto el elemento una vez analizado se dice que el muestreo se realizó con reemplazo; si el elemento seleccionado no es regresado al conjunto, el muestreo es sin reemplazo. Esta condición resulta muy importante cuando se desea asignar un valor de probabilidad a la selección.

Su ventaja es que todos los datos tienen la misma posibilidad de ser seleccionados y en consecuencia podemos obtener información importante de la población de la cual fue extraída la muestra y, su desventaja es que si la población es heterogénea o que se encuentre agrupada en segmentos de diferentes tamaños, entonces la muestra

puede no ser representativa de la población, debido a que si uno de los segmentos de la población es muy pequeño entonces cabe la posibilidad de que ninguno de sus elementos pueda ser incluido en la muestra y en consecuencia no ser tomado en cuenta.

- **Muestreo aleatorio sistemático**

Aclaremos esto observando que el procedimiento en este tipo de muestreo, se acomodan los elementos o personas de la población de forma ascendente de preferencia y se selecciona un punto de partida aleatorio y luego se toma cada k-esimo miembro para formar la muestra.

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; posteriormente, se seleccionaría luego una muestra de, por ejemplo, 100 números utilizando una tabla de números aleatorios; luego, en los cajones deberá localizarse una factura que concuerde con cada uno de estos 100 números; en fin, esta tarea puede requerir mucho tiempo. En lugar de ello, se podría seleccionar una **muestra aleatoria sistemática** utilizando el siguiente método: se recorren simplemente los cajones y se cuentan las facturas; finalmente, se toman las que coincidan con el número 20 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, etcétera.

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 cuasi-aleatorio.

La ventaja para este tipo de muestreo sería que es más rápido que un muestreo aleatorio formal y su desventaja es que puede no reflejar información importante contenida en el conjunto de datos debido a que no todos los elementos estrictamente hablados, tienen la misma oportunidad de ser seleccionados.

- **Muestreo aleatorio estratificado**

Otro tipo de muestreo es el aleatorio estratificado[1], que divide una población en subgrupos llamados **estratos** y se selecciona una muestra de cada uno de ellos con lo cual se garantiza la representación de cada subgrupo o estrato.

Una vez que la población se divide en estratos, es posible seleccionar una **muestra proporcional** o **no proporcional**. Como el nombre señala, 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 en mercadotecnia de las 352 empresas mexicanas más grandes seleccionadas por la revista "*Fortune*". Suponga que el objetivo de 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 mercadotecnia que las empresas que tienen un menor rendimiento o incluso un déficit.

Suponga que las 352 empresas se dividieron en cinco estratos; si seleccionamos una muestra de 50 empresas, entonces de acuerdo al muestreo aleatorio estratificado se deberían incluir:

[1] Douglas A. Lind *et al.*, *Estadística para administración y economía*, p. 226

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 la quinta columna de la tabla anterior, podemos observar los cálculos realizados para determinar el número de elementos muestreados por estrato, garantizando con este procedimiento, que cada uno de los estratos de interés, se encuentra representado en la muestra a estudiar.

Una muestra estratificada no proporcional es aquella en la cual, la cantidad de elementos que se seleccionan en cada estrato no guarda proporción con la cantidad de elementos 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, dado que puede darse el caso en ambos muestreos (aleatorio simple o sistemático), de que alguno de los estratos de interés no quede considerado en la muestra al no ser elegido al menos alguno de sus elementos y la desventaja para este tipo de muestreo estratificado es que puede caerse en el exceso de estratos haciendo el proceso de muestreo más difícil y tardado que si aplicamos un muestreo aleatorio simple.

- **Muestreo por conglomerados**[2]

Otro tipo de muestreo que es común es el **muestreo por conglomerados**. Se entiende como conglomerado de elementos de una población, a cualquier subconjunto de la misma, que se defina como tal, es decir, como un conglomerado.

[2] Douglas A. Lind. *et al.*, *Estadística para administración y economía*, pp. 227.

La definición de un conglomerado[3], así como su tamaño, se definen y dependen de los objetivos del estudio que se esté realizando, y en general, los conglomerados definidos en un estudio pueden o no tener el mismo tamaño.

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 2004. 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).

- **Tamaño de la muestra**

Para la determinación del tamaño de la muestra se requiere tomar en consideración la mayor cantidad posible de los siguientes elementos:[4]

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.

[3] Rosalinda Flores García. et al., *Estadística aplicada a la administración*. pp. 225.
[4] Jesús Galindo Caceres, *Técnicas de investigación en sociedad, cultura y comunicación*, pp. 49-62.

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.
12. Recursos disponibles.

Dependiendo del problema mismo, no todos los problemas incluyen la totalidad de los elementos mencionados. Como es de observarse, dentro de las teorías del muestreo y probabilidad existen diversos procedimientos para el cálculo de los tamaños de la muestra; todos ellos consideran a la mayoría de 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:

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

Ejemplo. Se requiere 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 como: 0.05
Nc	95% de nivel de confianza o exactitud. Que expresado como valor “z” que determina el área de probabilidad buscada queda como: 1.96[1]

Al sustituir estos valores en la fórmula, tenemos:

$$n = \frac{(3{,}000{,}000)(0.5)(0.5)}{\left[\frac{(0.05)^2}{(1.96)^2}(3{,}000{,}000-1)\right]+(0.5)(0.5)}$$

de donde, al realizar las operaciones indicadas, el valor de “n” obtenido es de 384.1. y finalmente haciendo un redondeo, el tamaño de la muestra será de 384 elementos.

El valor de “z” se busca en las tablas de distribución normal estándar y la forma de encontrarlo es la siguiente:

1º. El porcentaje deseado entre 2 (debido a la simetría de la curva de distribución normal), en este caso el resultado sería:

$$\frac{95}{2} = 47.5$$

2º. Este resultado (47.5) se divide entre 100 para convertirlo de porcentaje a decimal, es decir:

$$\frac{47.5}{100} = 0.475$$

3º. Este valor de 0.475 se busca en el cuerpo de la tabla de la curva de distribución normal estándar (La mayoría de los textos de probabilidad y estadística contienen esta tabla), donde encontramos el valor correspondiente de z = 1.96.

- **Error en el muestreo**[5]

Se denomina así a las diferencias que se presentan entre los resultados obtenidos en el análisis de las muestras respecto de los que en realidad corresponden a la población. Este tipo de error, se presenta con mayor intensidad cuando las muestras no son representativas de la población de la cual fueron extraídas, sin embargo, aún cuando las muestras son extraídas utilizando técnicas de muestreo aleatorio, el llamado **error de muestreo** se presenta, y aunque su presencia es azarosa, se tiene la ventaja de que en este caso el error puede ser calculado y analizado con el objetivo claro de minimizarlo.

Así, la diferencia entre un estadístico de la muestra y un parámetro de la población se conoce como **error de muestreo**.

Es importante mencionar que también existen los **errores no muestrales**, entre los cuales podemos incluir los cometidos durante el procesamiento de los datos, cuando hacen falta datos, errores de registro, errores de respuesta, errores de definición, cuestionarios mal elaborados, etc. En la práctica es muy importante reducir estos lo más posible, debido a que prácticamente no existen métodos estadísticos para medir o controlar estos errores no muestrales.

[5] Douglas A. Lind *et al.*, *Estadística para administración y economía*, p. 229.

1.3 Estimación de parámetros

Desde un punto de vista práctico, es muy importante ser capaz de inferir información sobre una población a partir de muestras suyas. Con tal situación se enfrenta 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).

- **Método de máxima verosimilitud**

En cualquier situación de muestreo es posible encontrar un estimador de un parámetro, utilizando el **método de máxima verosimilitud** de R. A. Fisher, 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.
- 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**

Si denotamos a la función de verosimilitud con la letra “L” y la definimos como la probabilidad de observar los datos tomados de manera independiente de una variable aleatoria cualquiera, entonces dicha función de verosimilitud tendrá la forma siguiente:

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

en el caso discreto y la siguiente forma en el caso continuo:

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

Como podemos observar, independientemente de cual fuere el caso (variable aleatoria discreta o variable aleatoria continua), la función de verosimilitud se obtiene simplemente sustituyendo en la función original cada uno de los datos y multiplicando la función por si misma para cada uno de los casos.

Por ejemplo 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:

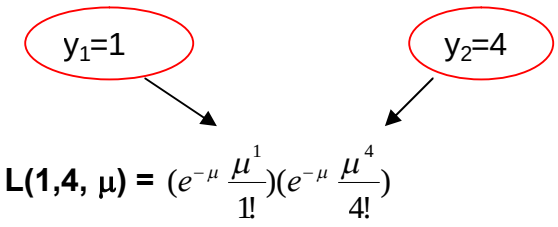
Primer paso. Debemos escribir la fórmula básica de la cual se parte y debemos identificar exhaustivamente todas sus variables; en este caso, la fórmula corresponde a una distribución de Poisson; por lo tanto, recordando que la distribución de Poisson es discreta con:

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

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, considerando la teoría de la función de verosimilitud. Los valores observados son $\mathbf{y_1=1}$ e $\mathbf{y_2=4}$; por lo tanto, la función de versomilitud 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:

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

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

La idea básica[6] del método de máxima verosimilitud es muy sencilla; su desarrollo es el siguiente:

Se elige aquella aproximación para el valor desconocido que en este caso y para efectos de explicación llamaremos **a** de manera que "**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:

Se escribe una derivada parcial debido a que "**L**" también depende de: y_1, y_2,...,y_n y una estimación de ésta ecuación: $\frac{\partial L}{\partial \mathrm{A}} = 0$ 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 la ecuación: $\frac{\partial L}{\partial \mathrm{A}} = 0$ por:

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

debido a que $f \geq 0$; un máximo de "f" en general es positivo y "ln (L)" es una función monótona creciente[7] de "L". Esto a menudo simplifica los cálculos.

---

[6] Erwin Kreyszig, *Matemáticas avanzadas para ingeniería*, vol. 2, p. 959.

[7] 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.

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 $\frac{\partial L}{\partial \mathrm{A}} = 0$ 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 $\frac{\partial \ln(L)}{\partial \mathrm{A}} = 0$ tenemos:

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

Por lo tanto, continuando con el ejemplo anterior tenemos que la función de verosimilitud era:

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

En donde el valor desconocido es en este caso μ

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

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

en 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:

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

Posteriormente, al obtener la **primera derivada** a esta ecuación, ésta cobra la siguiente forma:

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

Si a la ecuación anterior le aplicamos las leyes de la derivación matemática, tenemos que esta expresión se convierte en:

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

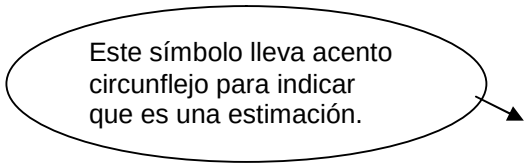
Continuando con el proceso, **igualamos a "cero" esta primera derivada**, por lo que la expresión resultante se indica a continuación:

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

que es lo mismo que:

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

Resolviendo la **última ecuación de primer grado** con una incógnita tenemos que:



$$\hat{u} = 2.5$$

De modo que la estimación de máximo verosímil o de máxima verosimilitud de μ es **$\hat{u}$=2.5.**

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

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 y considerando la teoría de la función de verosimilitud).
Tercer paso	Aplicar la función del 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.

- **Estimación por el método de momentos**

Otra forma de hacer una estimación puntual de un parámetro es a través del llamado método de los momentos, el cual es otra metodología utilizada, en la cual, se igualan los momentos muestrales con los momentos poblacionales.

Si consideramos que el primer momento poblacional es E(X) (valor esperado de X), el segundo momento poblacional es $E(X^2)$ y 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 $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ y así sucesivamente.

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

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

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.

$$\mathrm{E}(\mathrm{x}) = \int x f_x(x) dx$$

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

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, si la variable estudiada X es el porcentaje de agrado de un producto y dicho porcentaje (de 0 a 100) se distribuye de acuerdo con 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 el comportamiento de la población y en la cual el promedio es $E(X) = \frac{\theta+1}{\theta+2} = \frac{-0.36+1}{-0.36+2} = 0.39$; asimismo se puede calcular la mediana, moda, varianza, entre otras características.

Resulta claro que siendo un **estimador puntual**, un estadístico tomado de una muestra que es utilizado para estimar un parámetro, dicho estimador es tan bueno como lo sea la muestra de la cual proviene, sin embargo, para diferentes muestras representativas de la misma población, se tendrán diferentes estimaciones puntuarles. Así las cosas, estimar un parámetro utilizando una estimación de

intervalo (que veremos en el tema 3) resulta muchas veces preferible a utilizar una estimación puntual.

Bibliografía del tema 1

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

FLORES, Rosalinda. et al., *Estadística aplicada a la administración*. Grupo Editorial Iberoamericana, 1998.

GALINDO Caceres Jesús, *Técnicas de investigación en sociedad, cultura y comunicación*, Addison Wesley Longman, 1998.

HILDEBRAN, David K. y Ott, R. Lyman, *Estadística aplicada a la administración y a la economía*, Addison Wesley, , 1998

KREYSZIG Erwin*, Matemáticas avanzadas para ingeniería*, *vol. 2*, Limusa, 1996.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill. Alfaomega, 2004.

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica, 1980.

WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.1.1. Elabora un glosario conceptual de los conceptos vistos en el tema, con los libros de referencia.

A.1.2. Elabora un resumen de las recomendaciones del muestreo del libro *Teoría del muestreo.*

A.1.3. Investiga en que tipo de estudios es conveniente utilizar los diferentes tipos de muestreo.

A.1.4. Consulta en Internet la siguiente dirección *www.fao.org*, en su buscador escribe muestreo y revisa los apartados que se desarrollaron en el tema.

A.1.5. Consulta los periódicos: “El Reforma” “El Financiero” y “El Economista” y encuentra y compara en al menos tres artículos donde se dé la metodología de muestreo utilizada en sus artículos.

A.1.6. Investiga cuál fue la metodología de muestreo utilizada en “el conteo rápido” de elecciones pasadas y realiza comparaciones de las mismas.

A.1.7. Elabora un cuadro con los puntos de vista de los autores que se citan en *Estadística aplicada a la administración y a la economía,* con respecto al método de máxima verosimilitud en la práctica.

Cuestionario de autoevaluación

1. ¿Qué es la teoría del muestreo?
2. ¿En qué situaciones es conveniente recurrir al muestreo?
3. ¿Cuáles son los soportes de la teoría del muestreo?
4. ¿Qué es un muestreo aleatorio simple?
5. ¿Para qué se utiliza la teoría del muestreo?
6. ¿Qué es un muestreo aleatorio sistemático?
7. ¿Qué es un muestreo aleatorio estratificado?
8. ¿Qué es un muestreo por conglomerados?
9. ¿Qué es el nivel de confianza?
10. ¿Qué es el error de muestreo?

Examen de autoevaluación

1. **A los valores numéricos obtenidos del análisis estadístico descriptivo de una muestra se les denomina:**

a. Población
b. Parámetros
c. Estadísticos
d. Sesgo
e. Desviación estándar

2. **Cuando se selecciona una muestra con el fin de realizar una análisis estadístico debe cuidarse que los elementos:**

a. Tengan características similares entre sí
b. Se encuentren dentro del mismo lote
c. Sean seleccionados de manera aleatoria
d. Sean lo más parecidos a la población
e. Estén lo más alejados del centro de la población

3. **Al proceso mediante el cual se obtienen los elementos de una muestra representativa de la población se le denomina:**

a. Proceso estadístico
b. Procedimiento de muestreo
c. Proceso de selección
d. Muestreo aleatorio
e. Seccionamiento

4. **Al obtener una muestra se debe asegurar que durante el proceso todos los elementos:**

a. Resulten del mismo tipo
b. Resulten como deseamos
c. Se encuentren del intervalo seleccionado
d. Resulten sin defectos
e. Tengan la misma probabilidad de ser escogidos

5. **Una técnica para muestrear, en la cual se asegura la no intervención de la mano del hombre, es:**

a. El uso de un dado
b. Una moneda
c. Una tabla de números aleatorios
d. El criterio del analista a cargo
e. El criterio del cliente

6. **Una población finita en la que se realiza un muestreo con reemplazamiento puede ser considerada como:**

a. Modelo
b. Infinita
c. Muestra
d. Acotada
e. Estratificada

7. **El muestreo realizado mediante la aplicación de un criterio personal de preferencia o aversión hacia determinados elementos constituye un método:**

a. Probabilístico
b. Aleatorio simple
c. Aleatorio directo
d. De conglomerados
e. No probabilístico

8. **Suponga que hay un inventario con 15 diferentes líneas de producto. Si para efectuar un muestreo tomamos una sola línea de producto se dice que el muestreo fue:**

a. Probabilístico
b. Por conglomerados
c. Aleatorio simple
d. Aleatorio sistemático

9. **Se denomina así a la diferencia entre un estadístico y su parámetro poblacional correspondiente:**

a. Media poblacional
b. Proporción
c. Error de muestreo
d. Parámetro poblacional
e. Sesgo

10. Un auditor va a realizar una prueba donde espera una tasa de error no mayor al 5%. Si fija una precisión de ± 3% y un nivel de confianza de 95% en una población de 15 000 facturas, si la prueba se realizará en el mes de marzo y si la última factura del mes de febrero es la No. 28 974, el tamaño de la muestra es de:

a. 15 000
b. 375
c. 7 500
d. 28 974
e. 1 500

Tema 2. Distribuciones muestrales y el teorema central del límite

Objetivo particular

El alumno analizará los conceptos fundamentales acerca de la distribución de muestreo, así como la aplicación de intervalos de confianza para la media poblacional.

Temario detallado

2. Distribuciones muestrales y el teorema central del límite

2.1 Distribuciones relacionadas con la normal: j^2, t y F. Propiedades y manejo de tablas.

2.2 Teorema Central del límite

2.3 Distribución muestral para la media

2.4 Distribución muestral para la proporción.

Introducción

La distribución de la población de la cual extraemos la muestra con la que trabajamos en estadística, es importante para saber que tipo de distribución debemos aplicar en cada una de las situaciones que se nos presenten en la práctica; en el presente tema veremos algunas de estas distribuciones que se encuentran relacionadas con la distribución normal, además de observar la distribución muestral para la media y para la proporción y su relación con el teorema central del límite.

2.1 Distribuciones relacionadas con la normal: j^2, t y F. Propiedades y manejo de tablas.

- **Distribución Chi-Cuadrado (j^2 o χ^2)**

En ocasiones los investigadores muestran más interés en la varianza poblacional que en la proporción o media poblacionales y las razones llegan desde el campo de la calidad total, donde la importancia en demostrar una disminución continua en la variabilidad de las piezas que la industria de la aviación llega a solicitar es de vital

importancia. Por ejemplo, el aterrizaje de un avión depende de una gran cantidad de variables, entre las que encontramos la velocidad y dirección del aire, el peso del avión, la pericia del piloto, la altitud, etc.; si en el caso de la altitud, los altímetros del avión tienen variaciones considerables, entonces podemos esperar con cierta probabilidad un aterrizaje algo abrupto, por lo tanto la variabilidad de estos altímetros debe mostrar un disminución continua; y que decir de los motores que impulsan al avión mismo, si las piezas que los conforman son demasiado grandes, el motor puede incluso no poder armarse y si son demasiado pequeñas, entonces los motores tendrán demasiada vibración y en ambos casos las perdidas de la industria son cuantiosas.

Así, la relación entre la **varianza de la muestra** y la **varianza de la población** está determinada por la distribución Chi-cuadrada (χ^2) siempre y cuando la población de la cual se toman los valores de la muestra se encuentre normalmente distribuida. Y aquí debemos tener especial cuidado, pues la distribución Chi-cuadrada es sumamente sensible a la suposición de que la población está normalmente distribuida y por ejemplo construir intervalos de confianza para estimar una varianza poblacional, puede que los resultado no sean correctos dependiendo de si la población no está normalmente distribuida.

La distribución Chi-cuadrada (χ^2) es la razón que existe entre la varianza de la muestra (s^2) multiplicada por los grados de libertad y la varianza de la población. Es decir:

$$\chi^2 = \frac{s^2(gl)}{\sigma^2}$$

El término grados de libertad[8] se refiere al número de observaciones independientes para una fuente de variación menos el número de parámetros independientes estimado al calcular la variación.

Para la distribución Chi-cuadrada (χ^2), los grados de libertad vienen dados por (n – 1), por lo tanto, la formula anterior quedaría expresada como:

$$\chi^2 = \frac{s^2(n-1)}{\sigma^2}$$

donde podemos observar que la variación de la distribución Chi-cuadrada (χ^2) depende del tamaño de la muestra y de los grados de libertad que posea.

En general y debido a que la distribución Chi-cuadrada (χ^2) no es simétrica a medida que se incrementa el número de grados de libertad, la curva característica de la distribución se vuelve menos sesgada.

La distribución Chi-cuadrada (χ^2), es en sí toda una familia de distribuciones por lo que, existe una distribución Chi-cuadrado para cada grado de libertad.

Algebraicamente podemos manipular la formula anterior $\chi^2 = \frac{s^2(n-1)}{\sigma^2}$ con el objetivo de que nos sea de utilidad para construir intervalos de confianza para varianzas poblacionales, quedando de la siguiente manera:

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$$

Veamos un ejemplo de cómo se utiliza esta formula:

Suponga que una muestra de 7 pernos especiales utilizados en el ensamblado de computadoras portátiles arrojo los siguientes resultados:

2.10 mm; 2.00 mm, 1.90 mm, 1.97 mm, 1.98 mm, 2.01 mm, 2.05 mm

[8] Ken, Black. "*Estadística en los negocios*", editorial CECSA, pp. 264

Si quisiéramos una estimación puntual de la varianza de la población, sería suficiente con calcular la varianza de la muestra, de la siguiente manera:

Primero calculamos la **media aritmética** de los datos utilizando la siguiente formula:

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

por lo tanto sustituyendo datos tenemos que:

$$\bar{X} = \frac{2.10 + 1.90 + 1.98 + 2.05 + 2.00 + 1.97 + 2.01}{7}$$

y al efectuar cálculos el resultado de la media aritmética (redondeado a 2 decimales) es de:

$$\bar{X} = 2.00$$

a continuación elaboramos una tabla como la indicada a continuación para facilitar el calculo de la varianza de los datos:

i-dato I	**DATOS** x_i	**Dato-media** $(x_i - \mu)$	**(Dato - media) elevado al cuadrado** $(x_i - \mu)^2$
1	2,10	0,10	0,00972
2	1,90	-0,10	0,01029
3	1,98	-0,02	0,00046
4	2,05	0,05	0,00236
5	2,00	0,00	0,00000
6	1,97	-0,03	0,00099
7	2,01	0,01	0,00007
	14,01	0,01	**0,02389**

Recordando ahora la formula correspondiente a la varianza de una muestra:

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

y sustituyendo datos en esta formula, podemos ver que el valor obtenido en la

esquina inferior derecha de la tabla anterior corresponde a: $\sum_{i=1}^{n}(X_i - \bar{X})^2$ por lo tanto:

$$s^2 = \frac{1}{7-1}(0.02389)$$

de donde al efectuar cálculos vemos que:

$$s^2 = 0.003981$$

Es decir, la varianza de la muestra tiene un valor de: 0.003981, pero si consideramos que el valor de la **estimación puntual** puede cambiar de una muestra a otra, entonces será mejor construir un intervalo de confianza, para lo cual debemos suponer que la población de los diámetros de los pernos esta normalmente distribuida, y como vemos que n=7 entonces los grados de libertad serán: gl=7-1=6, si queremos que el intervalo sea del 90% de confianza, entonces el nivel de significancia α será de 0.10 siendo esta la parte del área bajo la curva de la distribución Chi-cuadrada que está fuera del intervalo de confianza, esta área es importante porque los valores de la tabla de distribución Chi-cuadrada están dados de acuerdo con el área de la cola derecha de la distribución. Además en nuestro caso $\alpha/2$ = 0.05 es decir, 0.05 del área está en la cola derecha y 0.05 está en la cola izquierda de la distribución.

Es importante hacer notar que debido a la forma de curva de la distribución Chi-cuadrada, el valor para ambas colas será diferente, así, el primer valor que se debe de obtener es el de la cola derecha, mismo que se obtiene al ubicar en el primer renglón de la tabla el valor correspondiente al nivel de significancia, que en este caso es de 0.05 y, posteriormente se ubica en el lugar de las columnas los correspondientes grados de libertad ya calculado, que en este caso es de 6 grados de libertad, por lo tanto el valor de Chi-cuadrada obtenido es de:

$$\chi^2_{0.05,6} = 12.5916$$

observe que en la nomenclatura se escribe la denotación de Chi-cuadrada teniendo como subíndice el nivel de significancia y los grados de libertad y, a continuación se escribe el valor correspondiente.[9]

El valor de Chi-cuadrada para la cola izquierda se obtiene al calcular el área que se encuentra a la derecha de la cola izquierda, entonces:

$\mathcal{A}$ a la derecha de la cola izquierda = 1 – 0.05

$\mathcal{A}$ a la derecha de la cola izquierda = 0.95

por lo tanto, el valor de Chi-cuadrada para la cola izquierda será, utilizando el mismo procedimiento anterior para un área de 0.95 y 6 grados de libertad, de:

$$\chi^2_{0.95,6} = 1.63538$$

incorporando estos valores a la formula, tenemos que el intervalo de 90% de confianza para los 7 pernos utilizados en el ensamblado de computadoras portátiles tendrá la forma mostrada a continuación:

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$$

$$\frac{0.0034122(7-1)}{12.5916} \leq \sigma^2 \leq \frac{0.0034122(7-1)}{1.63538}$$

$$0.0001625 \leq \sigma^2 \leq 0.0125189$$

este intervalo de confianza nos dice que con 90% de confianza, la varianza de la población está entre 0.0001625 y 0.0125189.

[9] Nota: el valor se obtuvo utilizando la tabla correspondiente a la Chi-cuadrada en el libro: "*Estadística en los negocios*" del autor: Ken Black, pp 779.

- **Distribución "t"**[10]

Cuando las muestras se toman de una población normal, la distribución muestral de la media es normal, sin embargo, si la desviación estándar de la población es desconocida, no podemos transformar la media muestral en un puntaje estándar.

En muchas situaciones prácticas la desviación estándar poblacional es desconocida, y se usa la desviación estándar muestral para estimar σ, en consecuencia, el estadístico siguiente no tiene la distribución muestral normal estándar:

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

este estadístico se denota por "t" y se denomina el **estadístico t**. Así, el estadístico "t" esta dado por la fórmula:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

En 1908, W. Gosset, un dirigente judío de una planta cervecera, publicó un artículo de investigación relativo a la ecuación para la distribución de probabilidad de "t", como los empleados de la planta cervecera no tenían permitido publicar los resultados de sus investigaciones, Gosset publicó sus resultados firmándolos bajo el nombre de student; desde entonces, la distribución muestral del estadístico "t" se conoce como la **distribución "t" de student**, o simplemente la **distribución t**[11]**.**

La distribución muestral de "**t**" es parecida a la distribución normal; ambas tienen formas acampanadas, media igual a cero y son simétricas respecto a sus medias. La distribución muestral de "t" es más variable que la normal estándar. Para el estadístico z, $\bar{X}$ es la única cantidad que varía de muestra a muestra, mientras que para "t" tanto $\bar{X}$ como "s" lo hacen.

[10] Weimer, Richard, C. "*Estadística*". Editorial: CECSA. pp 373-375
[11] Weimer, Richard, C. "*Estadística*". Editorial: CECSA. pp 374

La forma exacta de una distribución "**t**" está especificada completamente por un único valor, parámetro conocido como el: número de grados de libertad (**gl**); el tamaño de la muestra "n" se relaciona con "gl" por:

gl = n – 1

La formula anterior se debe a que normalmente se considera como parámetro independiente a la media poblacional **μ ,** misma que se estima con $\bar{X}$ al calcular "s" por lo tanto, la formula para los grados de libertad será igual a "n" observaciones independientes menos un parámetro independiente al ser estimada la variación.

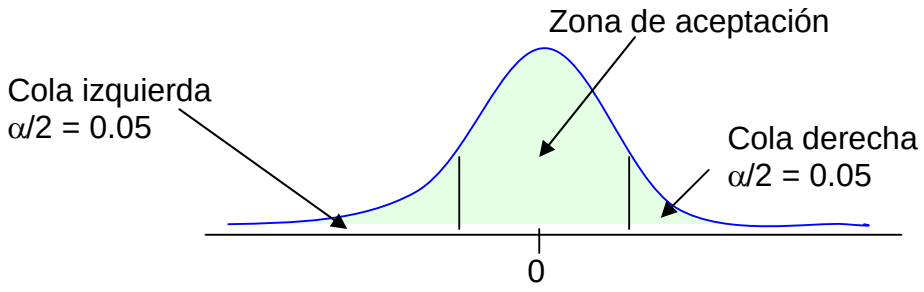
Las distribuciones muestrales "t" tienen las propiedades siguientes:

1. Media cero
2. Son simétricas respecto a **μ = 0**
3. Son más variables que la distribución normal estándar
4. Forma acampanada
5. Su forma exacta depende de **gl = n – 1**
6. Sus varianzas dependen de: gl y $\sigma^2 = \frac{gl}{gl-2}$ si gl>2
7. Cuando "**n**" crece, la distribución muestral de "**t**" se aproxima a la distribución normal estándar "**z**"
8. Como las distribuciones muestrales de "t" son más variables que la distribución normal estándar, tienen las áreas de las colas más grandes que la distribución normal estándar.

En las tablas de la **distribución "t"** los grados de libertad están en la primera columna (de izquierda a derecha), sin embargo hay que hacer notar que para esta distribución, la tabla no utiliza el área entre el estadístico y la media como lo hace la distribución normal estándar, sino más bien utiliza el área de la cola de la distribución, así, la relevancia de la tabla se encuentra en el nivel de significancia α y cada cola de la distribución contiene $\alpha/2$ del área bajo la curva cuando se construyen

intervalos de confianza. Es decir, la construir intervalos de confianza, el valor del estadístico “t” se encuentra en la tabla, en la intersección de la columna bajo el valor de $\alpha/2$ y el renglón del valor de grados de libertad (gl).

Así por ejemplo, si calculamos un intervalo de confianza de 90%, el área total de las dos colas será de 10% y $\alpha/2$ será de 0.05, es decir:



por lo tanto si tuviéramos 10 grados de libertad, entonces la intersección de $\alpha/2 = 0.05$ y gl = 10 nos arroja un valor de[12] t = 1.812.

- **Distribución F**

La distribución F es la distribución de pares repetidos calculados de la razón que existe entre las varianzas de dos muestras extraídas de la misma población (también puede darse el caso que las dos muestras sean extraídas de poblaciones diferentes siempre y cuando las dos poblaciones tengan el mismo valor de la varianza).

Las aplicaciones principales de la distribución F se encuentran también en el control de calidad, donde resulta importante comparar las variabilidades o varianzas de dos maquinas diferentes que fabrican el mismo producto, con el objetivo de analizar primero si existe diferencia en la variabilidad de las maquinas y después en caso de existir, las razones por las cuales una maquina llega a tener más variabilidad que otra.

[12] Los valores de la tabla pueden variar por algunas décimas dependiendo del autor del libro, sin embargo estos valores siempre serán muy próximos.

- **Valor F**

El valor F es la razón que existe entre las varianzas de dos muestras extraídas de la misma población; es decir:

$$F = \frac{S_1^2}{S_2^2}$$

Esta razón estrictamente hablando debería ser muy próxima a la unidad, sin embargo, debido al error de muestreo algunas veces estas varianzas son diferentes.

La distribución F no es simétrica y tiene asociados grados de libertad tanto con el numerador como con el denominador de la razón anterior. El punto de partida para la aplicación de la distribución F es el supuesto de que la población o poblaciones de donde se extrajeron las muestras a analizar, están normalmente distribuidas.

La formula a utilizar en pruebas de hipótesis que comparan dos varianzas poblacionales es:

$$F = \frac{S_1^2}{S_2^2}$$

$$v_1 = gl_{de\,\mathrm{ln}umerador} = n_1 - 1$$

$$v_2 = gl_{del deno\,\mathrm{min}\,ador} = n_2 - 1$$

Las tablas de la distribución F contienen valores para α= 0.10, 0.05, 0.025, 0.01, 0.005 y para diferentes grados de libertad tanto del numerador como del denominador. Además, estos valores están calculados para la cola superior de la curva y como la razón F siempre es positiva, el problema de asignar valores críticos a la cola inferior se resuelve utilizando la siguiente formula:

$$F_{1-\alpha,v_2,v_1} = \frac{1}{F_{\alpha,v_1,v_2}}$$

esta formula nos indica que el valor critico de F para la cola inferior (1-α) se encuentra al tomar el inverso multiplicativo del valor de F para la cola superior (α), teniendo cuidado en respetar los grados de libertar tanto del numerador como del denominador del valor F.

Ejemplo: suponga usted que dos maquinas fabrican el mismo producto, de tornillos que deben medir 20 mm de diámetro y el dueño de la fabrica, preocupado por la variabilidad de ambas maquinas ha solicitado un estudio en el que se muestrean al azar 10 tornillos fabricados por la maquina 1 y 12 tornillos fabricados por la maquina 2 y los resultados se presentan en la siguiente tabla:

Maquina 1	Maquina 2
21.3	21.8
22.1	22.3
20.8	20.9
20.5	22.7
20.6	21.4
21.6	22.0
20.4	21.9
22.1	21.5
21.7	22.9
22.4	20.8
	21.2
	22.4

Si el diámetro de los tornillos está normalmente distribuido, podemos aplicar una prueba de hipótesis para determinar si las varianzas de ambas maquinas son iguales o no lo son.

Resolviendo el problema, primero planteamos nuestras hipótesis opuestas, y en este caso serían:

$$H_0 : \sigma^2_1 = \sigma^2_2 \qquad \text{y} \qquad H_1 : \sigma^2_1 \neq \sigma^2_2$$

aquí, podemos observar que de acuerdo con el signo de igualdad incluido en la hipótesis nula, se trata de una prueba de dos colas.

El estadístico de prueba a utilizar es:

$$F = \frac{S_1^2}{S_2^2}$$

si utilizamos un nivel de significancia de α=0.05, como estamos realizando un prueba de dos colas entonces: $\frac{\alpha}{2} = 0.025$ y teniendo en cuenta que el tamaño de la muestra de la maquina 1 es de $n_1 = 10$ y el tamaño de la muestra de la maquina 2 es $n_2 = 12$, entonces el numero de grados de libertad para el valor crítico de la cola superior es:

$$v_1 = n_1 - 1$$
$$v_1 = 10 - 1$$
$$v_1 = 9$$

y en el denominador, el numero de grados de libertad para el valor critico de la cola inferior es de:

$$v_2 = n_2 - 1$$
$$v_2 = 12 - 1$$
$$v_2 = 11$$

por lo tanto, el valor crítico de F para la cola superior obtenido de la tabla es:

$$F_{1-\alpha,v_1,v_2} = F_{0.025,9,11} = 3.59$$

claro esta que este valor lo obtuvimos de la tabla de distribución F teniendo cuidado en buscarlo en que corresponde a $\alpha = 0.025$, el valor se encuentra en la intersección de los grados de libertad del numerador (9) con los grados de libertad del denominador (11).

Y el valor crítico de la cola inferior lo calculamos desde el valor de la cola superior utilizando la formula:

$$F_{1-\alpha,v_2,v_1} = \frac{1}{F_{\alpha,v_1,v_2}}$$

$$F_{0.975,11,9} = \frac{1}{F_{0.025,9,11}}$$

$$F_{0.975,11,9} = \frac{1}{3.59}$$

$$F_{0.975,11,9} = 0.28$$

Entonces, la regla de decisión es: rechazar la hipótesis nula si el valor de F que se observa es mayor a 3.59 o menor a 0.28

Si efectuamos lo cálculos para las varianzas tendríamos que para la maquina 1 la varianza es de: $s_1^2 = 0.545$ y para la maquina dos, la varianza es: $s_2^2 = 0.46333333$ por lo tanto el valor de F es de:

$$F = \frac{S_1^2}{S_2^2}$$

$$F = \frac{0.545}{0.46333333}$$

$$F = 1.1762$$

este valor de la razón de las varianzas muestrales 1.1762 cae dentro de la zona de aceptación que nos indica la regla de decisión, por lo que: como resultado del estudio aceptamos tentativamente la hipótesis nula, es decir: las varianzas de las dos muestras son iguales.

2.2 Teorema central del límite[13]

El enunciado formal del teorema del límite central es el siguiente: si en cualquier población se seleccionan muestras de un tamaño específico, la distribución muestral de las medias de muestras es aproximadamente una distribución normal. Esta aproximación mejora con muestras de mayor tamaño.

Ésta es una de las conclusiones más útiles en estadística pues nos permite 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, de acuerdo con el teorema del límite central, es válido aproximar la distribución de probabilidad normal a cualquier distribución de valores medios muestrales, siempre y cuando se trate de una muestra suficientemente grande.

El teorema central del límite o teorema del límite central se aplica a la distribución muestral de las medias de muestras que veremos a continuación y permite utilizar la distribución de probabilidad normal para crear **intervalos de confianza** (que se verán en el apartado 3.4) para la media de la población.

[13] Douglas A. Lind., et al. "*Estadística para administración y economía*" p.p 234

2.3 Distribución muestral para la media

Si 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 muestral para la media** 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.

Los resultados que nos da una muestra para estimar el parámetro de una población se utilizan (en aplicaciones avanzadas de la estadística) cuando se quiere saber lo siguiente:

- Hacer una predicción precisa sobre el éxito de algún producto de reciente desarrollo sólo con base en los resultados de la muestra.
- ¿Cómo puede el departamento de control de calidad de una empresa maquiladora liberar un embarque de un producto determinado con base en una muestra de sólo unas cuantas unidades?
- ¿Cómo puede “Encuestas Mitovsky” hacer una predicción precisa de una votación presidencial con base en una muestra de sólo una muestra de los votantes registrados que proceden de una población de alrededor de 100 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 para la media o distribución muestral de las medias de las muestras.

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

Veamos un ejemplo sencillo, que si bien es cierto que no responde a las preguntas tan complejas del inicio del tema, si ayuda a entender el concepto y la importancia de la distribución muestral para la media.

Ejemplo[14]: 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.

[14] Problema tomado con ligeros cambios del libro: "*Probabilidad y Estadística*" de Stephen S. Willoughby. p.p 126

Solución al problema propuesto:

a) Para encontrar la **media aritmética**, 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:

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

en donde al sustituir los datos tenemos que:

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

solución al a) $\mu = 74$

b) Para este inciso 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:

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

respuesta al b) σ **= 4.9**

c) Para dar respuesta a este inciso, debemos aplicar la fórmula correspondiente al cálculo de combinaciones; es decir:

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

en donde sustituyendo los valores tenemos que:

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

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

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

respuesta al c) $C_r^n = 20$

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 la siguiente manera: $80_1, 80_2, 76, 70_1, 70_2, 68$.

2. 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, nos saltamos el tercer dato; por lo tanto, 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:

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:

Entonces, la terna sería: 80_1, 76, 70_1

Siguiendo este procedimiento, podemos encontrar fácilmente las siguientes ternas; es importante considerar que los datos son**:** 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:

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 tres 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:

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

donde sustituyendo los datos tenemos que:

respuesta al f) $\mu_{\bar{x}}$ **= 74**

g) Para calcular la desviación estándar de las medias de las muestras, es necesario elaborar la siguiente tabla:

M U E S T R A S	**Promedio de la muestra (Datos)**	**Media aritmética de la distribución de las muestras:**	**Dato-media**	**(Dato - media) elevado al cuadrado**
I	x_i	μ	$(x_i - \mu)$	$(x_i - \mu)^2$
1	**78 2/3**	74	4 2/3	21 7/9
2	**76 2/3**	74	2 2/3	7 1/9
3	**76 2/3**	74	2 2/3	7 1/9
4	**76**	74	2	4
5	**75 1/3**	74	1 1/3	1 7/9
6	**75 1/3**	74	1 1/3	1 7/9
7	**74 2/3**	74	2/3	4/9
8	**73 1/3**	74	- 2/3	4/9
9	**72 2/3**	74	-1 1/3	1 7/9
10	**72 2/3**	74	-1 1/3	1 7/9
11	**75 1/3**	74	1 1/3	1 7/9
12	**75 1/3**	74	1 1/3	1 7/9
13	**74 2/3**	74	2/3	4/9
14	**73 1/3**	74	- 2/3	4/9
15	**72 2/3**	74	-1 1/3	1 7/9
16	**72 2/3**	74	-1 1/3	1 7/9
17	**72**	74	-2	4
18	**71 1/3**	74	-2 2/3	7 1/9

19	**71 1/3**	74	-2 2/3	7 1/9
20	**69 1/3**	74	-4 2/3	21 7/9
Sumatoria	1480		0	96

Para efectuar este cálculo, lo primero que hacemos es escribir la formula correspondiente, que en este caso quedaría de la siguiente forma:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{N}\sum_{1}^{N}(\bar{x}_i - \mu_{\bar{x}})^2}$$

A continuación sustituimos los datos correspondientes

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{20}(96)}$$

solución al g) $\sigma_{\bar{x}} = 2.19$

Como podemos observar, el valor de la desviación estándar de las medias de las muestras es de $\sigma_{\bar{x}} = 2.19$

h) Compare los resultados de los incisos a y f

En el inciso a calculamos el valor de la media aritmética de la población, obteniendo un valor de $\mu = 74$ mientras que en el inciso f calculamos el valor de la media de la distribución de las medias de las muestras, para la encontramos un valor de $\mu_{\bar{x}}$= **74**, con lo cual podemos concluir que la media de la población y la media de la distribución de las medias tienen el mismo valor.

i) Compare los resultados de los incisos b y g.

En el inciso b determinamos la desviación estándar de la población, obteniendo un valor de σ= **4.9** mientras que en el inciso g encontramos que el valor de la desviación estándar de las medias de las muestras fue de $\sigma_{\bar{x}} = 2.19$ con lo cual

podemos decir que el valor de la desviación estándar de la población y el de la desviación estándar de las medias de las muestras son diferentes.

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

- 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 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:

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

donde $\sqrt{\frac{N-n}{N-1}}$ se conoce como factor de población finita y se utiliza cuando el tamaño de la muestra es mayor al 5% del tamaño de la población. Esto es debido a que los resultados obtenidos con un muestreo con y sin reemplazo son distintos. Esto ocurre porque las probabilidades cambian significativamente cuando se trabaja con muestras pequeñas. Para considerar esta situación en los análisis con distribuciones muestrales es necesario corregir el error estándar de manera que refleje el cambio que pueden tener las probabilidades.

Si en el ejercicio anterior del obrero que fabrica latas para refresco se calcula la desviación estándar de las medias de las muestras $\sigma_{\bar{x}}$ mediante la formula: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ se obtiene exactamente el mismo resultado de $\sigma_{\bar{x}} = 2.19$. (se deja al estudiante que realice la comprobación).

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

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

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 (siempre y cuando 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).

2.4 Distribución muestral de la proporción

Hoy es bien sabido[15] que si la investigación produce datos mensurables tales como el peso, distancia, tiempo e ingreso, la media muestral es en ocasiones el estadístico más utilizado, pero, si la investigación resulta en artículos “contables” como por ejemplo: cuántas personas de una muestra escogen la marca “Peñafiel” como su refresco, o cuantas personas de una muestra tienen un horario flexible de trabajo, la proporción muestral es generalmente el mejor estadístico a utilizar.

Mientras que la media se calcula al promediar un conjunto de valores, la “**proporción muestral**” se calcula al dividir la frecuencia con la cual una característica dada se presenta en una muestra entre el número de elementos de la muestra. Es decir:

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

donde: x = número de elementos de una muestra que tienen la característica.

n = numero de elementos de la muestra.

[15] Black, Ken. “*Estadística en los negocios*” pp. 241-242

Ejemplo; suponga que una comercializadora pretende establecer un nuevo centro y desea saber la proporción del consumidor potencial que compraría el principal producto que vende para lo cual realiza un estudio de mercado mediante una encuesta a 30 participantes, lo cual permitirá saber quiénes lo comprarían y quiénes no; se obtuvieron 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 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 con el 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:

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

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

Dado que todas las muestras aleatorias que sean tomadas de una misma población en general serán distintas y tendrán por ende diferentes valores para sus estadísticos tales como la media aritmética o la desviación estándar, entonces resulta importante estudiar la distribución de todos los valores posibles de un estadístico, lo cual significa estudiar las distribuciones muestrales para diferentes estadísticos[16]. La importancia de éstas distribuciones muestrales radica en el hecho de que en estadística inferencial, las inferencias sobre poblaciones se hacen utilizando estadísticas muestrales pues con el análisis de las distribuciones asociadas con éstos estadísticos se da la confiabilidad del estadístico muestral como instrumento para hacer inferencias sobre un parámetro poblacional desconocido.

Bibliografía del tema 2

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill.

[16] Weimer, Richard, C. "*Estadística*". pp 353

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica.
WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.2.1. Elabora un glosario con los conceptos básicos de “distribución muestral para la media” “distribución muestral de la proporción” y “teorema central de límite” en los libros de la bibliografía del tema.

A.2.2. Realiza los ejercicios sobre este tema que se presentan en el libro de *Probabilidad y Estadística* de Stephen S. Willoughby.

A.2.3. Investiga y elabora un cuadro con las ventajas y desventajas de la distribución muestral para la media y de la distribución muestral de la proporción.

A.2.4. Investiga las aplicaciones prácticas para distribución muestral para la media.

A.2.5. Investiga en que situaciones se utiliza la distribución t de Student en lugar de una distribución normal.

A.2.6. Revisa el uso y manejo de las tablas para las distribuciones: t de Student, “F” y normal estándar, en base a los ejercicios que se presentan en los libros citados en la bibliografía del tema.

Cuestionario de evaluación

1. ¿Qué es una distribución de muestreo?

2. Si el estadístico utilizado es la media muestral, ¿qué nombre recibe la distribución de este estadístico?

3. ¿Qué es la distribución muestral de las medias de las muestras?

4. ¿Qué relación existe entre la media de las medias de la muestra y la media de la población?

5. ¿Cómo es la dispersión de las medias de la muestra en comparación con la de los valores de la población?

6. ¿Cómo es 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?

7. ¿Cómo es la desviación estándar de las medias de las muestras comparada con

la desviación estándar de la población?

8. Para una población infinita ¿qué implicación tiene el hecho de que la distribución de muestreo sea asintóticamente normal?
9. ¿Cómo es la distribución de muestreo de medias cuando la población de origen está normalmente distribuida?

10. En una empresa se tienen 4 puestos de gerente nivel C disponibles y 7 candidatos que pueden ocupar esos puestos, ¿de cuántas formas podemos tomar la decisión correspondiente?

Examen de autoevaluación

1. **Al considerar todas las muestras de tamaño "n" que pueden extraerse de una población, si se calcula el valor medio para cada una de ellas y se integran estos valores en un solo conjunto de datos es posible obtener una:**
   a. Campana de Gauss
   b. Tendencia paramétrica
   c. Curva de ajuste
   d. Distribución muestral
   e. Parámetro muestral

2. **En el proceso de inferencia estadística paramétrica existen dos maneras de estimar los parámetros de una población, una de ellas es la:**
   a. Estadística descriptiva
   b. Estimación puntual
   c. Prueba de significancia
   d. Medida de sesgo
   e. Medida de tendencia central

3. **Calcular el factor de corrección para la población finita de un inventario que consta de 250 productos y a la cual se le efectuará un muestreo de 40%:**
   a. 0.881
   b. 0.918
   c. 0.819
   d. 0.991
   e. 0.989

4. **Qué concepto establece que si se selecciona una muestra aleatoria suficientemente grande de n observaciones, la distribución muestral de las medias de las muestras se aproxima a una distribución normal.**
   a. Definición de distribución muestral
   b. Proceso aleatorio
   c. Proceso de muestreo
   d. Teorema del límite central
   e. Distribución de probabilidad

5. **Si una población se distribuye normalmente (con media μ y desviación estándar σ), la distribución muestral de las medias construida a partir de la misma población también se distribuye normalmente. Esta definición corresponde a:**
   a. El teorema de Bayes
   b. La ley de las probabilidades
   c. El teorema del límite central
   d. La ley de la distribución normal
   e. El teorema de Markov

6. **Una población se compone de los siguientes cinco números 2, 3, 6, 8, y 11. Calcule la media de la distribución muestral para tamaños de muestra 2 con reemplazamiento:**
   a. 6.2
   b. 5.7
   c. 6.0
   d. 6.1
   e. 5.8

7. **Cuando se lleva a cabo un estudio estadístico paramétrico se requiere una muestra suficientemente grande, lo cual significa que debe tener un tamaño igual o mayor a:**
   a. 64
   b. 50
   c. 40
   d. 30
   e. 20

8. **Si las distribuciones muestrales tienen la misma media, la elección de una de ellas deberá entonces basarse en la que tenga el menor valor del estadístico. Esta definición corresponde a:**
   a) Rango
   b) Varianza
   c) Sesgo
   d) Mediana
   e) Moda

9. **Se tiene una lista de 120 estudiantes, 60 de ellos son de Contaduría y el resto de Administración. Si se toma una muestra al azar, halle la probabilidad de que se escojan entre el 40% y el 60% de contadores del tamaño de la muestra:**

   a. 98.5%
   b. 96.7%
   c. 95.8%
   d. 97.7%
   e. 99.1%

10. **De un lote muy grande (población infinita) de facturas, la desviación estándar es $10. Se extraen diversas muestras; cada una de ellas es de 200 facturas y se calculan las desviaciones estándar de cada muestra. Hallar la media de la distribución muestral de desviaciones estándar:**

    a. 0.30
    b. 0.50
    c. 2.77
    d. 7.41
    e. 10.0

Tema 3. Estimación de parámetros e intervalos de confianza

Objetivo particular

El alumno analizará los conceptos fundamentales de estimación de parámetros e intervalos de confianza y los aplicará en la práctica a problemas de su área profesional laborable.

Temario detallado

3. Estimación de parámetros e intervalos de confianza

- 3.1 Definición de estimador y estimación
- 3.2 Propiedades de los estimadores
- 3.3 Estimación de media, varianza y proporciones
- 3.4 Intervalo de confianza para la media y para proporciones
- 3.5 Determinación del tamaño de la muestra

Introducción

En el momento de tomar decisiones el conocimiento de los parámetros de población es de vital importancia, tal conocimiento generalmente solo se puede tener al estimar el valor de dichos parámetros, sin embargo, la estimación es mejor cuando se da un margen de confianza y uno de error, siendo de vital importancia la correcta estimación de dichos parámetros a través de la construcción de intervalos de confianza que puedan sustentar la toma de decisiones de manera eficiente.

En el momento de tomar decisiones, es de vital importancia tener el conocimiento de los parámetros de población aunque éstos solo pueden ser estimados sus valores, sin embargo, la estimación es mejor cuando se tiene un margen de confianza y uno de error, para ello se debe tener una correcta estimación de los parámetros por medio de la construcción de intervalos de confianza que puedan sustentar la toma de decisiones de manera más eficiente.

3.1 Definición de estimador y estimación

Para realizar un análisis requerimos de una definición técnica. 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 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.

Se llama **estimador puntual**[17] al número (punto sobre la recta real o recta de los números reales), 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, es un valor que se calcula a partir de la información de la muestra, y que se usa para estimar el parámetro de la población.

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 sólo podemos definir buenos procesos (estimadores), mas no garantizar buenos resultados (estimaciones).

Por ejemplo, la media muestral $\bar{x}$ 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 (es decir, puede darse el caso en el que la mediana muestral esté más próxima a la media poblacional que la media muestral). Así, lo más que podemos hacer es encontrar estimadores que den buenos resultados en el límite.

[17] Erwin. Kreyszig, *Matemáticas avanzadas para ingeniería*, vol. 2, p. 958.

Como una aproximación[18] de la media **μ** de una población, puede tomarse la media $\bar{x}$ de una muestra correspondiente, lo cual da la estimación: **$\hat{u}=\bar{x}$,** para **μ**, es decir:

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

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:

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

Evidentemente, (1) y (2) son estimaciones de los parámetros para distribuciones en las que tanto la media como 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 la que los parámetros que van a estimarse se expresan en términos de los momentos de la distribución[19] en las fórmulas resultantes; esos momentos se reemplazan por los momentos correspondientes de la muestra, lo cual proporciona las estimaciones deseadas. Aquí, el k-ésimo momento de una muestra x_1, $x_{2,}$...x_n, es:

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

[18] Kreyszig. Erwin. "*Matemáticas avanzadas para ingeniería vol. 2"*. pp 958
[19] Para mayor información consulte la sección 19.8 del libro: "*Matemáticas avanzadas para ingeniería Vol. 2.*" de Erwin Kreyszig.

3.2 Propiedades de los estimadores

- **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 o esperanza matemática 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 el estimador es sesgado, la magnitud del sesgo está dada por la siguiente fórmula:

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

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**; en caso contrario, 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 esperanza matemática, 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. De acuerdo con la teoría estadística clásica, en términos generales se debe preferir el estimador insesgado más eficiente sobre cualquier otro. Más adelante veremos que las **hipótesis** nos dicen cuál 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.

Por 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.

De todos los estadísticos que estiman la media de la población, la media muestral proporciona la mejor (la más eficiente) estimación.

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. Por lo anterior, surge un tema de investigación en la teoría estadística: el de los llamados **estimadores robustos**, estadísticos casi insesgados y casi eficientes para una gran variedad de distribuciones poblacionales.

- **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. De acuerdo con la definición, la media muestral $\mu_{\bar{x}}$ es consistente.

Un estimador inconsistente es evidentemente un mal estimador y no es aconsejable dar una estimación imprecisa basada en una infinidad de datos, lo cual puede suceder si el sesgo de un estimador se aproxima a cero a medida que “n” tiende a infinito. Por ejemplo, utilizar el percentil 25 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.

- **Estimaciones de 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[20], 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 para este propósito.

3.3 Estimación de media, varianza y proporciones

- **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.

Por ejemplo: si decimos que una distancia se ha medido como 5.28 metros (m), estamos dando una **estimación de punto**. Por otra parte, si decimos que la distancia es 5.28±0.03 m (o sea, que esta entre 5.25 y 5.31 m) estamos dando una **estimación de intervalo.**

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

[20] Douglas A. Lind *et al*, *Estadística para administración y economía*, pp. 242.

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”.

Así, por ejemplo:

- A una empresa le puede interesar el número promedio de piezas defectuosas producidas por una cierta máquina.
- A un ingeniero especialista en vehículos 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 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, es decir, de 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: se escoge una muestra aleatoria de 36 recién egresados en la carrera de contaduría de cierta universidad; 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.

El procedimiento anterior es satisfactorio tal como se ha presentado. El problema estriba en el contenido de las palabras alrededor de y aproximadamente.

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 $\sigma_{\bar{x}}$.

Puesto que la distribución de las medias de las muestras es aproximadamente normal, se puede utilizar de manera efectiva el conocimiento sobre este tipo de distribución.

3.4 Intervalo de confianza para la media y para proporciones

Una estimación de un parámetro de la población dada por un solo número se llama una **estimación de punto del parámetro**. No obstante[21], 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 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 puntuales.

Por ejemplo: si decimos que el porcentaje de productos defectuosos que produce una máquina es del 6%, entonces el nivel se ha medido en 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**.

[21] Douglas A. Lind *et al.*, *Estadística para administración y economía*, pp. 242.

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

- **Intervalo para estimar la media**

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

$$-1 < z < +1$$

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 **z** 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:

$\sigma_{\bar{x}}$ por $\frac{s}{\sqrt{n}}$ y $\mu_{\bar{x}}$ por μ_x

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 que con una probabilidad de 0.68 que μ_x se encuentre dentro del intervalo:

(69– 0.58, 69+0.58)

es decir: $68.42 < \hat{\mu}_x < 69.58$ aquí, la media aritmética de la población lleva un acento circunflejo debido a que se trata de una estimación.

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 pueden construir otros intervalos con porcentajes de confianza que sean más útiles.

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

$$-k < z < +k$$ **----------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 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 "1.96". Es decir, el área bajo la curva

normal estándar entre –1.96 y +1.96 es 0.9544, o sea, aproximadamente 0.95. Así, la probabilidad de que **z** se encuentre dentro del intervalo:

(-1.96, +1.96)

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

-1.96 <z<+1.96

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 **z** 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, se obtiene:

$$-1.96 < \frac{\bar{X}-\mu_x}{\frac{s}{\sqrt{n}}} < +1.96$$

Resolviendo estas desigualdades para μ , se tiene que:

$$\bar{X}-\frac{1.96s}{\sqrt{n}} < \mu_x < \bar{X}+\frac{1.96s}{\sqrt{n}} \text{ ----------2}$$

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:

$$\bar{X}-\frac{1.96s}{\sqrt{n}} \qquad \textbf{y} \qquad \bar{X}+\frac{1.96s}{\sqrt{n}}$$

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 tamaño de muestra 36) en 2 se tiene que el intervalo con 95% de confianza es:

$$69 - \frac{1.96(3.5)}{\sqrt{36}} < \mu_x < 69 + \frac{1.96(3.5)}{\sqrt{36}}$$

$$67.8 < \mu_x < 70.1$$

(67.8 , 70.1)

- **Intervalo para estimar la varianza**

En el apartado 3.2 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; 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 con lo estudiado en el tema 2 tiene una distribución Chi-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 una máquina que llena envases con capacidad de 500 ml.; para ello, se extrae una muestra periódicamente; si la muestra indica que hay una variación de ±5 ml. alrededor de los 500 y 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 ello, 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 con la estimación por intervalos de confianza, se tendrá que:

$$X^2_{0.025} < \frac{(n-1)S^2}{\sigma^2} < X^2_{0.975}$$

El resultado anterior de acuerdo con tablas de Chi-cuadrada con 40 grados de libertad $X^2_{0.025}$=24.433 y $X^2_{0.09750}$ = 59.342.
(Recuerda que el uso de las tablas y de los grados de libertad se encuentra en el apartado 3.2)

Entonces el intervalo es:

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

Sustituyendo los resultados de la muestra se tiene:

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

Al obtener inversos multiplicativos tenemos:

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

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

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

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

Obteniendo raíz cuadrada, se tiene:

$$4.555 > \sigma > 2.92$$

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

- **Intervalo para estimar la proporción**

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

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

El cual, de acuerdo con el 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; 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 observó que la proporción que bajo de precio son el 0.10 (10%). En el día se estima que no se presenten 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 con 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 con tablas 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:

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

$$que\ equivale\ a:$$

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

$$y\ despejando\ P\ se\ tiene:$$

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

$$igual\ a:$$

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

$$Por\ lo\ cual\ el\ \text{int}ervalo\ es:$$

$$0.169 > P > 0.0304$$

Es decir aproximadamente entre el 3% y 17%.

3.5 Determinación del tamaño de la muestra

- **Tamaño de muestra para la media**

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

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

Lo anterior representa el número de desviaciones estándar alrededor de la media μ dado el nivel de confianza 1-α, por lo que 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, 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:

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

La fórmula anterior representa el número de desviaciones estándar alrededor de la

media P dado el nivel de confianza 1-α; así, 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 que el promedio de las que bajaron son $\bar{p} = 0.17$; 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).

Como podemos observar el método estadístico nos permite realizar estudios tales que nos permite autocorregir en un momento dado nuestra apreciación no solo en cuanto al tamaño de una muestra, sino también a la hora de dar una confianza el momento de emitir nuestros resultados.

Bibliografía del tema 3

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

KREYSZIG Erwin*, Matemáticas avanzadas para ingeniería*, *vol. 2*, Limusa, 1996.
LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill.
RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica,1980.
WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.3.1. Elabora un resumen de las ventajas y desventajas al utilizar una estimación de intervalo sobre una estimación del tipo puntual, con los libros citados en la bibliografía.

A.3.2. Elabora un cuadro que permita comparar los procedimientos para calcular los intervalos de confianza para la media, la varianza y para la proporción.

A.3.3. Elabora un ensayo de la importancia de tener un buen conocimiento sobre las distribuciones: “t de student”, “Chi-cuadrada” “F” y “normal estándar”.

A.3.4. Realiza los ejercicios sobre la construcción de intervalos de confianza para la media, para la varianza y para la proporción que vienen en los libros de la bibliografía del tema.

A.3.5. Realiza los ejercicios sobre el cálculo del tamaño de la muestra tanto para la media como para la proporción que están citados en la bibliografía del tema.

A.3.6. Investiga en revistas especializadas la importancia de las estimaciones de intervalo y las razones de tal necesidad.

A.3.7. Elabora un resumen del tema.

Cuestionario de autoevaluación

1. ¿Cuál será la probabilidad de que un auditor tenga cuatro éxitos si va a realizar cinco auditorias, suponiendo que las probabilidades de éxito y por ende las de fracaso son independientes de una auditoria a otra?

2. Suponga usted que la llegada de trabajos a un despacho contable obedece a una distribución de Poisson y que en dicho despacho realizamos un muestreo;

durante el primer día llegaron dos trabajos; en el segundo, cuatro; en el tercero, tres; en el cuarto, cinco; y en el quinto, dos. Encuentre usted el estimador de máxima verosimilitud correspondiente.

3. Si realizamos un muestreo en un autolavado donde durante la primera hora llegaron dos automóviles; en la segunda, cuatro; en la tercera, tres; en la cuarta, cinco; y en la quinta, dos; encuentre usted el estimador de máxima verosimilitud correspondiente.

4. Una muestra aleatoria de tamaño 49 tiene una media de 157 y una desviación estándar de 14.7. Determine un intervalo con 95% de confianza para la media verdadera de la muestra.

5. Suponiendo que 64 mediciones de la densidad del cobre dieron por resultado una media de 8.81 y una desviación estándar de 0.24, calcule un intervalo con 99% de confianza para la densidad verdadera.

6. En una muestra aleatoria de 125 llantas para automóvil, se encontró que la vida media fue de 35,000 km. y la desviación estándar de 4,000. Determine un intervalo con 68% de confianza para la vida media.

7. Un estudio sobre ciertas acciones comunes permitió conocer que en una muestra aleatoria de 100 acciones la rentabilidad anual promedio fue de 4.2%, mientras que su desviación estándar es de 0.6%. Determine un intervalo, con 95% de confianza, para la rentabilidad promedio.

8. ¿Cuál es la diferencia entre una estimación y un estimador?

9. ¿Qué es un intervalo de confianza?

10. Señale, ¿por qué son preferibles las estimaciones de intervalo a las estimaciones puntuales?

Examen de autoevaluación

1. **Cuando se define un intervalo de valores de la muestra y se menciona que dentro del mismo es muy probable que se encuentre el parámetro poblacional, se dice que se está realizando:**
   a. Un análisis estadístico
   b. Una estimación de punto
   c. Una estimación de intervalo
   d. Una prueba de hipótesis
   e. Un estudio inferencial

2. **Suponga que estamos realizando la estimación por intervalo del valor de la media poblacional considerando un nivel de confianza del 99%, ¿cuál de los intervalos siguientes expresa nuestra intención?**
   a. $\mu_s \pm \sigma_s$
   b. $\mu_s \pm 1.96\sigma_s$
   c. $\mu_s \pm 2\sigma_s$
   d. $\mu_s \pm 2.58\sigma_s$
   e. $\mu_s \pm 3\sigma_s$

3. **Determine el intervalo de valores correspondiente a un nivel de confianza del 99% para el valor de la media poblacional si una muestra de 200 datos dieron una media de 0.824 pulgadas con una desviación estándar de 0.042 pulgadas:**
   a. 0.824 ± 0.005
   b. 0.824 ± 0.006
   c. 0.824 ± 0.009
   d. 0.824 ± 0.008
   e. 0.824 ± 0.003

4. **La corrección que se realiza al valor de la desviación estándar por la consideración de población finita depende de la relación que guardan los tamaños de la población y de la muestra, ¿cuál es la relación por considerar?**
   a. n/N >5%
   b. n/N < 5%
   c. n/N =5%
   d. n/N = 10%
   e. n/N >10%

5. **Una muestra al azar de 50 calificaciones de proyectos de inversión de un total de 200 arrojó una media de 75 y una desviación estándar de 10. ¿Con qué nivel de confianza podrá decirse que la media de las 200 calificaciones es de 75±1?**
   a. 73.2%
   b. 46.9%
   c. 63.4%
   d. 56.52%
   e. 81.0%

6. **Durante el envase de mermeladas se obtuvo en un envase un peso de 216.48 gramos. Si se sabe que el error probable es de 0.272 gramos. ¿Cuáles son los límites de confianza del 95% (en gramos) para dicho peso?**
   a. 216.48 ± 0.56
   b. 216.48 ± 0.57
   c. 216.48 ± 0.55
   d. 216.48 ± 0.53
   e. 216.48 ± 0.54

7. Si las distribuciones muestrales de dos estadísticos tienen la misma media, entonces el estadístico más eficiente es el que tenga:

a) Solo una frecuencia modal
b) Menor varianza
c) Sesgo hacia la derecha
d) Mediana y media más parecidas
e) Más intervalos de clase

8. La precisión o margen de error de una estima se conoce como:

a. Seguridad
b. Variación
c. Aproximación
d. Desviación estándar
e. Varianza

9. Si en una muestra grande con distribución normal se toma un estadístico S, ¿qué porcentaje de la muestra se encuentra en el intervalo $\mu_s \pm \sigma_s$?:

a. 60%
b. 68.27%
c. 75%
d. 95.45%
e. 90%.

10. Si en una muestra grande con distribución normal se toma un estadístico S, ¿con qué intervalo se obtiene el 99% de nivel de confianza?:

a) $\mu_s \pm \sigma_s$
b) $\mu_s \pm 1.96\sigma_s$
c) $\mu_s \pm 2\sigma_s$
d) $\mu_s \pm 2.58\sigma_s$
e) $\mu_s \pm 3\sigma_s$

Tema 4. Prueba de hipótesis

Objetivo particular

El alumno analizará y entenderá la importancia que tienen las pruebas de hipótesis en la toma de decisiones dentro de las empresas en general.

Temario detallado

4. Pruebas de hipótesis

4.1 Etapas básicas en pruebas de hipótesis

4.2 Concepto de hipótesis nula y alternativa

4.3 Error tipo I y tipo II, nivel de significación, curva operativa característica, potencia de una prueba.

4.4 Comprobación de hipótesis referentes a la media aritmética de una población, con muestras grandes y pequeñas.

Introducción

En este tema, el alumno investigará y analizará el concepto de prueba de hipótesis y lo aplicará sobre varianzas, medias, etc.; ello le permitirá percatarse de la importancia que tienen las pruebas de hipótesis para la toma de decisiones dentro de las empresas.

Actualmente, sabemos que la matemática es una herramienta importante en la toma de decisiones, y la estadística junto con todos sus procesos no es la excepción; así, es importante que el alumno desarrolle todos los conceptos y ejercicios planteados en la presente unidad, enriqueciendo su cultura para su futuro desempeño profesional.

Sabemos que cuando las personas **toman decisiones**, inevitablemente lo hacen con base en las creencias que tienen en relación al mundo que los rodea; llevan en la mente una cierta imagen de la realidad, piensan que algunas cosas son verdaderas y otras falsas y actúan en consecuencia, así, los ejecutivos de empresas toman todos

los días decisiones de importancia crucial porque tienen ciertas creencias tales como:

- De que un tipo de máquina llenadora pone al menos un kilogramo de detergente en una bolsa.
- De que cierto cable de acero tiene una resistencia de 100 kg. o más a la rotura.
- De que la duración promedio de una batería es igual a 500 horas.
- De que en un proceso de elaboración de cápsulas éstas contengan precisamente 250 miligramos de un medicamento,
- Que la empresa de transportes de nuestra competencia tiene tiempos de entrega más rápidos que la nuestra.
- De que la producción de las plantas de oriente contiene menos unidades defectuosas que las de occidente.

Incluso los estadistas basan su trabajo en creencias tentativas:

- Que dos poblaciones tienen varianzas iguales.
- Que esta población está normalmente distribuida.
- Que estos datos muestrales se derivan de una población uniformemente distribuida.

En todos estos casos y en muchos más, las personas actúan con base en alguna creencia sobre la realidad, la cual 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.

Por lo tanto, de una manera sencilla podemos decir que una **prueba de hipótesis** es un método sistemático de evaluar creencias tentativas sobre la realidad, dicho método requiere de la confrontación de tales creencias con evidencia real y decidir, en vista de esta evidencia, si dichas creencias se pueden conservar como razonables o deben desecharse por insostenibles.

A continuación estudiaremos la forma en que las creencias de las personas pueden ser probadas de manera sistemática.

4.1 Etapas básicas en pruebas 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: Formulación de dos hipótesis opuestas

El primer paso para probar una hipótesis es siempre formular dos hipótesis opuestas, que sean mutuamente excluyentes y, también colectivamente exhaustivas, del experimento que estemos evaluando. 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, simbolizada por H_1 o bien por Ha, es la **hipótesis alternativa.**

Paso 2: Selección de un estadístico de prueba

El segundo paso para probar una hipótesis es la selección de un estadístico de prueba. Un **estadístico de prueba** es aquel calculado con base en una sola muestra aleatoria simple tomada de la población de interés; en una prueba de hipótesis sirve para establecer la verdad o falsedad de la hipótesis nula.

Paso 3: Derivación de una regla de decisión

Una vez que hemos formulado de manera apropiada las 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 nos permite determinar si la hipótesis nula debe ser **aceptada** o si debe ser **rechazada** a favor de la alternativa.

Se dice que los valores numéricos del estadístico de prueba para los que $\mathbf{H_0}$ es aceptada están en la **región de aceptación** y son considerados **no significativos estadísticamente**.

Por el contrario, si el valor numérico del estadístico de prueba se encuentra en la región de rechazo, esto aconseja que la hipótesis alternativa sustituya a la desacreditada hipótesis nula; entonces este valor es considerado estadísticamente significativo.

Es importante notar que la aceptación o rechazo se refiere a la hipótesis nula $\mathbf{H_0}$.

Paso 4[22]: 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).
c) Confrontar con la regla de decisión (derivada en el paso 3).

[22] Heinz Kohler, *Estadística para negocios y economía*, p. 384.

4.2 Concepto de hipótesis nula y alternativa

La **hipótesis nula**, $\mathbf{H_0}$, 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 H_0 se corrobora en una prueba de hipótesis, no es necesario tomar ninguna acción.

La **hipótesis alternativa**, $\mathbf{H_1}$, es la segunda de dos opuestas en una prueba de hipótesis. Es un medio para hacer aseveraciones sorprendentes que contradicen la sabiduría convencional. Si $\mathbf{H_0}$ no se puede corroborar en una prueba de hipótesis, $\mathbf{H_1}$ se acepta tentativamente y esto requiere iniciar una acción. Por lo tanto, se puede considerar a $\mathbf{H_1}$ como la hipótesis de acción.

Por ejemplo:

Establezca las dos hipótesis para cada una de las situaciones siguientes:

1. Un fabricante de láminas de aluminio que se utilizan para la elaboración de la latas para refrescos asegura que éstas tienen 1 milímetro de espesor en promedio.

   Solución:

   $$\mathbf{H_0 : \mu_0 = 1\ mm} \qquad \mathbf{H_1 : \mu_0 \neq 1\ mm}$$

2. Un fabricante de varillas de acero especial que son utilizadas en la construcción de edificios muy altos asegura que éstas poseen una resistencia promedio a la tracción de al menos 2000 libras.

   Solución**:**

   $$\mathbf{H_0 : \mu_0 \geq 2000} \qquad \mathbf{H_1 : \mu_0 < 2000}$$

3. Un fabricante de computadoras desea probar lo dicho por un supervisor acerca de que el ensamble de una computadora promedia al menos 50 minutos.

Solución:

$\mathbf{H_0:}\ \mu_0 \geq 50$ $\mathbf{H_1:}\ \mu_0 < 50$

- **Tipos de pruebas de hipótesis**

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

Si la afirmación de H_0 contiene el signo de igualdad, entonces la prueba se llama no direccional**,** mientras que si tal afirmación no contiene el signo de igualdad (esto es, si involucra los signos menor o mayor que), entonces la prueba se llama direccional**.** Las pruebas no direccionales se llaman también **pruebas de dos colas** y las direccionales se nombran pruebas de una cola**.**

Así, si la afirmación de "H_0" contiene el símbolo ">", entonces la prueba se llama prueba direccional de cola izquierda; por el contrario Si la afirmación de H_0 tiene el símbolo "<", entonces la prueba se denomina prueba direccional de cola derecha.

Quienes investigan el mercado de consumo tienen una **hipótesis alternativa o de investigación**: 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 H_0, es la negación de la hipótesis alternativa H_1. La estrategia básica en las pruebas de hipótesis es tratar de apoyar la hipótesis alternativa **"contradiciendo"** la hipótesis nula.

4.3 Error tipo I y tipo II, nivel de significación, curva operativa característica, potencia de una prueba

- **Error tipo I**[23]

En una prueba estadística, rechazar la hipótesis nula cuando ésta es verdadera se denomina **error tipo I**. Y 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.

Problema ejemplo: **Incurrir en un riesgo α**

Un fabricante de varillas de acero especial que son utilizadas en la construcción de edificios muy altos ha contratado a un estadista para que pruebe si sus varillas ciertamente tienen un promedio de resistencia a la tensión de al menos 2000 libras ¿Cuáles son las implicaciones si el nivel de significancia de la prueba de hipótesis se fija en: α = 0.08?

Solución:

Dadas las hipótesis: $H_0 : \mu_0 \geq 2000$ y $H_1 : \mu_0 < 2000$

el procedimiento asegura lo siguiente: aun cuando las varillas tengan un promedio de resistencia a la tensión de 2000 libras o más, en el 8% de todas las pruebas la

[23] Mendenhall/Reinmuth, *Estadística para administración y economía*, p. 149.

conclusión será lo contrario.

- **Error tipo II**[24]

En una prueba estadística, aceptar la hipótesis nula cuando ésta es falsa se denomina **error tipo II.** 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 con ellas.

Ejemplo: **incurrir en un riesgo β**

El fabricante de computadoras ha contratado a un estadista para probar si el ensamble de una computadora toma un promedio de al menos 50 minutos. ¿Cuáles son las implicaciones si el riesgo β de la prueba es igual a 0.2?

Solución:

Dadas las hipótesis: **H_0 :** $\mu_0 \geq 50$ y **H_1 :** $\mu_0 < 50$

El procedimiento asegura lo siguiente: incluso si el tiempo de ensamble en efecto promedia más de 50 minutos, en el 20% de todas las pruebas la conclusión será lo contrario. Sin embargo, en el 80% de dichas pruebas este tipo de error se evita, lo que indica la **potencia de la prueba**.

[24] Mendenhall/Reinmuth, *Op cit*,. p. 149.

- **Nivel de significancia**

El nivel de significancia o significación es la probabilidad de cometer un error tipo I, es decir, el valor que se le asigna a **α**.

- **Potencia de la prueba**

Es posible[25] 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 $P_{(\text{rechazar Ho cuando es verdadera})} = \alpha$,

tenemos:

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

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

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:

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

[25] Richard C. Weimer, *Estadística*, p. 461.

La probabilidad de rechazar la hipótesis nula H_0 cuando es falsa se llama **potencia de la prueba.**

SÍMBOLO DE LA PROBABILIDAD	DEFINICIÓN
α	Nivel de significancia. Probabilidad de un error tipo I
β	Probabilidad de un error tipo II
1-α	Nivel de confianza. Probabilidad de no rechazar H_0 cuando es verdadera
1-β	Potencia de la prueba. Probabilidad de rechazar H_0 cuando es falsa.

Cuadro 4.1. Probabilidades asociadas con los cuatro resultados posibles de una prueba de hipótesis.

4.4 Comprobación de hipótesis referentes a la media aritmética de una población, con muestras grandes y pequeñas

Hasta aquí, hemos visto las dos técnicas clásicas para hacer inferencias sobre el valor de un parámetro desconocido: **la estimación y 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-α)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; un laboratorio farmacéutico anuncia que una de sus tableta para bajar la temperatura contiene 10 miligramos de aspirina. El estudio de 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%?

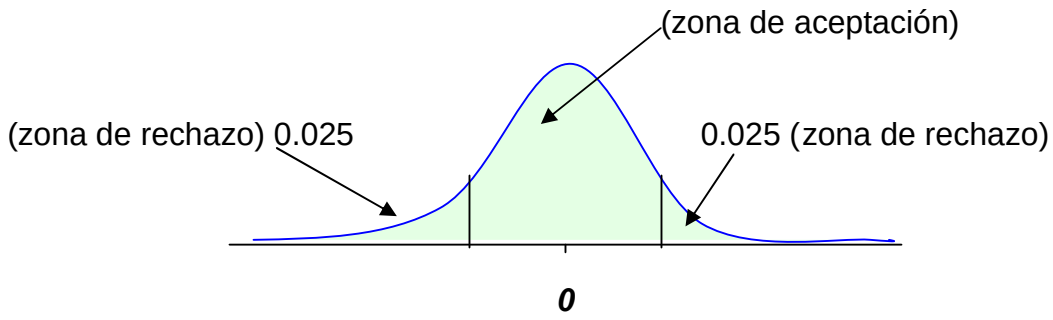
Resolvamos este ejemplo, utilizando la prueba de hipótesis:

Paso 1. Establecemos las dos hipótesis opuestas y dado que se supone que la tableta contiene10 miligramos de aspirina entonces:

$$Ho: \mu=10$$

$$H_1: \mu\neq10$$

Observemos 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:

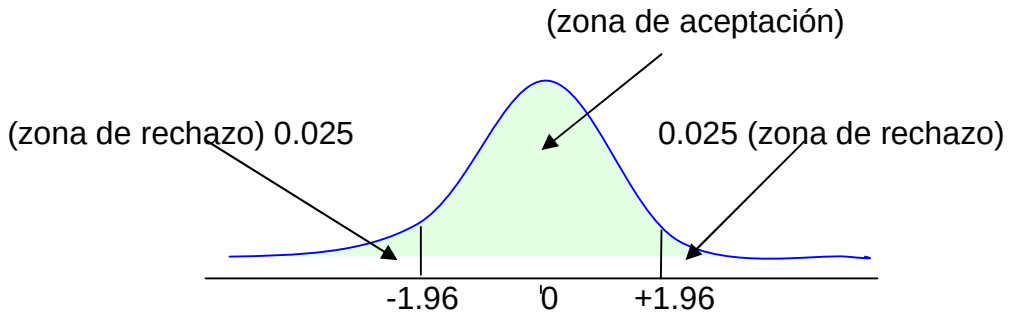


Curva de la distribución normal estándar,
se puede apreciar las zonas de aceptación y de rechazo

Paso 2. Seleccionar el estadístico de prueba, que 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 $z<-z_{0.025}$ ó $z>z_{0.025}$ resulta claro al utilizar una tabla de la distribución normal estándar en la 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

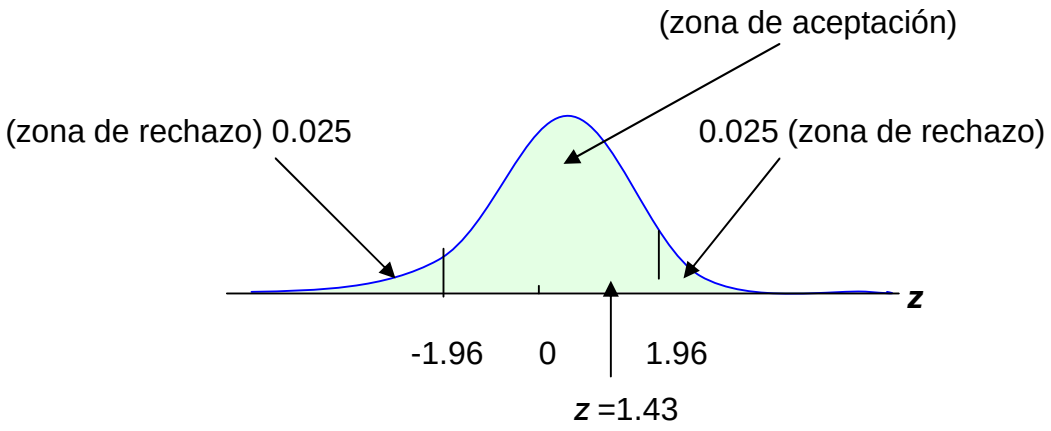
Considerando que el estadístico de prueba es:

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

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:



Curva de la distribución normal estándar,
se puede apreciar la confrontación del estadístico de prueba con la regla de decisión.

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 recordar 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, 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).

Como podemos observar, 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.

Así

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

nula.

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

a. Pruebas de hipótesis (muestras pequeñas)[26]

En las pruebas de hipótesis que hemos venido realizando, 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.

Ejemplo: los **datos muestrales** pueden sugerir que algo relevante está sucediendo en la población o que no esta sucediendo. Por ejemplo, el estudio de una muestra de clientes potenciales puede sugerir que la tendencia del mercado es preferir una nueva marca de algún producto sobre la ya existente. Resulta claro que en este caso, los datos provienen de una muestra limitada y por lo mismo están sujetos a cierto grado de variación aleatoria. Probar hipótesis estadísticas es una manera de estimar si los resultados aparentes en una muestra indican de manera concluyente que en realidad algo está pasando.

[26] R. D. Mason, *et al.*, *Estadística para administración y economía*, p. 307.

Bibliografía del tema 4

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

KOHLER Heinz, *Estadística para negocios y economía*, Editorial CECSA, 1996.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill.

MENDENHALL, William, REINMUTH, James, *Estadística para administración y economía*, Grupo Editorial Iberoamericana, 1981.

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica.

WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.4.1. Elabora un resumen de las etapas básicas en pruebas de hipótesis que vienen propuestas en los diferentes libros de la bibliografía del tema.

A.4.2. Elabora un cuadro comparativo de los conceptos de hipótesis nula y alternativa que vienen en los diferentes libros de la bibliografía del tema.

A.4.3. Investiga los diferentes conceptos sobre: error tipo I y error tipo II que dan los diferentes libros de la bibliografía del tema.

A.4.4. Investiga la importancia de las hipótesis en la toma de decisiones.

A.4.5. Elabora cinco ejercicios de las etapas básicas en pruebas de hipótesis

A.4.6. Elabora un cuadro con las implicaciones de tener muestras grandes y pequeñas a la hora de hacer la prueba de una hipótesis.

A.4.7. Investiga en revistas especializadas como formulan y se presentan las hipótesis nula y alternativa.

Cuestionario de autoevaluación

1. Una hipótesis nula se establece como:

2. El nivel de significancia en una prueba de hipótesis es:

3. Un estadístico de prueba en una prueba de hipótesis es:

4. ¿Cuáles son las etapas básicas en pruebas de hipótesis?

5. En una prueba de una cola el signo de la hipótesis nula puede ser:

6. El nivel de significancia en una prueba de hipótesis corresponde a:

7. Un artículo de prensa señaló que la edad promedio de los accionistas de empresas está decreciendo. El gerente de una de ellas decide realizar una prueba de hipótesis para verificar si este señalamiento aplica a su empresa. Se considera una desviación estándar de 12 años y una muestra de tamaño 250, cuya media muestral es de 53 años. Para un nivel de significancia del 5%, ¿cuál es el valor crítico para la prueba?

8. La Ingeniería de Control de Calidad probó un lote de tubos fluorescentes y encontró una vida promedio de 1,570 horas con desviación estándar de 120 horas. Con un nivel de significación del 5%, determinar la regla de decisión.

9. Se prueba un lote de un nuevo modelo experimental de 100 lámparas de vapor de sodio; su vida es de 43,000 horas y su desviación estándar, de 2,000 horas. Si la vida normal de las lámparas es de 40,000 horas. Probar con un nivel de significación del 10%

10. En una planta embotelladora de leche se toma una muestra de 500 botellas; 40 de ellas se obtienen con impurezas. Si se supone que el límite máximo de impurezas es 7%. Establezca la regla de decisión para un nivel de significancia del 4%

Examen de autoevaluación

1. Suponga que usted forma parte de un grupo de protección al consumidor, y está interesado en determinar si el peso promedio de cierta marca de arroz, empacado en paquetes de 1 kg, es menor que el peso anunciado; para ello, usted elige una muestra aleatoria de 50 bolsas, de las cuales obtiene una media de 980gr. y una desviación estándar de 70 gr. Para este problema, cuál es la hipótesis nula:

a) H_0: $\mu \leq 1$ kg.

b) H_0: $\mu \geq 1$ kg.

c) H_0: $\mu = 1$ kg.

d) H_0: $\mu \neq 1$ kg.

2. ¿Cuál es la hipótesis alternativa?

a) H_1: $\mu \geq 1$ kg.

b) H_1: $\mu \neq 1$ kg.

c) H_1: $\mu = 1$ kg.

d) H_1: $\mu < 1$ kg.

3. Para este problema, se dice que la prueba de hipótesis es de

a) cola izquierda

b) cola derecha

c) dos colas

d) sin cola

4. Si el nivel de significancia es del 5%, entonces la hipótesis nula se:

a) acepta

b) es indiferente

c) rechaza

d) debe replantear

5. Se supone que un medicamento que sirve como antibiótico contiene 1000 unidades de penicilina. Una muestra aleatoria de 100 de estos antibióticos produjo una media de 1020 gramos y una desviación estándar de 140 gramos. Para este problema, la hipótesis nula es:

a) H_0: $\mu \neq 1000$

b) H_0: $\mu = 1000$

c) H_0: $\mu \leq 1000$

d) H_0: $\mu \geq 1000$

6. La hipótesis alternativa es:

a) H_1: $\mu \geq 1020$

b) H_1: $\mu = 1000$

c) H_1: $\mu < 1020$

d) H_1: $\mu \neq 1000$

7. Para este problema se dice que la prueba de hipótesis es de:

a) una cola

b) dos colas

c) cola izquierda

d) cola derecha

8. **A un nivel de significancia del 5%, la hipótesis nula se:**

a. acepta

b. rechaza

c. es indiferente

d. replantea

9. **Se sabe que los voltajes de una marca de pilas "AAA" para calculadora se distribuyen normalmente con un promedio de 1.5 volts; se probó una muestra aleatoria de 15 y se encontró que la media fue de 1.3 volts y que la desviación estándar fue de 0.25 volts. Para este problema, la hipótesis nula es:**

a) H_0: $\mu \leq 1.5$

b) H_0: $\mu \geq 1.5$

c) H_0: $\mu = 1.5$

d) H_0: $\mu \neq 1.4$

10. **La hipótesis alternativa es:**

a) H_1: $\mu \geq 1.4$

b) H_1: $\mu > 1.5$

c) H_1: $\mu \neq 1.5$

d) H_1: $\mu < 1.5$

Tema 5. Estadística no paramétrica

Objetivo particular

El alumno analizará los fundamentos de la estadística no paramétrica, su importancia, desarrollo y evolución, así como su aplicación en las áreas económico-administrativas.

Temario detallado

5. Estadística no paramétrica

5.1 Características de las pruebas no paramétricas

5.2 Pruebas de bondad de ajuste

5.3 Tablas de contingencia

5.4 Prueba de los signos de Wilcoxon

5.5 Prueba de rachas

5.6 Otras pruebas

Introducción

En el Tema 4 correspondiente a “Pruebas de Hipótesis”, se estudiaron pruebas tanto para las medias poblacionales como para las proporciones poblacionales. En algunos casos el tamaño de la muestra era mayor que 30, mientras que en otras la muestra era pequeña.

Sin embargo, todas estas situaciones de pruebas presentaron una característica común: necesitaban de ciertos supuestos respecto a la población. Por ejemplo, las pruebas “t” y las pruebas “F” requerían el supuesto de que la población estuviese distribuida normalmente. Debido a que tales pruebas dependen de postulados sobre la población y sus parámetros, se denominan **pruebas paramétricas**.

En la práctica, surgen muchas situaciones en las cuales simplemente no es posible hacer de forma segura ningún supuesto sobre el valor de un parámetro o sobre la forma de la distribución poblacional, por lo que la mayoría de las pruebas descritas en los capítulos anteriores no son aplicables. Más bien se deben utilizar otras pruebas que no dependan de un solo tipo de distribución o de valores de parámetros específicos; estas pruebas se denominan **pruebas no paramétricas** (o libres de distribución).

5.1 Características de las pruebas no paramétricas

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.

Generalmente son más fáciles de realizar y comprender ya que no requieren cálculos laboriosos ni el ordenamiento o clasificación formal de datos o mediciones más exactas de parámetros poblacionales.

5.2 Pruebas de bondad de ajuste

Pruebas de bondad de ajuste. Medidas sobre qué tan cerca se ajustan los datos muestrales observados a una forma de distribución particular planteada como hipótesis. Si el ajuste es razonablemente cercano, puede concluirse que si existe la forma de distribución planteada como hipótesis.

Con frecuencia, las decisiones en los negocios requieren que se pruebe alguna hipótesis sobre la distribución poblacional desconocida. Por ejemplo, se puede plantear la hipótesis que la distribución poblacional es uniforme y que todos los valores posibles tienen la misma probabilidad de ocurrir.

Las hipótesis que se probarían son las siguientes:

H_0: La distribución poblacional es uniforme.

H_1: La distribución poblacional no es uniforme.

La prueba de bondad de ajuste se utiliza entonces para determinar si la distribución de los valores en la población se ajusta a una forma en particular planteada como hipótesis —en este caso, una distribución uniforme. De la misma manera que con todas las pruebas estadísticas de esta naturaleza, los datos muestrales se toman de la población y éstos constituyen la base de los hallazgos.

Si existe gran diferencia entre lo que realmente se observa en la muestra y lo que se esperaría observar si la hipótesis nula fuera correcta, es menos probable que la hipótesis nula sea verdadera. Es decir, la hipótesis nula debe rechazarse cuando las observaciones obtenidas en la muestra tienen diferencias significativas del patrón que se espera que ocurra en la distribución planteada como hipótesis.

Por ejemplo, si se hace rodar un dado “bueno”, es razonable plantear como hipótesis un patrón de resultados tal que cada resultado (números del 1 al 6) ocurra aproximadamente un sexto de las veces. Sin embargo, si un porcentaje significativamente grande o pequeño de número pares ocurre, puede concluirse que el dado no está balanceado adecuadamente y que la hipótesis es falsa. Es decir, si la diferencia entre los patrones de eventos que en realidad se observaron y el patrón de eventos que se espera que ocurra si la hipótesis nula es correcta, prueba ser demasiado grande como para atribuirlo a un error de muestreo debe concluirse entonces que la población presenta una distribución distinta de la especificada en la hipótesis nula.

Para **contrastar la hipótesis relativa** a una distribución poblacional, se debe analizar la diferencia entre las expectativas con base en la distribución planteada como hipótesis y los datos reales que aparecen en la muestra.

Para lo anterior, se utiliza la distribución χ^2 (Chi-cuadrada) como prueba estadística de bondad de ajuste y se utiliza alguna de las siguientes fórmulas:

$$\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} \quad \text{o} \qquad \chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n$$

En donde:

χ_e^2 Es el estadístico de prueba.

f_o Es la frecuencia de los eventos observados en los datos muestrales

f_e Es la frecuencia de los eventos esperados si la hipótesis nula es correcta

k Es el número de categorías o clases.

n Es el número de datos.

La prueba tiene K-m-1 grados de libertad, en donde “m” es el número de parámetros por estimar.

Por ejemplo si se desconoce la media o varianza de la población y se tienen que “estimar” cada uno representa un grado menos de libertad.
En la fórmula podemos observar que el numerador mide la diferencia entre las frecuencias de los eventos observados y las frecuencias de los eventos esperados. Para este tipo de pruebas no paramétricas se establecen los siguientes 5 pasos:

Paso 1. Establecer la hipótesis nula (H_o) y la hipótesis alternativa (H_1).

La H_o indica que no hay diferencias significativas entre las frecuencias observadas y las frecuencias esperadas. Cualquier diferencia puede atribuirse al muestreo o a la casualidad. La H_i indica por lo tanto que si hay diferencias significativas entre una distribución esperada y la estimada para la población.

Paso 2. Elegir un nivel de significación (α).

Paso 3. Elegir y calcular el estadístico de prueba χ_e^2

Paso 4. Establecer la regla de decisión.

Paso 5. Calcular el valor de Chi-cuadrada crítica (χ_c^2) y tomar la decisión.

Se estudiarán 3 tipos de pruebas de bondad de ajuste:

I. Prueba para un ajuste uniforme.

II. Prueba de ajuste para un patrón específico.

III. Prueba de normalidad.

I. Prueba para un ajuste uniforme

Se pretende probar que la distribución de datos es uniforme.

Ejemplo de aplicación; un nuevo director de mercadotecnia tiene la responsabilidad de controlar el nivel de existencias para 4 tipos (A, B, C ,D) de automóviles vendidos por su empresa de distribución. Le han informado que la demanda de cada tipo de automóviles es la misma. Para probar esta hipótesis se selecciona una muestra aleatoria de 100 automóviles vendidos en los últimos meses. Se requiere un nivel de significación del 10%.

Se cuenta con la siguiente información:

Tipo Automóvil	Ventas Observadas
A	32
B	21
C	19
D	28

Solución*:*

Paso 1. H_o = La demanda es uniforme para los 4 tipos de automóviles.

H_1 = La demanda no es uniforme para los 4 tipos de automóviles.

Paso 2. $\alpha = 0.10$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$ y se comprueba con:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n$$

Tipo Automóvil	Ventas Observadas	Ventas Esperadas	$\frac{(f_o - f_e)^2}{f_e'}$	$\frac{f_o^2}{f_e}$
A	32	25	1.96	40.96
B	21	25	0.64	17.64
C	19	25	1.44	14.44
D	28	25	0.36	31.36
Suma	100	100	4.40	104.40

Tabla de frecuencias observadas y esperadas

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 4.40$; utilizando la otra fórmula, comprobamos:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n = 104.4 - 100 = 4.40$$

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl=k-m-1 = 4-1=3$ y el nivel de significación es de *0.10*, se observa:

$$\chi_{c,0.10,3}^2 = 6.251$$

Por lo tanto como $\chi_e^2 < 6.251$, la hipótesis nula de que la demanda es uniforme, no se rechaza. Las diferencias no son lo suficientemente grandes para refutar la hipótesis nula; las diferencias no son significativas y pueden atribuirse simplemente a un error de muestreo.

Veamos otro ejemplo, una tienda vende 6 tipos de tarjetas de onomástico y se quiere saber si todas se venden en las mismas cantidades. Si en el siguiente día se vendieron 120 tarjetas, se esperaría que se vendieran 20 de cada una. Sin embargo, el número de tarjetas que se vendieron de cada tipo fueron: A – 13; B – 33; C – 14; D – 14; E – 36; F – 17.
Con esta información, probar que no hay diferencias significativas en el número de ventas de las tarjetas en estudio a un nivel de significación del 5%.

Solución:

Paso 1. H_o = Las tarjetas se venden en la misma cantidad.

H_1 = Las tarjetas no se venden en la misma cantidad.

Paso 2. $\alpha = 0.05$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$ y se comprueba con:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n$$

Tipo Tarjeta	Ventas Observadas	Ventas Esperadas	$\frac{(f_o - f_e)^2}{f_e'}$	$\frac{f_o^2}{f_e}$
A	33	20	2.45	8.45
B	13	20	8.45	54.45
C	14	20	1.80	9.80
D	7	20	8.45	2.45
E	36	20	12.80	64.80
F	17	20	0.45	14.45
Suma	120	120	34.40	154.40

Tabla de frecuencias observadas y esperadas:

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 34.40$; utilizando la otra fórmula, comprobamos:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n = 154.40 - 120 = 104.40$$

Paso 4. Regla de decisión: Si χ_e^2 es $\leq$ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl=k-m-1 = 6-153$ y el nivel de significación es de *0.10*, se observa:

$$\chi^2_{c,0.05,5} = 11.070$$

Por lo tanto como χ_e^2 > 11.070, se encuentra en la zona de rechazo. Las diferencias son lo suficientemente grandes para considerarlas significativas. Se concluye que es improbable que todas las tarjetas se vendan en el mismo número.

II. Prueba de ajuste a un patrón específico

Existen muchos casos en los cuales las frecuencias se prueban contra un patrón determinado en las que las frecuencias esperadas no son todas iguales.

Las frecuencias esperadas se calculan con datos porcentuales de la siguiente forma:

$f_e = np_i$; en donde

n= Tamaño de la muestra

p_i = Probabilidad de cada categoría como se especifica en la hipótesis nula.

Ejemplo de aplicación; un director de un banco trata de seguir una política de extender un 35% de sus créditos a empresas industriales; un 20% a empresas comerciales; un 18% a empresas de servicios; un 25% a empresas maquiladoras; y un 5% a empresas extranjeras.

Para demostrar que la política se está siguiendo, se seleccionaron 113 créditos que se aprobaron recientemente. Se encontró que 28 créditos se otorgaron a empresas

industriales; 22 a comerciales; 25 a empresas de servicios; 30 a maquiladoras; y 8 a empresas extranjeras. Probar esta hipótesis a un nivel de significación del 20%.

Solución:

Paso 1. H_o = Se mantuvo el patrón deseado.

H_1 = No se mantuvo el patrón deseado.

Paso 2. $\alpha = 0.20$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$ y se comprueba con:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n$$

Tipo de Empresa	Frecuencias Observadas	Frecuencias Esperadas	$\frac{(f_o - f_e)^2}{f_e'}$	$\frac{f_o^2}{f_e}$
Industrial	28	39.55	3.37	19.82
Comercial	22	22.60	0.02	21.42
De servicios	25	20.34	1.07	30.73
Maquiladoras	30	24.86	1.06	36.20
Extranjeras	08	05.65	0.98	11.33
Suma	113	113.00	6.50	119.50

Tabla de frecuencias observadas y esperadas:

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 6.50$; utilizando la otra fórmula, comprobamos:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n = 154.40 - 120 = 104.40$$

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl=k-m-1 = 6-153$ y el nivel de

significación es de *0.10*, se observa:

$$\chi^2_{c,0.05,5} = 11.070$$

Por lo tanto como χ^2_e > 11.070, se encuentra en la zona de rechazo. Las diferencias son lo suficientemente grandes para considerarlas significativas. Se concluye que es improbable que todas las tarjetas se vendan en el mismo número.

Otro ejemplo sería el de tres monedas fueron lanzadas 80 veces y se registró el número de veces que salieron "águilas":

x	0	1	2	3
f	20	38	18	4

Siendo *"x"* el "lado águila" y *"f"* el número de veces que salió "águila".

Con esta información, poner a prueba la hipótesis nula de que "x" es binomial con *n=3 y p= 0.5.* Usar un nivel de significación del 5%.

Solución:

Paso 1. H_o = *"x"* sigue una distribución binomial.

H_1 = *"x"* no sigue una distribución binomial.

Paso 2. $\alpha = 0.05$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$ y se comprueba con:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n$$

Se calculan las probabilidades de éxito binomiales para 0, 1, 2, y 3.

Fórmula: $P(x) = C_x^n \cdot p^x \cdot q^{n-x}$

$$P(0) = \frac{3!}{0!(3-0)!} 0.5^0 \cdot 0.5^3 = 0.125$$

$$P(1) = \frac{3!}{1!(3-1)!} 0.5^1 \cdot 0.5^2 = 0.375$$

$$P(2) = \frac{3!}{2!(3-2)!}0.5^2 \cdot 0.5^1 = 0.0.375$$

$$P(3) = \frac{3!}{3!(3-3)!}0.5^3 \cdot 0.5^0 = 0.125$$

Tabla de frecuencias observadas y esperadas:

x	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e'}$	$\frac{f_o^2}{f_e}$
0	20	10	10.00	40.00
1	35	30	02.13	48.13
2	18	30	04.80	10.80
3	04	10	03.60	01.60
Suma	80	80	20.53	100.53

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 20.53$; utilizando la otra fórmula, comprobamos:

$$\chi_e^2 = \frac{\sum_{i=1}^{k} f_o^2}{f_e} - n = 100.53 - 80 = 20.53$$

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen *gl=k-m-1 = 4-0-1=3* y el nivel de significación es de *0.20*, se observa: $\chi_{c,0.05,3}^2 = 7.815$

Por lo tanto como χ_e^2 > 7.815, se encuentra en la zona de rechazo. En consecuencia se concluye que *"x"* no sigue una distribución binomial con *n=3* y *p=0.50*.

III. Pruebas de normalidad

Se requiere probar que una serie de elementos de una población sigue una distribución normal por medio de una muestra.

Ejercicio de aplicación; en clases de buceo, los tanques de inversión se llenan a una presión promedio de 600 libras por pulgada cúbica (*psi*). Se permite una desviación estándar de 10 *psi.* Las especificaciones de seguridad permiten una distribución normal en los niveles de llenado. Probar la hipótesis a un nivel de significación del 5% si en una muestra se miden 1,000 tanques con los siguientes resultados:

Evento *psi*		**Frecuencia**
A	<580	020
B	580 a 590	142
C	590 a 600	310
D	600 a 610	370
E	610 a 620	128
F	>620	030

Solución:

Paso 1. H_o = Los datos siguen una distribución normal

H_1 = Los datos no siguen una distribución normal

Paso 2. $\alpha = 0.05$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$ Se calculan las probabilidades de éxito de una distribución normal para cada evento.

Fórmula: $z = \frac{x - \mu}{\sigma}$

Para el evento A: $P(A) = 0.5000 - P(z_1)$

$$z_1 = \frac{580 - 600}{10} = -2.0$$

En la tabla de distribución normal se encuentra: $P(z_1) = 0.4772$

Por lo tanto : $P(x \leq 580) = 0.5000 - 0.4772$

Para el evento B: $P(B) = P(\ddot{z}_1) - P(z_2)$

$$z_2 = \frac{590 - 600}{10} = -1.0$$

En la tabla de distribución normal se encuentra: $P(z_2) = 0.3413$

Por lo tanto: $P(580 \leq x \leq 590) = 0.4772 - 0.3413 = 0.1359$

Para el evento C: $P(C) = P(z_2)$

$$z_2 = \frac{590 - 600}{10} = -1.0$$

En la tabla de distribución normal se encuentra: $P(z_2) = 0.3413$

Por lo tanto : $P(590 \leq x \leq 600) = 0.3413$

Tabla de frecuencias observadas y esperadas:

Evento	f_o	$P(x)$	f_e	$\frac{(f_o - f_e)^2}{f_e'}$
A	020	0.0228	0022.8	0.344
B	142	0.1359	0135.9	0.274
C	310	0.3413	0341.3	2.870
D	370	0.3413	0341.3	2.413
E	128	0.1359	0135.9	0.459
F	030	0.0228	0022.8	2.274
Suma	1000	1.0000	1000.0	8.634

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 8.634$;

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl=k-m-1 = 6-0-1=5$ y el nivel de significación es de *0.05*, se observa: $\chi^2_{c,0.05,5} = 11.070$

Por lo tanto como χ_e^2 < 11.070, se encuentra en la zona de aceptación. En consecuencia se concluye que los datos de la población siguen una distribución normal.

Se presenta un siguiente ejemplo; los fabricantes de una marca de computadoras reportan en su publicidad que su vida media útil es de 6 años con una desviación estándar de 1.4 años. En una muestra de 90 computadoras vendidas hace 10 años se encontraron los siguientes tiempos de vida útil:

Evento	Tiempo de vida (años)	Frecuencia
A	Hasta 4	07
B	De 4 a 5	14
C	De 5 a 6	25
D	De 6 a 7	22
E	De 7 a 8	16
F	8 o más	06

Con esta información, ¿puede concluir el fabricante, con un nivel de significación del 5% que la vida útil de las computadoras tiene una distribución normal?

Solución:

Paso 1. H_o = La vida útil de las computadoras sigue una distribución normal.

H_1 = La vida útil de las computadoras no sigue una distribución normal

Paso 2. $\alpha = 0.05$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e}$. Se calculan las probabilidades de éxito de una distribución normal para cada evento.

Fórmula: $z = \frac{x - \mu}{\sigma}$

Para el evento A: $P(A) = 0.5000 - P(z_1)$

$$z_1 = \frac{4-6}{1.4} = -1.43$$

En la tabla de distribución normal se encuentra: $P(z_1) = 0.4236$

Por lo tanto : $P(x \leq 4) = 0.5000 - 0.4236 = 0.0764$

Para el evento B: $P(B) = P(z_1) - P(z_2)$

$$z_2 = \frac{5-6}{1.4} = -0.71$$

En la tabla de distribución normal se encuentra: $P(z_2) = 0.2611$

Por lo tanto: $P = 0.4236 - 0.2611 = 0.1625$

Para el evento C: $P(C) = P(z_2)$

$$z_2 = \frac{5-6}{1.4} = -0.71$$

En la tabla de distribución normal se encuentra: $P(z_2) = 0.2611$

Por lo tanto: $P(5 \leq x \leq 6) = 0.2611$

Tabla de frecuencias observadas y esperadas:

Evento	f_o	$P(x)$	f_e	$\frac{(f_o - f_e)^2}{f_e'}$
A	07	0.0764	06.876	0.0022
B	14	0.1625	14.625	0.0267
C	25	0.2611	23.499	0.0959
D	22	0.3413	23.499	0.0959
E	16	0.1359	14.625	0.1293
F	06	0.0228	06.876	0.1116
Suma	90	1.0000	90.000	0.4613

Por lo tanto: $\chi_e^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} = 0.4613$;

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl=k-m-1 = 6-0-1=5$ y el nivel de significación es de *0.05*, se observa: $\chi^2_{c,0.05,5} = 11.070$

Por lo tanto como χ_e^2 < 11.070, se encuentra en la zona de aceptación. En consecuencia se concluye que los datos de la población siguen una distribución normal.

5.3 Tablas de contingencia

En aplicaciones estadísticas es frecuente interesarse en calcular si 2 variables de clasificación, cuantitativas o cualitativas, son independientes o si están relacionadas.

Las hipótesis son:

H_o: Las variables de clasificación son independientes.

H_1: Las variables de clasificación son dependientes.

Estos modelos se basan también en la prueba JI-cuadrada por lo que se procede a comparar las frecuencias esperadas con las observadas para determinar que tan grande debe ser el alejamiento permitido para que la hipótesis de independencia pueda rechazarse.

Si el valor del estadístico de prueba JI-cuadrada es mayor que el valor critico, no se puede suponer que las 2 variables de clasificación sean independientes.
La fórmula del estadístico de prueba es la siguiente:

$$\chi_e^2 = \sum_1^{rc} \frac{(f_o - f_e)^2}{f_e}$$

en donde:

r es el N° de renglones de una tabla de contingencia.

c es el N° de columnas de una tabla de contingencia.

Los grados de libertad serán: $gl = (r-1)(c-1)$

Ejemplo; un director de investigación de productos debe determinar si existe alguna relación entre la clasificación de efectividad que los consumidores asignan a un nuevo insecticida y el sitio (urbano o rural) en los cuales se utilizan. De los 120 consumidores de la encuesta, 90 viven en zonas urbanas y 30 en rurales. El nivel de significación es del 1%.

En la siguiente tabla de contingencia se muestran las clasificaciones.

Atributo "A": Clasificación	**Atributo "B": Urbano**	**Ubicación Rural**
Por encima del promedio	24	13
En el promedio	48	10
Por debajo del promedio	18	07

Probar esta hipótesis con un nivel de significación es del 1%.

Solución*:*

Paso 1. H_o = La clasificación y la ubicación son independientes.

H_1 = La clasificación y la ubicación no son independientes.

Paso 2. $\alpha = 0.01$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_1^{rc} \frac{(f_o - f_e)^2}{f_e}$

Tabla de contingencia

Atributo "A": Clasificación	Atributo "B": Urbano	Ubicación Rural	Total
Por encima del promedio	24	13	037
En el promedio	48	10	058
Por debajo del promedio	18	07	025
Total	90	30	120

Se calculan las frecuencias esperadas relacionando las 2 variables. En un esquema matricial se utiliza su nomenclatura:

Calculo de: $f_{e_{11}} = \frac{f_{t_{13}}}{f_{t_{43}}} \cdot f_{t_{41}} = \frac{37}{120} \cdot 90 = 27.75$ $\quad f_{e_{12}} = \frac{f_{t_{13}}}{f_{t_{33}}} \cdot f_{t_{42}} = \frac{37}{120} \cdot 30 = 9.25$

$f_{e_{21}} = \frac{f_{t_{32}}}{f_{t_{43}}} \cdot f_{t_{41}} = \frac{58}{120} \cdot 90 = 43.50$ $\quad f_{e_{22}} = \frac{f_{t_{23}}}{f_{t43}} \cdot f_{t_{42}} = \frac{58}{120} \cdot 30 = 14.5$

$f_{e_{31}} = \frac{f_{t_{33}}}{f_{t_{43}}} \cdot f_{t_{33}} = \frac{25}{120} \cdot 90 = 18.75$ $\quad f_{e_{32}} = \frac{f_{t_{33}}}{f_{t_{43}}} \cdot f_{t_{42}} = \frac{25}{120} \cdot 30 = 6.25$

Tabla de frecuencias de clasificación

Atributo "A": Clasificación	Atributo "B": Urbano		Ubicación Rural	
	f_o	f_e	f_o	f_e
Por encima del promedio	24	24.75	13	09.25
En el promedio	48	43.50	10	14.50
Por debajo del promedio	18	18.75	07	06.25

Por lo tanto utilizando la fórmula del estadístico de prueba:

$$\chi_e^2 = \sum_{1}^{rc} \frac{(f_o - f_e)^2}{f_e} = \frac{(24-27.75)^2}{27.75} + \frac{(13-9.25)^2}{9.25} + \frac{(48-43.5)^2}{43.5} + \ldots\ldots + \frac{(7-6.25)^2}{6.25} = 4.009$$

Paso 4. Regla de decisión: Si χ_e^2 es $\leq$ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2:, si se tienen $gl = (r-1)(c-1) = (3-1)(2-1) = 2$ y el nivel de significación es de *0.01*, se observa: $\chi^2_{c,0.01,2} = 9.210$

Por lo tanto como $\chi_e^2 < 9.210$, la hipótesis nula es aceptada y se concluye que tanto la clasificación como la ubicación son factores independientes.

Un siguiente ejemplo, es el director de Marketing de un diario metropolitano de gran circulación, estudia la relación entre el tipo de actividad y la sección del periódico que es de su preferencia. De una muestra de lectores se obtuvo la siguiente información:

Tabla de frecuencias de clasificación

Lectores	Noticias Nacionales	Sociales	Deportes
Profesionistas	170	124	090
Estudiantes	112	100	120
Otros	130	088	090

¿Podemos concluir con un nivel de significación del 10% que si hay relación entre el tipo de actividad y la sección del periódico de su preferencia?

Solución*:*

Paso 1. H_o= No hay relación entre el tipo de actividad y la sección de su preferencia.

H_1= Si hay relación entre el tipo de actividad y la sección de su preferencia.

Paso 2. $\alpha = 0.10$

Paso 3. Se elegirá el estadístico de prueba: $\chi_e^2 = \sum_1^{rc} \frac{(f_o - f_e)^2}{f_e}$

Tabla de contingencia

Lectores	Noticias Nacionales	Sociales	Deportes	Total
Profesionistas	170	124	090	0384
Estudiantes	112	100	120	0332
Otros	130	088	090	0308
Total	**412**	**312**	**300**	**1,024**

Se calculan las frecuencias esperadas relacionando las 2 variables. En un esquema matricial se utiliza su nomenclatura:

Calculo de: $f_{e_{11}} = \frac{f_{t_{13}}}{f_{t_{43}}} \cdot f_{t_{41}} = \frac{384}{1024} \cdot 412 = 154.5$ $f_{e_{12}} = \frac{f_{t_{13}}}{f_{t_{33}}} \cdot f_{t_{42}} = \frac{384}{1024} \cdot 312 = 117.0$

$f_{e_{21}} = \frac{f_{t_{32}}}{f_{t_{43}}} \cdot f_{t_{41}} = \frac{332}{1024} \cdot 412 = 133.6$ $f_{e_{22}} = \frac{f_{t_{23}}}{f_{t43}} \cdot f_{t_{42}} = \frac{332}{1024} \cdot 312 = 101.2$

$f_{e_{31}} = \frac{f_{t_{33}}}{f_{t_{43}}} \cdot f_{t_{33}} = \frac{308}{1024} \cdot 412 = 123.9$ $f_{e_{32}} = \frac{f_{t_{33}}}{f_{t_{43}}} \cdot f_{t_{42}} = \frac{308}{1024} \cdot 312 = 93.8$

Se realizan cálculos similares para la sección de deportes.

Lectores	Noticias nacionales		Sociales		Deportes	
	f_o	f_e	f_o	f_e	f_o	f_e
Profesionistas	170	154.5	124	117.0	090	112.5
Estudiantes	112	133.6	100	101.2	120	097.2
Otros	130	123.9	088	93.8	090	090.3

Por lo tanto utilizando la fórmula del estadístico de prueba:

$$\chi_e^2 = \sum_{1}^{rc} \frac{(f_o - f_e)^2}{f_e} = \frac{(170-154.5)^2}{154.5} + \frac{(112-133.6)^2}{133.6} + \frac{(130-123.9)^2}{123.9} + \ldots\ldots + \frac{(90-90.3)^2}{90.3} = 15.936$$

Paso 4. Regla de decisión: Si χ_e^2 es ≤ que χ_c^2 no se rechaza la H_o. En caso contrario rechazar la H_o.

Paso 5. En la tabla de la distribución χ^2, si se tienen $gl = (r-1)(c-1) = (3-1)(3-1) = 4$ y el nivel de significación es de *0.10*, se observa: $\chi^2_{c,0.10,4} = 7.779$

Por lo tanto como χ_e^2 > 7.779, se rechaza la hipótesis nula por lo que se puede afirmar que si hay una relación entre la actividad de los lectores y la sección del periódico de su preferencia.

5.4 Prueba de los signos de Wilcoxon

Se utiliza como una **alternativa no paramétrica** cuando se trata de comparar los datos de 2 poblaciones o de una misma población mediante una muestra apareada en la que cada unidad experimental genera 2 observaciones pareadas o ajustadas, una de la población 1 y una de la población 2. Las diferencias entre las observaciones pareadas permiten tener una buena perspectiva respecto de la

diferencia entre las 2 poblaciones.

La metodología del análisis paramétrico de una muestra pareada requiere de datos de intervalo y de la suposición de que la población de las diferencias entre los pares de observaciones tenga una distribución normal. Con este supuesto se puede usar la distribución “t” para probar la hipótesis nula es decir que no hay diferencias entre las medias poblacionales. Si no es así se debe utilizar la prueba de rango con signo de Wilcoxon.

La prueba de los rangos con signo usa los rangos de los valores absolutos de las diferencias pareadas, asignando el rango 1 a la diferencia con valor absoluto mínimo, el rango 2 a la siguiente diferencia con menor valor absoluto y así se procede sucesivamente. Se deben descartar los rangos con diferencias de cero y en caso de valores absolutos repetidos, a cada uno de ellos se les otorga el valor promedio de los rangos ocupados por los valores repetidos. A cada uno de los rangos positivos o negativos, se les asocia el signo correspondiente.

La suma de los rangos positivos se indica por T^+, la suma de los rangos negativos se denota por T^- y el máximo valor entre estos 2 valores se escribe solamente *“T”* y se utiliza generalmente como estadístico de prueba. Si el número de diferencias es igual o mayor de 15 entonces la distribución muestral de *“T”* es aproximadamente normal por lo que se utilizará la variable parametrizada “z”. Si es menor se deberán utilizar tablas especiales que proporcionan los valores críticos para la prueba de rangos con signo.

La suma de los rangos es: $S = \frac{n(n+1)}{2}$ y deberá ser igual a $T^+ + T^-$

Las fórmulas de la media y desviación estándar de la distribución muestral *“T”* son las siguientes:

Media: $\mu_T = \frac{n(n+1)}{4}$

Desviación estándar: $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

y el estadístico de prueba es: $z = \frac{T - \mu_T}{\sigma_T}$

Ejemplo de aplicación; se desea saber si un programa de capacitación en cómputo en una empresa especializada, mejoró las habilidades de los empleados en dicha materia. Por ello se observa el nivel de habilidades antes del programa y después del programa en una muestra de 22 empleados, obteniéndose los siguientes resultados y probar la hipótesis a un nivel de significación del 1%.

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	18.5
19	43	39	-4	24	19	18.5
20	90	98	8	27	20	20
21	40	60	20	30	21	21
22	50	60	10	60	22	22

Se obtienen las diferencias de los puntajes antes y después, sus diferencias, las diferencias absolutas ordenadas, sus rangos y los rangos con signos correctos.

La suma de rangos positivos es: $T^+ = 225.5$

La suma de rangos negativos es: $T^+ = 27.5$

Comprobación: $$S = T^+ + T^- = \frac{n(n+1)}{2} = \frac{22(22+1)}{2} = 253.0$$

Por lo tanto $T = 225.5$

La hipótesis por probar son:

Ho: No hay diferencia significativa debido al tratamiento.

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 son signos correctos se 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.

Estadístico de prueba: $$z = \frac{T - \mu_T}{\sigma_T}$$

La media es: $$\mu_T = \frac{n(n+1)}{4} = \frac{22 \cdot 23}{4} = 126.5$$

La desviación estándar: $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{22 \cdot 23 \cdot 43}{24}} = 30.1$

Por lo tanto: $z = \frac{T - \mu_T}{\sigma_T} = \frac{225.5 - 126.5}{30.1} = 3.29$

Nivel de significación: $\alpha = 0.01$ por lo que $z_c = 2.33$

Como $z > z_c$ cae en la zona de rechazo, se puede concluir que el programa de capacitación de computo en esta empresa si mejoró las habilidades del personal.

5.5 Prueba de rachas

Es una prueba que se utiliza para **comprobar la aleatoriedad** de muestras. Es muy importante demostrar la aleatoriedad de las muestras en los estudios estadísticos. Si no es así se crea una gran desconfianza en los procesos de muestreo.

En una prueba de rachas, se asigna a todas las observaciones de la muestra uno o dos símbolos. Una racha se designa como una secuencia de uno o más símbolos similares y también se expresa como una serie continua de uno o más símbolos. Si el número de rachas es menor de 20, se utilizan tablas específicas en donde se muestran valores críticos mínimos y máximos por lo que si el número de rachas (*r*) es menor o excede de esos valores críticos, se indica una ausencia de aleatoriedad.

Si se tienen 2 categorías y los datos muestrales no caen en alguna de ellas, se puede utilizar la mediana como valor de referencia. Una importante aplicación de la prueba de rachas es en el método de mínimos cuadrados en el análisis de regresión. Una propiedad básica en estos modelos de regresión es que los errores son aleatorios.

Las hipótesis para probar son:

H_o : Existe aleatoriedad en las muestras.

H_1 : No existe aleatoriedad en las muestras.

Si el número de datos en 2 categorías n_1 y n_2 son mayores a 20, la distribución de muestreo para "*r*" se aproxima a una distribución normal.

Las fórmulas son:

Media de la distribución muestral del número de rachas:

$$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1$$

Desviación estándar: $$\sigma_r = \sqrt{\frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}}$$

Estadístico de prueba: $$z = \frac{r-\mu_r}{\sigma_r}$$

Ejemplo de aplicación; en una campaña a 100 posibles compradores de un producto especializado, se realizaron 52 ventas, 48 no ventas y 40 rachas. A un nivel de significación del 1% probar la hipótesis que la muestra es aleatoria.

Las hipótesis son:

H_o : La muestra es aleatoria.

H_1 : La muestra no es aleatoria.

Estadístico de prueba: $$z = \frac{r-\mu_r}{\sigma_r}$$

La media es: $$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2\cdot 52\cdot 48}{52+48} + 1 = 50.92$$

La desviación estándar:

$$\sigma_r = \sqrt{\frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right)}} = \sqrt{\frac{2 \cdot 52 \cdot 48\left(2 \cdot 52 \cdot 48 - 52 - 48\right)}{\left(52 + 48\right)^2\left(52 + 48 - 1\right)}} = \sqrt{24.67} = 4.97$$

Por lo tanto: $$z = \frac{r - \mu_r}{\sigma_r} = \frac{40 - 50.92}{4.97} = -2.20$$

Nivel de significación: $\alpha = 0.01$ por lo que $z_c = \pm 2.58$ ya que es una prueba de 2 colas. Como $z < z_c$ cae en la zona de aceptación se puede concluir que no hay evidencia suficiente para rechazar la hipótesis nula, por lo que se puede indicar que la muestra es aleatoria.

5.6 Otras pruebas

- **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 esta prueba se considerará que el estadístico de prueba se comportará como una distribución de Mann-Whitney y en ocasiones se prefiere en lugar de la "***t***" de Student debido a que la varianza de las dos poblaciones son independientes o los datos son de tipo ordinal.

Las hipótesis son las siguientes:

H_o : Las medias o medianas son iguales.

H_1 : Las medias o medianas no son iguales.

Se utilizan los estadísticos U_1 y U_2 para la primera y la segunda muestra respectivamente.

Fórmulas: $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum Rangos_1$ y $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum Rangos_2$

Media de la distribución muestral "*U*":

$$\mu_u = \frac{n_1 n_2}{2}$$

Desviación estándar.

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Estadístico de prueba:

$$z = \frac{U_i - \mu_u}{\sigma_u}$$

Veamos un ejemplo; en una empresa se está efectuando 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 nueve hombres y cinco mujeres y se les calificó en puntos el nivel de aptitud; este último varia 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), 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	

800 *	10.5			9
1 320	5		TOTAL	
1 150	7			34
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.

Realizar una prueba de hipótesis a un nivel de significación del 10%.

Cálculo de U_1 y U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum Rangos_1 \quad y \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum Rangos_2$$

que en este caso es igual a:

$$U_1 = (9)(5) + \frac{9(10)}{2} - 71 = 19 \quad y \quad U_2 = (9)(5) + \frac{5(6)}{2} - 34 = 26$$

Media de la distribución: $\mu_u = \frac{n_1 n_2}{2} = \frac{9 \cdot 5}{2} = 22.5$

Desviación estándar.

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{9 \cdot 5 (9 + 5 + 1)}{12}} = \sqrt{56.25} = 7.50$$

Estadístico de prueba:

$$z_1 = \frac{U_1 - \mu_u}{\sigma_u} = \frac{19 - 22.5}{7.50} = -0.47$$

$$z_2 = \frac{26 - 22.5}{7.50} = 0.47$$

Nivel de significación: $\alpha = 0.10$ por lo que $z_c = \pm 1.96$ ya que es una prueba de 2 colas.

Como $z_1 \; y \; z_2 < z_c$ cae en la zona de aceptación se puede concluir que no hay evidencia suficiente para rechazar la hipótesis nula, por lo que se puede indicar que la aptitud de los hombres es muy similar a las de las mujeres.

Es importante para los alumnos profundizar en el manejo de estas pruebas no parámetricas y estudiar otras más que tienen múltiples aplicaciones en la administración y economía.

Bibliografía del tema 5

BERENSON L., Mark, David LEVINE M. y Timothy Krehbiel C., *Estadística para administración*, 2ª edición, Prentice Hall, 2001.

LEVIN, Richard I. y David S. Rubin, *Estadística para administradores*, 6a. Edición, México, Prentice Hall, 1996.

MASON D., Robert, Douglas LIND A. y William MARCHAL G, *Estadística para administración y economía*, 11ª edición, Colombia, Alfaomega, , 2004.

Actividades de aprendizaje

A.5.1. A partir del estudio de la bibliografía específica sugerida, elabora un glosario para cada uno de los conceptos principales y sus aplicaciones.

A.5.2. Investiga y realiza por lo menos tres ejercicios o aplicaciones estadísticas de acuerdo con los conceptos estudiados utilizando la bibliografía sugerida.

A.5.3. Proponga ejemplos de aplicación de pruebas de bondad de ajuste a una distribución binomial.

A.5.4. Investiga y proporciona ejemplos prácticos de pruebas de bondad de ajuste a una normalidad.

A.5.5. Estudia y elabora un resumen el tema de tablas de contingencia y otros tipos de pruebas de independencia.

A.5.6. Estudia y proponga ejercicios prácticos de prueba de signo para muestras pequeñas.

A.5.7. Estudia y proponga ejercicios prácticos de prueba de U de Mann–Whitney para muestras pequeñas.

A.5.8. Investiga la prueba de Kruskal-Wallis para un análisis de varianza por rangos.

A.5.9. Visita la página www.aulafacil.com y compare los temas estudiados con la propuesta que se expresa e indique sus conclusiones.

Cuestionario de autoevaluación

1. ¿Qué pruebas son más efectivas: las paramétricas o las no paramétricas?

2. El entrenador de un equipo de ciclismo determina al azar la presión de las llantas de las bicicletas antes de la carrera. Si la presión no es correcta la registra como muy baja (B) o muy alta (A). A continuación se dan los datos. Utilice la prueba correcta para determinar, con un nivel de significancia de 0.10, si la presión de las llantas tiende a ser muy alta o muy baja, o si la ocurrencia de alta y baja se puede considerar igual.

A A B B B A A B B B B A A B B B

3. ¿Qué prueba paramétrica es similar a la prueba de U de Mann–Whitney?

4. El número de juegos de video vendidos por semana durante varias semanas se organiza en una secuencia de baja a alta; se designan por A o B, que representan a los dos vendedores clave de la compañía. El distribuidor de videos está interesado en analizar su volumen de ventas.

¿Pueden los vendedores considerarse igualmente efectivos? Pruebe con un nivel de significancia de 0.05.

A,A,B,A,A,B,B,A,A,A,A,B,B,A,A,B

A,B,A,B,B,B,A,B,A,B,B,B,A,B,B,B

5. ¿Cuál es la diferencia esencial entre los métodos estadísticos paramétricos y los no paramétricos?

6. Enumere las razones por las que elegiría un método no paramétrico para analizar datos muestrales.

7. ¿Qué prueba no paramétrica es similar a la prueba del signo de una muestra?

8. Un generador de números aleatorios genera números positivos y negativos en forma aleatoria. Después de verificar la primera serie de números, el analista piensa que la serie parece aleatoria, pero decide que debe realizarse una prueba estadística antes de usar el programa en toda la empresa. Se presenta la serie de números observada, donde P representa a un número positivo y N a uno negativo. ¿Parece ser aleatorio el programa?

a) Pruebe con un nivel de significancia de 0.5

b) Pruebe con un nivel de significancia de 0.10

9. Use la prueba de U de Mann-Whitney para determinar si hay diferencia significativa entre los valores del grupo 1 y 2; utilice nivel de significancia de 0.05

Grupo 1	Grupo 2
15	23
17	14
26	24
11	13
18	22
21	23
13	18
29	21

10. Utilice la prueba de Kruskal – Wallis para determinar si los grupos del 1 al 5 provienen de diferentes poblaciones. Utilice nivel de significancia de 0.01

1	**2**	**3**	**4**	**5**
157	165	219	286	197
188	197	257	243	215
175	204	243	259	235
174	214	231	250	217
201	183	217	279	240
203				233
				213

Examen de autoevaluación

1. **En estadística no paramétrica, una muestra es grande cuando su tamaño es mayor de:**

a. 64
b. 50
c. 40
d. 30
e. 20

2. **En una prueba de signo:**

a. Se manejan sólo valores absolutos
b. Se ignora la magnitud de la diferencia
c. Es importante la dispersión de los datos
d. Se manejan sólo números o cantidades positivas
e. Se manejan sólo números o cantidades negativas

3. **Una manufacturera automotriz desea conocer la preferencia de los clientes por los colores ocre o índigo del modelo de lujo, pues sólo uno saldrá al mercado. Se invitó a los 20 mejores vendedores para que opinaran y se encontró que doce prefirieron el color ocre, siete el índigo y uno indeciso. En un nivel del 10% probar si :**

H_0: Cualquier color gustará por igual a los clientes

H_1: Hay preferencia por alguno de los colores de los clientes

a. Se rechaza H_0 pues la región de aceptación es entre 8 y 11 signos “+”
b. Se rechaza H_0 pues la región de aceptación es entre 7 y 12 signos “+”
c. Se rechaza H_0 pues la región de aceptación es entre 6 y 13 signos “+”
d. Se acepta H_0 pues la región de aceptación es entre 5 y 14 signos “+”
e. Se acepta H_0 pues la región de aceptación es entre 4 y 15 signos “+”

4. **En un grupo piloto de 14 competidores del equipo olímpico mexicano se efectuó una prueba de "confianza en sí mismo" antes de cursar un seminario. La autoconfianza se clasificó como negativa, baja, alta y muy alta. Usando un nivel de significación del 5% y analizando la tabla diga si el seminario ayuda a mejorar la autoconfianza:**

H_0: p=0.50

H_1: p>0.50

Nombre	Antes del curso	Después del curso
Elizabeth	Negativa	Alta
Luisa	Baja	Muy alta
Mario	Baja	Alta
René	Negativa	Baja
Cristina	Baja	Alta
Eloísa	Negativa	Baja
Arturo	Baja	Alta
Luis	Baja	Muy alta
Xóchitl	Negativa	Baja
Mónica	Negativa	Negativa
Jaime	Baja	Alta
Soledad	Muy alta	Baja
Estrella	Baja	Alta
Francisco	Baja	Alta

a. Se acepta H_0 pues hubo 9 signos "+"
b. Se rechaza H_0 pues hubo 10 signos "+"
c. Se rechaza H_0 pues hubo 11 signos "+"
d. Se rechaza H_0 pues hubo 12 signos "+"
e. Se rechaza H_0 pues hubo 13 signos "+"

5. **Para el aniversario de la empresa se organizó una convención y se dio a escoger entre el menú tradicional o uno especial. La muestra fue de 81 clientes de los cuales 42 prefirieron el especial. Utilizando la prueba del signo y un nivel de 0.02, pruebe si a los clientes les gustó más el menú especial que el tradicional:**

H_0: Ambos menús gustaron por igual (p=0.50)

H_1: Gustó más el menú especial (p>0.50)

a. Se acepta H_0 pues Zc es 0.44
b. Se rechaza H_0 pues Zc es 1.03
c. Se acepta H_0 pues Zc es 1.69
d. Se rechaza H_0 pues Zc es 0.23
e. Se acepta H_0 pues Zc es 1.45

6. **El sindicato de taxistas afirma que la mediana del recorrido semestral de cada unidad es de 40,000 km. Sin embargo, esto es rebatido por algunos dirigentes que afirman que la mediana es mayor; por lo tanto, se tomó una muestra aleatoria de 205 taxis y se encontró que 170 recorren una mayor distancia; 5 recorren 40,000 km. y los restantes menos de 40,000 km. Utilizando la prueba del signo y un nivel de 0.05, pruebe si:**

H_0: La mediana semestral es igual a 40,000 km.

H_1: La mediana semestral es mayor de 40,000 km

a. Se acepta H_0 pues la Zc es 1.509
b. Se rechaza H_0 pues la Zc es 3.748
c. Se rechaza H_0 pues la Zc es 10.540
d. Se acepta H_0 pues la Zc es 0.841
e. Se rechaza H_0 pues la Zc es 6.492

7. **En dos equipos de vendedores de puerta en puerta se analizaron los resultados de su entrenamiento. Las puntuaciones del equipo "A" son: 186, 212, 97, 141, 160, 122, 180 y 121; las del grupo "B" son: 147, 99, 167, 126, 180, 197 y 128. Como la población de puntuaciones no se distribuye normalmente, use la prueba de Mann-Whitney con un nivel de significación de 5% y determine el valor crítico de la prueba:**

H_0: No hay diferencia entre los dos grupos

H_1: Existe una diferencia entre ambos grupos

a. 7
b. 8
c. 9
d. 10
e. 11

8. **En dos equipos de vendedores de puerta en puerta se analizaron los resultados de su entrenamiento. Las puntuaciones del equipo "A" son: 186, 212, 97, 141, 160, 122, 180 y 121; las del grupo "B" son: 147, 99, 167, 126, 180, 197 y 128. Como la población de puntuaciones no se distribuye normalmente, use la prueba de Mann-Whitney con un nivel de significación de 5%:**

H_0: No hay diferencia entre los dos grupos

H_1: Existe una diferencia entre ambos grupos

a. Se rechaza H_0 pues la U calculada es 15.4
b. Se rechaza H_0 pues la U calculada es 17.8
c. Se acepta H_0 pues la U calculada es 21.9
d. Se acepta H_0 pues la U calculada es 24.6
e. Se acepta H_0 pues la U calculada es 27.5

9. **En la Olimpiada de Matemáticas compitieron 20 ingenieros contra 15 actuarios, las calificaciones fueron las siguientes: ingenieros: 35, 11, 8, 21, 4, 18, 15, 36, 13, 28, 31, 23, 5, 9, 27, 29, 37, 2, 1, 20, 19, 16, 39, 33 y 7; actuarios: 12, 34, 32, 25, 22, 3, 40, 17, 30, 24, 14, 10, 38, 6 y 26. Use la prueba de Mann-Whitney con un nivel de significación de 0.05 y demuestre si hay alguna diferencia en su aptitud hacia las matemáticas de alguno de los grupos:**

H_0: No hay diferencia entre ingenieros y actuarios
H_1: Hay una diferencia significativa de alguno

a. Se acepta H_0 pues la Zc es –0.42
b. Se acepta H_0 pues la Zc es –0.71
c. Se acepta H_0 pues la Zc es 1.21
d. Se rechaza H_0 pues la Zc es 2.72
e. Se rechaza H_0 pues la Zc es –0.35

10.En los botes pesqueros se desea probar la efectividad de una nueva técnica. Así que al tomar una prueba con 11 botes los resultados fueron los siguientes:

No.	Normal	Nueva
1	15	25
2	10	28
3	16	16
4	10	22
5	20	19
6	17	20
7	24	30
8	23	26
9	17	18
10	21	23
11	25	22

Aplicando la prueba de Wilcoxon de rangos con signo y un nivel de significación del 5% probar si la nueva técnica incrementa o no la pesca:

H_0: Los volúmenes de pesca no cambian con la nueva técnica

H_1: Los volúmenes de pesca mejoran con la nueva técnica

a. Se acepta H_0 pues la T calculada es 18.1
b. Se acepta H_1 pues la T calculada es 6.5
c. Se acepta H_0 pues la T calculada es 2.7
d. Se acepta H_1 pues la T calculada es 5.1
e. Se acepta H_0 pues la T calculada es 7.9

Tema 6. Análisis de regresión simple y correlación

Objetivo particular

El alumno analizará los conceptos fundamentales de regresión simple y correlación, su desarrollo y aplicación dentro del ámbito empresarial.

Temario detallado

6. Análisis de regresión simple y correlación

- 6.1 Modelo lineal simple
- 6.2 Método de mínimos cuadrados
- 6.3 Inferencias relativas a la pendiente de la recta de regresión
- 6.4 Predicción de un valor particular de “y” para un valor dado de “x”
- 6.5 Coeficiente de correlación y coeficiente de determinación
- 6.6 Inferencias relativas al coeficiente de correlación

Introducción

El uso de la regresión lineal simple es muy utilizado para observar el tipo de relación que existe entre dos variables y poder llevar a cabo la toma de decisiones correspondiente dependiendo de la relación entre dichas variables, así por ejemplo pudiera darse el caso en el que después de aplicar la regresión lineal no exista relación entre las variables involucradas y en consecuencia la decisión podría ser buscar cual es la variable independiente que tiene influencia sobre la dependiente y volver a realizar el estudio completo; pero si fuera el caso en el cual si existiera una relación positiva entre las variables involucradas, la obtención del coeficiente de correlación nos daría más información sobre el porcentaje de relación existente y pudiendo determinar si es necesario la inclusión de otra variable independiente en el problema mismo, para lo cual el análisis de regresión ya sería del tipo múltiple.

6.1 Modelo lineal simple

El **análisis de regresión lineal o bivariada**[27] 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. 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**.

- **Naturaleza de la relación**[28]

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 exista, 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.

[27] Carl McDaniel, y Roger Gates, *Investigación de mercados contemporánea*, p. 558.
[28] *Op. Cit*

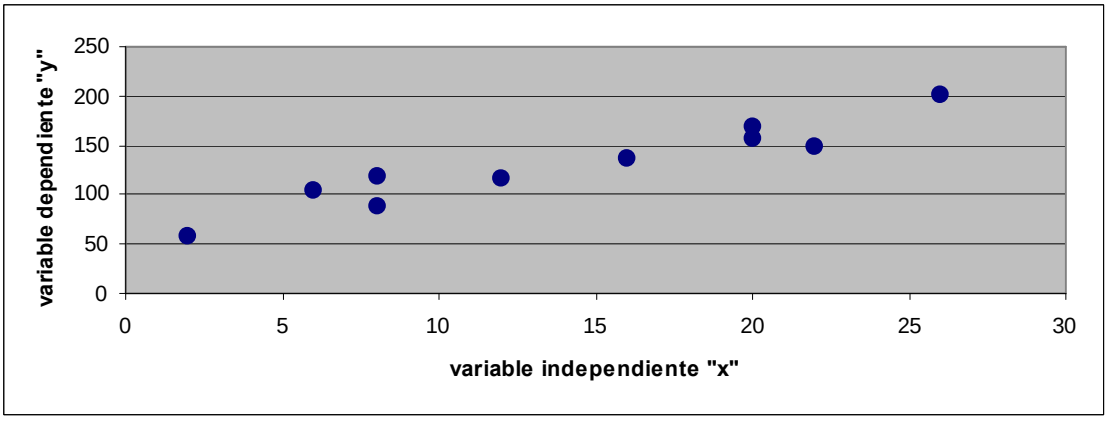


Figura 6.1 Diagrama de dispersión que muestra la relación entre la variable dependiente e independiente

6.2 Método de mínimos cuadrados

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**.

El método de **mínimos cuadrados**, es un procedimiento para encontrar la ecuación de regresión que se origina al estudiar la relación estocástica que existe entre dos variables. Fué Karl Friedrich Gauss (1777-1855) quien propuso el método de los mínimos cuadrados y fue el primero en demostrar que la ecuación estimada de regresión minimiza la suma de cuadrados de errores.

En el análisis de regresión[29], 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"**. Por el contrario, una variable cuyo valor se suponga desconocido y que se explique o prediga con ayuda de otra se llama **variable dependiente** y se simboliza por "**Y"**.

[29] Heinz Kohler, *Estadística para negocios y economía*, pp. 528-529.

Una **relación estocástica**[30] 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 cometidos cuando la ecuación de esa línea se utilice para estimar **y** a partir de ***x***.

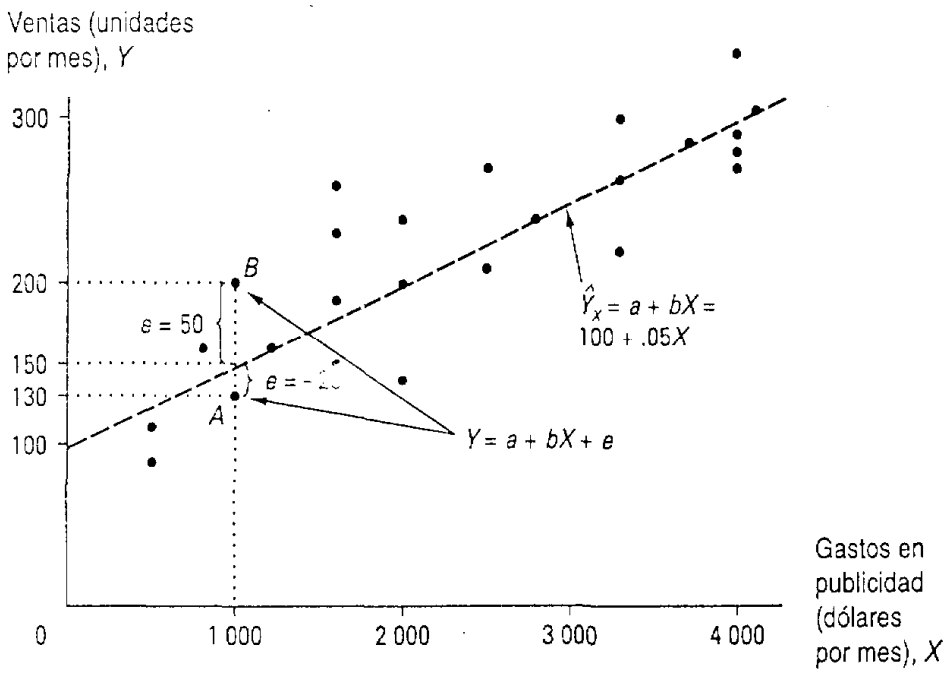


Figura 6.2 Grafica que muestra la relación existente entre los gastos de publicidad y las ventas.

De esta gráfica podemos ver claramente que las ventas dadas en unidades por mes (variable dependiente) en este caso, si guardan relación con los gastos en publicidad y, que dicha relación puede ser denotada por la "recta de regresión"

De este análisis de relación estocástica que se da entre dos variables, surgen las ecuaciones que nos provee el método de mínimos cuadrados, que a saber son:

[30] Heinz Kohler, *Estadística para negocios y economía*, p. 530.

Ecuación de la recta de regresión: $\hat{y}_i = b_0 + b_1 X_i$

En la que:

- $\mathbf{x_i}$ = es un valor dado de la variable independiente para el cual se quiere estimar el valor correspondiente de la variable dependiente
- $\mathbf{b_0}$ = ordenada al origen de la línea estimada de regresión,
- $\mathbf{b_1}$ = pendiente de la línea estimada de regresión,
- $\mathbf{\hat{Y}_i}$ = valor estimado de la variable dependiente, para el i-ésimo valor de la variable independiente

Resulta claro que para poder determinar la recta de regresión, es necesario que antes sean calculados los valores correspondientes a la pendiente de la recta y a la ordenada al origen.

La pendiente de la recta de regresión se calcula mediante la siguiente formula:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

y la ordenada al origen se calcula mediante la formula:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Antes de continuar, 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. Por ejemplo,

un estadista puede llegar a determinar que la relación entre las ventas y el presupuesto asignado a mercadotecnia es positiva y que se tiene un coeficiente de correlación de 0.96, lo cual prácticamente nos indica que es recomendable incrementar el presupuesto al departamento de mercadotecnia para obtener mejores ingresos dentro de la compañía, sin embargo el director de operaciones puede llegar a determinar que debido a condiciones internas del país en el que se encuentre la empresa, o bien la aparición de una nueva ley que regule los medios utilizados por el mencionado departamento de mercadotecnia, pueden llegar a frenar o incluso generar conflictos dentro de la empresa si incrementamos el presupuesto al departamento correspondiente.

a. Inferencias relativas a la pendiente de la recta de regresión

Las inferencias acerca de la pendiente de la recta de regresión son importantes dado que la relación entre las dos variables en cuestión depende de ella precisamente, es decir, si la pendiente de la recta de regresión es positiva, entonces la naturaleza de la relación entre ambas variables será positiva, y la pendiente de la recta es negativa, entonces la relación entre las variables será negativa también, con lo cual podemos iniciar la toma de decisiones dependiendo del contexto del problema mismo. Como se menciono anteriormente la ecuación de la recta deregresión: $\hat{y}_i = b_0 + b_1 X_i$

b_0 representa la ordenada al origen de la línea estimada de regresión, y

b_1 es la pendiente de la línea estimada de regresión.

Donde **b_0** es en sí, el punto donde la recta corta al eje de las “x” y **b_1** nos da el grado de inclinación de la recta, de tal forma que cuando la pendiente de la recta es positiva, se dice que la relación que existe entre las dos variables dependiente e independiente es de naturaleza positiva, es decir, que posee una grafica como la

indicada a continuación:

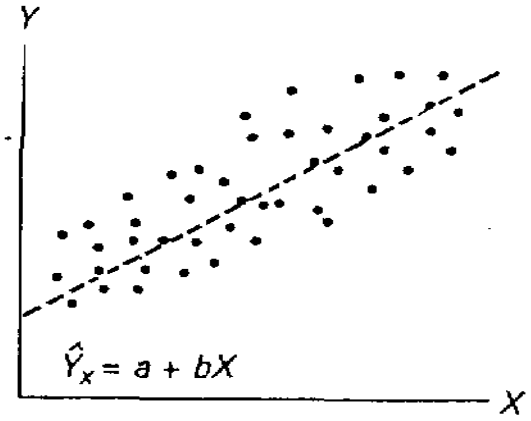


Figura 6.3 Relación positiva entre dos variables en regresión lineal

En este tipo de relación, los incrementos en los valores de la variable independiente traen como consecuencia un incremento en los valores correspondientes de la variable dependiente y la grafica tiene como podemos apreciar una forma ascendente.

Pero cuando la pendiente de la recta de regresión es negativa, es decir, que dicha ecuación tuviera la forma $\hat{y}_i = b_0 - b_1 X_i$ entonces la relación existente entre las variables es de tipo negativa, lo cual quiere decir, que a incrementos en los valores de la variable independiente, la variable dependiente responde con decrementos; la grafica resultante tendría la forma siguiente:

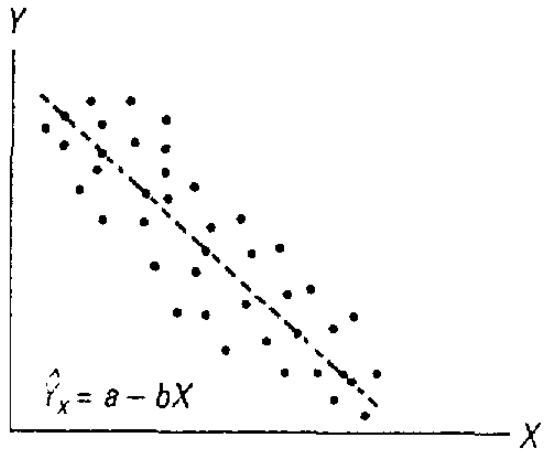


Figura 6.4 Relación negativa entre dos variables en regresión lineal.

En esta grafica podemos observar que la tendencia de la recta de regresión es descendente, lo cual implica como ya habíamos mencionado, que la relación entre ambas variables es negativa.

6.4 Predicción de un valor particular de “y” para un valor dado de “x”

Para predecir un valor particular de “y” para un valor dado de “x” utilizaremos el análisis de regresión simple y para ilustrarlo consideremos el siguiente ejemplo: Domin’s Pizza es una cadena de restaurantes dedicado exclusivamente a la distribución de pizzas. El director general cree que los lugares donde sus establecimientos han tenido más éxito están cercanos a establecimientos de educación superior y para sustentar su creencia ha solicitado un estudio de las ventas de sus restaurantes asociadas al tamaño de la población estudiantil de los centros educativos correspondientes.

Los administradores creen que las ventas mensuales 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.

Para ilustrarlo, supongamos que en el caso de Domin's Pizza se reunieron datos de una muestra de 10 restaurantes ubicados cerca de centros educativos.

Para el iésimo restaurante de la muestra, x_i es el tamaño de la población estudiantil, en miles, y y_i son las ventas mensuales, 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:

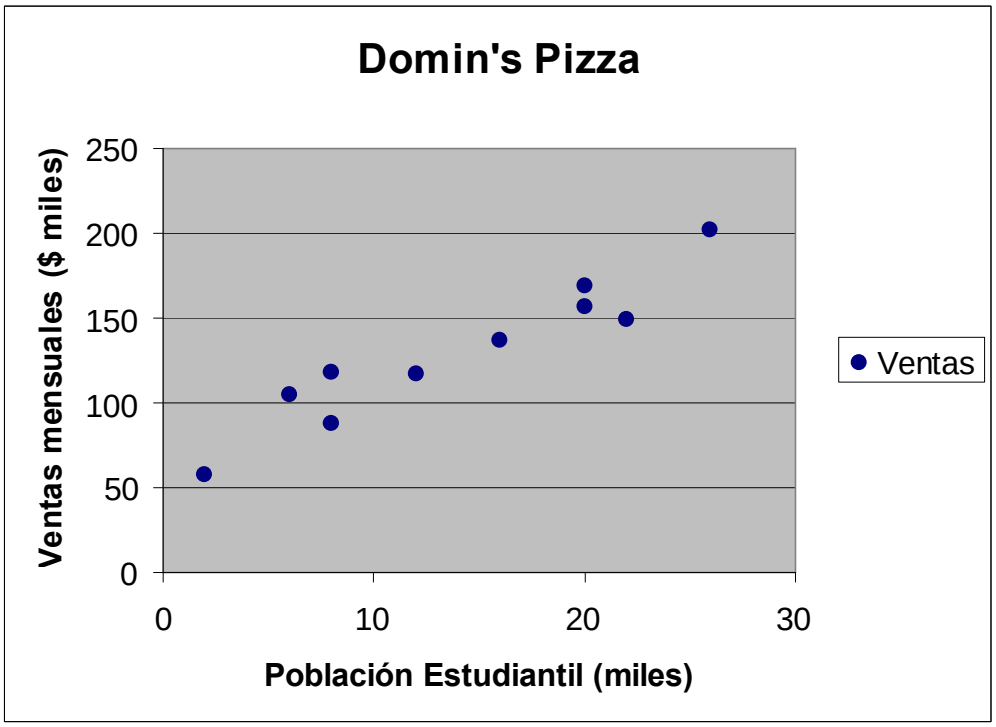
RESTAURANTE I	POBLACIÓN DE ESTUDIANTES x_i (MILES)	VENTAS MENSUALES y_i ($ MILES)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Datos de población estudiantil y ventas trimestrales para 10 restaurantes de Domin's Pizza.

Lo primero que hay que hacer para resolver un problema de regresión lineal, es definir nuestras variables involucradas (variable dependiente y variable independiente); en este caso particular los administradores de Domin`s Pizza ya lo han hecho, definiendo las ventas mensuales de los restaurantes como la variable dependiente y la población estudiantil dada en miles como la variable independiente, debido a que claro esta se supone que las ventas de los restaurantes dependen de la población estudiantil de los centros educativos.

La siguiente gráfica corresponde al diagrama de dispersión de los datos de la tabla

anterior:



¿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 iésimo 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 iésimo 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 mensuales, en miles, para el pésimo restaurante.

Para desarrollar nuestros cálculos de manera más sencilla, complementamos la tabla de inicio del problema, misa que quedaría de la siguiente forma:

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

Por lo tanto, calculando la pendiente de la recta tenemos que la formula es:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

donde al sustituir los datos obtenidos de la tabla anterior:

$$b_1 = \frac{(21040) - \frac{(140)(1300)}{10}}{(2528) - \frac{(140)^2}{10}}$$

y al realizar los cálculos pertinentes tenemos que:

$$b_1 = 5$$

calculando ahora la ordenada al origen tenemos que la formula es:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

en este caso podemos ver que nos falta el valor de la media en “Y” y el valor de la media de “X”, mismas que tienen un valor de:

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

donde al sustituir datos tenemos que:

$$\overline{X} = \frac{140}{10}$$

es decir:

$$\overline{X} = 14$$

y para la media de “Y” tenemos que la formula es:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

donde la sustituir datos vemos que:

$$\overline{Y} = \frac{1300}{10}$$

es decir:

$$\overline{Y} = 130$$

por lo tanto, sabiendo que:

$$b_1 = 5$$

$$\overline{X} = 14$$

$$\overline{Y} = 130$$

podemos sustituir estos valores en la formula:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

obteniendo:

$$b_0 = 130 - (5)(14)$$

finalmente realizando las operaciones indicadas:

$$b_0 = 60$$

por lo tanto, la ecuación de la recta de regresión sería:

$$\hat{y}_i = b_0 + b_1 X_i$$

donde al sustituir valores:

$$\hat{y}_i = 60 + 5 X_i$$

Esta recta de regresión nos sirve para predecir cuales serían las ventas mensuales en función del tamaño de la población estudiantil. 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

$$\hat{y}_i = 60 + 5(30)$$

$$\hat{y}_i = 210$$

es decir, las ventas estimadas para ese restaurante serían de $210,000.00 mensuales (recuerde que las ventas están dadas en miles).

Resulta claro que la predicción de las ventas en función de la cantidad de personas es importante porque me ayuda a tomar decisiones que van desde el importe de renta que puedo pagar por el local donde vaya a ubicar la Pizzería, hasta programar mi presupuesto de insumos, la cantidad de personal del equipo de reparto, etc.

6.5 Coeficiente de correlación y coeficiente de determinación

El coeficiente de determinación se utiliza para evaluar la bondad de ajuste para la ecuación de regresión y se define como:

$$r^2 = \frac{\text{Suma de Cuadrados de la regresión}}{\text{Suma de cuadrados Totales}} = \frac{SSR}{SST}$$

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n}(Y - \bar{Y}_i)^2}$$

La ecuación anterior se puede interpretar como el porcentaje de variación de la variable Y que se puede explicar con el modelo de regresión; en el ejemplo de las pizzas, r^2=0.9027 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.

- **Coeficiente de correlación**

Cuando es necesario resumir aún más los datos (de una gráfica por ejemplo) se utiliza un solo número, que de alguna forma mide la fuerza de asociación entre dos variables como son el ingreso real y el nivel de educación escolar en nuestro caso. El análisis de correlación nos ayuda a obtener dicho número que se conoce como: **coeficiente de correlación**. Los valores de coeficiente de correlación siempre están entre –1 y +1 un valor de +1 indica que las dos variables tienen una relación lineal positiva perfecta. Esto es, todos los puntos de datos están en una línea recta con pendiente positiva. Un valor de –1 indica que las variables tienen una relación lineal negativa perfecta, y que todos los puntos de datos están en una recta con pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que las variables no tienen relación lineal[31].

A continuación presentamos la ecuación para calcular el coeficiente de correlación de la muestra. Si ya se ha hecho un análisis de regresión y se ha calculado el coeficiente de determinación, entonces, el coeficiente de correlación se puede calcular como sigue:

$$r = (signo\ de\ b_1)\sqrt{r^2}$$

donde b1 es la pendiente de la ecuación de regresión.

De esta formula, resulta claro que el signo del coeficiente de correlación es positivo si la ecuación de regresión tiene pendiente positiva (b1 > 0), y negativo si la ecuación de regresión tiene pendiente negativa (b1 < 0).

En nuestro ejemplo de las pizzas tendríamos que:

$$r = (signo\ de\ b_1)\sqrt{r^2}$$

$$r = +\sqrt{0.9027}$$

$$r = +0.9501$$

[31] Anderson, Sweeney & Williams, 1999. *Estadística para administración y economía*, p.p. 555.

6.6 Inferencias relativas al coeficiente de correlación

Cuando se tiene una relación lineal entre dos variables, el coeficiente de determinación y el coeficiente de correlación permiten tener medidas de la intensidad de una relación. El coeficiente de determinación da una medida entre 0 y 1, mientras que el coeficiente de correlación da una medida entre –1 y +1 aunque el coeficiente de correlación se restringe a una relación lineal entre dos variables, el coeficiente de determinación se puede emplear en relaciones no lineales y en relaciones que tengan dos o más variables independientes. En ese sentido, el coeficiente de determinación tiene una aplicabilidad más amplia.

Obtener una ecuación que describa la relación existente entre dos variables en un análisis de regresión lineal es muy importante, descubrir que esa relación es positiva, negativa o inexistente también lo es, pero tener un indicador que nos diga que tan intensa es la relación si que la hay entre las variables en cuestión, no deja de ser importante, además de complementar y sustentar las decisiones que se deriven del análisis de regresión.

Como podemos apreciar en el desarrollo del tema, el análisis de regresión es una herramienta matemática muy importante que nos ayuda a la mejor toma de decisiones en un ámbito como el actual lleno de cambios y de una competencia muy cerrada donde una respuesta rápida a los cambios presentados por el medio empresarial, por el mercado, etc. Puede representar la aparición de nueva competencia o bien la extinción de las empresas.

Bibliografía del tema 6

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill.

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica.
WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.6.1. Con la bibliografía del tema estudia y elabora un cuadro de la relevancia que tiene el coeficiente de correlación contra el coeficiente de determinación.

A.6.2. Identifica en el contexto social en el que vive algún problema en el cual se pueda aplicar la regresión lineal.

A.6.3. Compara el método de mínimos cuadrados en los diferentes libros de la bibliografía del tema y elabora un resumen.

A.6.4. Estudia los ejercicios resueltos de libro *Estadística* de Weimer sobre regresión lineal.

A.6.5. Estudia y compara los conceptos y aplicaciones de los coeficientes: de correlación y de determinación en los diferentes libros de la bibliografía del tema.

A.6.6. Averigua algunas investigaciones que utilizaron el análisis de regresión simple.

A.6.7. Averigua que tipo de investigaciones utilizan el análisis de correlación.

Cuestionario de autoevaluación

1. Diga qué es el análisis de regresión lineal o bivariada.
2. ¿Cuándo se aplica la regresión múltiple?
3. ¿Qué es el método de los mínimos cuadrados?
4. ¿Quién propuso el método de los mínimos cuadrados?
5. ¿Qué es el coeficiente de determinación?
6. ¿Cuál es el rango del coeficiente de determinación?
7. ¿Qué es el coeficiente de correlación?
8. ¿Cuál es el rango del coeficiente de correlación?
9. ¿Quién desarrolló por primera vez los métodos estadísticos para el estudio de la relación entre dos variables?
10. ¿Es el análisis de regresión un procedimiento para establecer una relación de

causa y efecto?

Examen de autoevaluación

1. 1 **Considere que el departamento de recursos humanos de la empresa en la que laboramos está interesada en saber si el monto del salario guarda alguna relación directa con la el ahorro voluntario que los empleados sindicalizados de la empresa presentan. Para ello la empresa ha tomado una muestra aleatoria de 10 empleados quedando los datos obtenidos en la siguiente tabla:**

Sueldo del empleado	8000	7000	6500	9200	6000	12000	10300	8700	7500	6250
Ahorro del empleado	4000	2000	3200	4500	1200	1000	2500	1500	1700	2250

¿Para este problema la recta de regresión considerando el sueldo del empleado como variable independiente es?:

a) Y = -2607.25 + 0.58484 Xi

b) Y = -3500.50 + 0.51831 Xi

c) Y = 2687.23 – 0.03711 Xi

d) Y = 5000 + 0.12581 Xi

2. Para el problema de la pregunta 1, el coeficiente de determinación es:

a) 0.33985986

b) 0.00369796

c) –0.3398477

d) 2.45768779

3. Para el problema de la pregunta 1, el coeficiente de correlación es:

a) 9.4372838

b) 2.5465758

c) –1.247574

d) –0.060811

4. Para el problema 1 podemos decir que los empleados:

a) entre más ganan mas ahorran

b) entre más ganan menos dinero ahorran

c) ahorran de manera indiferente

d) no existe un patrón de ahorro en función del salario

5. El pronóstico de ahorro para un empleado que gana $15,000.00 será de:

a) 3243.83

b) 2543.83

c) 4243.83

d) 6243.83

6. Una tienda departamental, está considerando otorgar tarjetas de crédito a sus cliente, para lo cual realiza un estudio con el fin de observar el comportamiento de sus gastos en función de su salario. Los datos obtenidos en una muestra aleatoria de tamaño 11 se encuentran en la siguiente tabla:

Sueldo del cliente	18.0	15.0	19.0	9.2	8.6	12.0	10.7	14.3	17.8	16.0	15.0
Gastos del cliente	14.8	10.4	15.7	7.1	5.3	8.0	8.5	10.2	13.0	14.0	11.3

Nota: tanto el sueldo como los gastos del cliente son mensuales y están dados en miles de pesos.

¿Para este problema la recta de regresión considerando el sueldo del cliente como variable independiente es?:

a) Y = -1.3223 - 0.60253 Xi

b) Y = 1.2332 + 0.25360 Xi

c) Y = -2.2332 – 0.60253 Xi

d) Y = 1.3223 + 0.60253 Xi

7. Para el problema de la pregunta 6, el coeficiente de determinación es:

a) 0.37923

b) 2.10045

c) –1.2678

d) 3.42567

8. Para el problema de la pregunta 6, el coeficiente de correlación es:

a) 4.218923

b) –3.34567

c) 0.615817

d) –0.61587

9. Para el problema 6 podemos decir que los empleados:

a) entre más ganan más gastan

b) entre más ganan menos dinero gastan

c) gastan de manera indiferente

d) no existe un patrón de gasto en función del salario

10. El pronóstico de gasto para un cliente que gana $21,000.00 será de:

a) 18000.23

b) 15000.50

c) 11330.76

d) 14325.79

Tema 7. Series de tiempo

Objetivo particular

El alumno analizará el concepto de series de tiempo y su aplicación en el entorno profesional administrativo y contable.

Temario detallado

7. Series de tiempo

7.1 Análisis de tendencias
7.2 Variación cíclica
7.3 Variación temporal
7.4 Variación irregular
7.5 Análisis de predicciones.

Introducción

Una serie de tiempo es el conjunto de datos que se registran a través del tiempo sobre el comportamiento de una variable de interés, generalmente los registros se realizan en periodos iguales de tiempo.

Las series de tiempo resultan especialmente útiles cuando se requiere realizar un pronóstico sobre el comportamiento futuro que puede tener una variable determinada, imaginemos por ejemplo la necesidad de tomar una decisión sobre el comportamiento a futuro de la demanda, el precio y las ventas de un producto, los ingresos en el próximo año, los precios de bienes y servicios, los valores de los energéticos, etc. En todas estas situaciones resulta útil el análisis de las series de tiempo que los representan, bajo la hipótesis de que los factores que han influenciado su comportamiento en el pasado, estarán presentes de manera similar en el futuro. De esta manera, el objetivo principal del conocimiento de las series de tiempo es la identificación de los factores que intervienen y la separación de cada uno de ellos, con el fin de pronosticar cuál será el comportamiento en el futuro.

7.1 Análisis de tendencias

Generalmente los datos de una serie de tiempo pueden contener de manera implícita cuatro elementos, Tendencia (T), ciclos (C), estacionalidad (E), y una componente irregular (I), aunque no siempre están presentes todos ellos.

Para aislar y comprender los elementos de una serie, es necesario descomponerla, al hacerlo, se puede obtener una mejor idea del comportamiento de cada uno de sus elementos facilitándose el pronóstico, para llevar a cabo la separación, deben tenerse en cuenta las relaciones matemáticas que los unen, uno de los modelos mas utilizados para descomponer una serie de tiempo es el llamado **modelo multiplicativo**, en el cual se supone que la serie es el resultado del producto de sus cuatro componentes, el modelo se establece mediante la expresión siguiente:

$$\mathbf{Y = T} \; x \; \mathbf{C} \; x \; \mathbf{E} \; x \; \mathbf{I}$$

La interpretación de cada uno de los componentes es la siguiente:

Tendencia (**T**). La tendencia es la componente que representa el comportamiento (crecimiento o decrecimiento), en un periodo largo de tiempo. Generalmente se puede representar como una línea recta o curva, el valor de la pendiente indicará el sentido de dicho crecimiento.

Ciclo (**C**). La componente cíclica 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.

Estacionalidad (**E**). La 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.

Componente irregular (**I**). Es la componente que queda después de separar a las otras componentes, es el resultado de factores no explicables que siguen un comportamiento aleatorio, siendo por ello una parte no previsible de la serie.

Ejemplo:

Supongamos que tenemos la información siguiente, correspondiente al comportamiento del rendimiento de los Certificados de la Tesorería, denominados CETES a 90 días, el tiempo está expresado en trimestres y el valor de la variable en valores de la tasa de interés que ganan en cada trimestre.

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 representa 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:

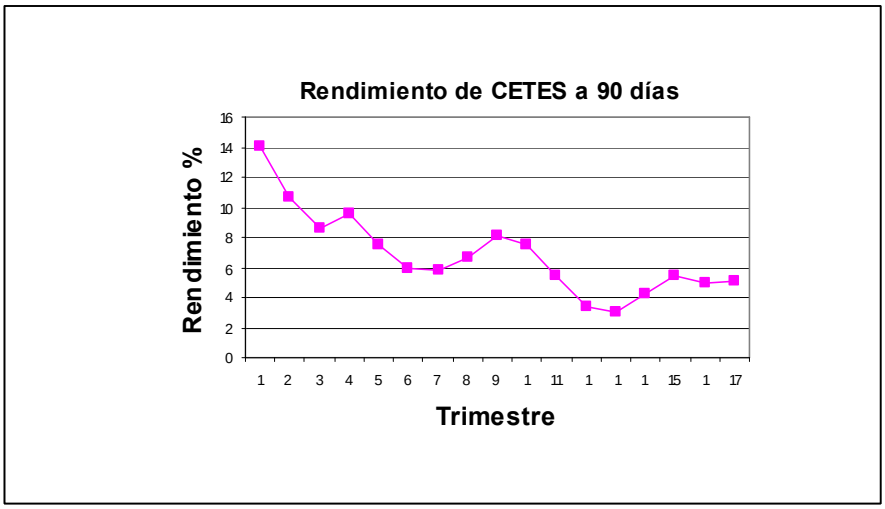


Figura 7.1. Rendimiento de los certificados de la tesorería a 90 días.

Utilizando el ejemplo anterior procederemos a descomponer la serie de tiempo en cada uno de sus componentes, lo cual haremos en los siguientes incisos.

La separación de la tendencia, utiliza la metodología de la línea de regresión, hemos mencionado que esta línea puede ser una recta o una curva, en este curso únicamente analizaremos el modelo lineal, por su simpleza y facilidad de cálculo, de esta manera podemos representar a la tendencia por medio de la expresión matemática siguiente:

$$\mathbf{Yt} = b_o + b_1\mathbf{X}$$

En donde:

Yt tasa de rendimiento calculada

X tiempo, en este caso expresado en trimestres

b_o valor de Y cuando el valor del tiempo es cero

b_1 pendiente de la recta de tendencia

Una vez definido el modelo, se procede a la determinación de los valores de los coeficientes **b_o** y **b_1** de la recta de regresión. En nuestro problema en particular, la ecuación de regresión, que representa a la tendencia del comportamiento de la tasa de rendimiento de los CETES a 90 días aplicando las formulas correspondientes para el calculo primero de "b_1"

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

y posteriormente para el calculo de "b_0"

$$b_0 = \overline{Y} - b_1 \overline{X}$$

es:

Yt = 10.8553676 - 0.44595588 **X**

Además, aplicando las formulas correspondientes primero al calculo del coeficiente de determinación:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

y finalmente al calculo del coeficiente de correlación:

$$r = (signo\ de\ b_1)\sqrt{r^2}$$

tenemos que el valor del coeficiente de correlación es de r = -0.8078, lo que nos indica que el ajuste logrado con la recta de regresión es adecuado, recordemos que el coeficiente de correlación es una medida de la precisión lograda en el ajuste, valores del coeficiente de correlación iguales a +1 ó -1 son la indicación de un ajuste perfecto, un valor igual a cero nos dirá que este no existe. (nota: se deja al estudiante corroborar los valores obtenidos de "b_1", "b_0" y "r")

Una vez definida la ecuación de la recta de tendencia, es posible compararla gráficamente con los valores de la serie, como se muestra en la gráfica siguiente (Figura 7.2), en ella podemos observar que la tendencia de las tasas de rendimiento es descendente, el signo del coeficiente b1, que representa la pendiente de la recta, ya nos lo había indicado. También podemos observar que son evidentes valores por arriba y por debajo de esta línea, estos representan a los valores cíclicos de la serie.

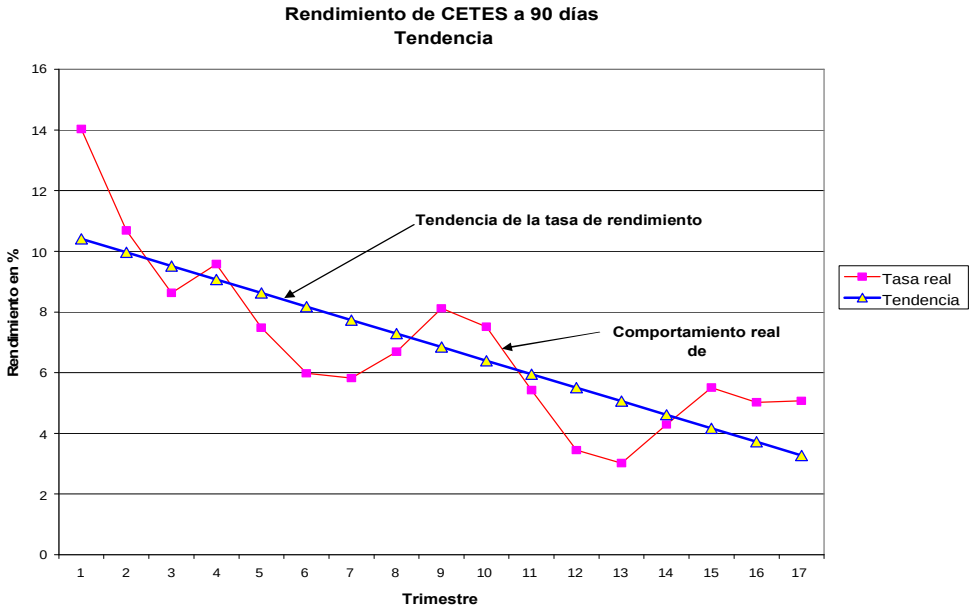


Figura 7.2 Gráfica de comparación de la recta de tendencia contra el comportamiento real de los CETES a 90 días.

En el análisis de tendencias podemos ver clara y rápidamente mediante el cálculo de la pendiente de la recta de regresión (b_1) si la tendencia de la variable de medición (en nuestro caso en particular “el rendimiento de los CETES a 90 días) es a la baja (pendiente negativa), a la alza (pendiente positiva) o a mantenerse sin variación (pendiente cero); lo cual dentro del análisis de la serie de tiempo, es muy importante.

7.2 Variación cíclica

Las **fluctuaciones de los valores de rendimientos** alrededor de la línea de tendencia, constituyen la componente cíclica, estas están son el resultado de la ocurrencia de fenómenos que pueden tener origen social, económico, político, costumbres locales, etc, pero que pueden afectar el comportamiento de la variable, de ahí que su separación resulte importante.

Supongamos ahora que nos interesa conocer la variación que han tenido los rendimientos respecto de la tendencia, es decir la **componente cíclica**, la cual queda representada en la gráfica (Figura 7.3) por los valores mayores y menores respecto de la tendencia. Si deseamos conocer el valor numérico de este comportamiento debemos proceder como sigue:

Calcular para cada trimestre 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.

Los valores así calculados se muestran en la tabla siguiente, expresados en porcentaje respecto del valor de la tendencia, los valores que estén por encima de la recta de tendencia alcanzarán un porcentaje superior a cien, mientras que los que se encuentren por debajo de ella tendrán valores inferiores a cien.

Trimestre	**Rendimiento**		**Componente cíclica %**
	Real **Y**	Tendencia **Yc**	
1	14.03	10.41	134.78
2	10.69	9.96	107.29
3	8.63	9.52	90.68
4	9.58	9.07	105.60
5	7.48	8.63	86.72
6	5.98	8.18	73.11
7	5.82	7.73	75.26
8	6.69	7.29	91.80
9	8.12	6.84	118.68
10	7.51	6.40	117.42
11	5.42	5.95	91.09
12	3.45	5.50	62.68
13	3.02	5.06	59.71
14	4.29	4.61	93.02
15	5.51	4.17	132.26
16	5.02	3.72	134.94
17	5.07	3.27	154.85

Cuadro 7.1 Valores de la componente cíclica

Las componentes cíclicas, pueden ser graficados para observar los posibles patrones que se presentan, la línea de la tendencia corresponde en la gráfica a la línea del 100%, observemos que la variación cíclica se presenta hacia arriba y hacia abajo de la recta de tendencia.

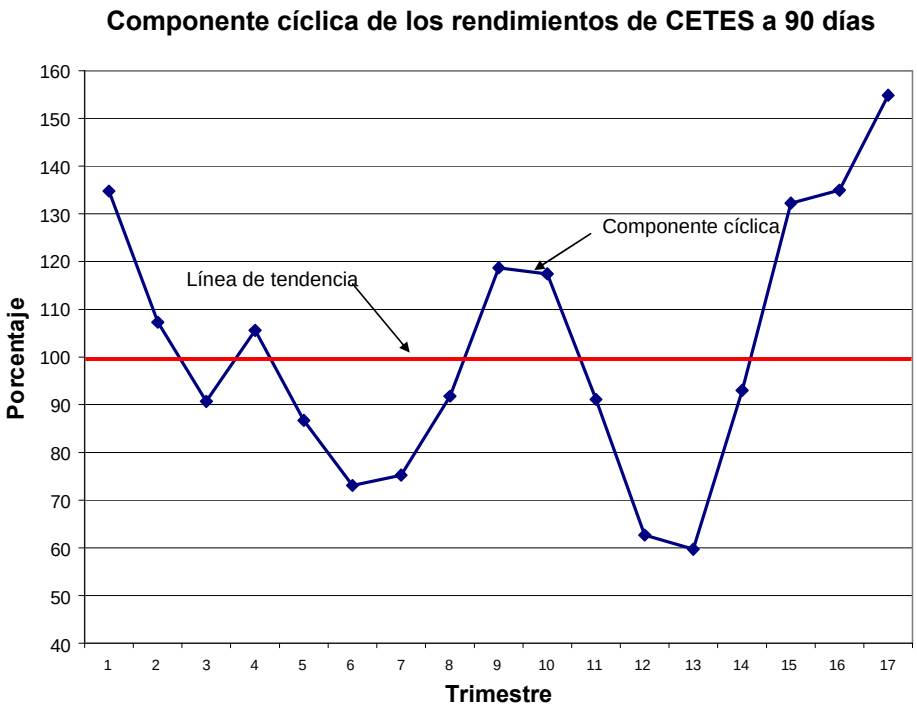


Figura 7.3 Gráfica de apreciación de la componente cíclica de los CETES a 90 dias.

Es posible ver con mucha claridad cual ha sido el comportamiento de los rendimientos respecto de la tendencia. Podemos observar que las fluctuaciones a la baja han sido más importantes que las correspondientes a la alza. Esto muy importante, pues si alguna persona compro CETES a 90 días durante el primer trimestre, podemos observar que el rendimiento de estos bajo a continuación y apenas pudieron igualarse los rendimientos alrededor del trimestre 16, presentando una alza alrededor del trimestre 17, lo cual puede representar una perdida de tiempo y dinero para la persona que bien pudo invertir algunos otros instrumentos que tuvieran mejores rendimientos.

7.3 Variación temporal

El análisis para lograr la separación de la de la componente temporal o estacional requiere disponer de un número importante de datos, por el método que se utiliza para descomponerla, cuando la serie de tiempo contiene datos diarios, semanales o mensuales, la primera componente que debe ser aislada es la temporal.

De acuerdo con el modelo multiplicativo de la serie de tiempo,

$Y = (T)(C)(E)(I)$ es posible plantear la siguiente operación:

$$\frac{(T)(C)(E)(I)}{(T)(C)} = (E)(I)$$

El resultado obtenido, contiene los efectos estacionales, junto con las fluctuaciones irregulares, para separar estas componentes procederemos aplicando el proceso siguiente:

Dependiendo de la temporalidad de la información, podremos tener datos diarios, semanales, mensuales, etc, se obtienen los promedios móviles de un año completo, en nuestro ejemplo del rendimiento de los CETES a 90 días contamos con datos trimestrales, entonces deberemos obtener los promedios móviles de cuatro trimestres. Para separar la componente estacional, se requiere contar con valor promedio para cada uno de los cuatro trimestres del año, por lo que utilizaremos dos totales móviles consecutivos de cuatro trimestres, los sumaremos y obtendremos su promedio, al realizar esta operación, estamos amortiguando los efectos temporales o estacionales, dejando únicamente las componentes de **T** y **C**, de acuerdo con la expresión que anotamos arriba, podremos entonces separar a los componentes, **E** e **I** del modelo multiplicativo, al establecer la relación de los valores reales del rendimiento (**T**,**C**,**E,I**), y los obtenidos (**T**,**I**), con el proceso de cálculo descrito, el cual se ilustra en la tabla siguiente. Los valores de (**E** x **I)** de la última columna de la tabla se expresan en porcentaje.

Trimestre	Rendimiento Real	Total móvil 4 trimestres	Total móvil 8 trimestres	Promedio 8 trimestres	Componentes Estacional e Irregular %
1	14.03				
2	10.69				
3	8.63	42.93	79.31	9.91	87.05
4	9.58	36.38	68.05	8.51	112.62
5	7.48	31.67	60.53	7.57	98.86
6	5.98	28.86	54.83	6.85	87.25
7	5.82	25.97	52.58	6.57	88.55
8	6.69	26.61	54.75	6.84	97.75
9	8.12	28.14	55.88	6.99	116.25
10	7.51	27.74	52.24	6.53	115.01
11	5.42	24.5	43.9	5.49	98.77
12	3.45	19.4	35.58	4.45	77.57
13	3.02	16.18	32.45	4.06	74.45
14	4.29	16.27	34.11	4.26	100.62
15	5.51	17.84	37.73	4.72	116.83
16	5.02	19.89			
17	5.07				

Cuadro 7.2 Separación de las componentestemporal e Irregular

Una vez que hemos logrado separar las componentes E e I, procedemos a calcular los valores de la componente estacional para cada trimestre del año, para lo cual nos basaremos en los valores calculados. Los valores de E e I se organizan de acuerdo al trimestre y año que corresponden, de esta manera podemos disponer de igual número de valores para cada trimestre, estos valores se promedian para obtener un valor único para cada trimestre, el cual representa a la componente estacional, considerando que al calcular el promedio son eliminadas las irregularidades que contenían.

Trimestre	Año				Componente estacional
	1	2	3	4	
1		98.86	116.25	74.45	96.52
2		87.25	115.01	100.62	100.96
3	87.05	88.55	98.77		91.46
4	112.62	97.75	77.57		95.98

Cuadro 7.3 Cálculo de la componente temporal

Para separar del modelo a la componente temporal, basta dividir todas las componentes del modelo, **T** *x* **C** *x* **E** *x* **I,** entre el valor de la componente estacional, divido entre 100, esta última operación se presenta en la tabla siguiente, en la última columna se presentan los datos reales separados de la componente estacional.

Trimestre	Rendimiento Real	Componente estacional %	Datos No estacionales
1	14.03	96.52	14.54
2	10.69	100.96	10.59
3	8.63	91.46	9.44
4	9.58	95.98	9.98
5	7.48	96.52	7.75
6	5.98	100.96	5.92
7	5.82	91.46	6.36
8	6.69	95.98	6.97
9	8.12	96.52	8.41
10	7.51	100.96	7.44
11	5.42	91.46	5.93
12	3.45	95.98	3.59
13	3.02	96.52	3.13
14	4.29	100.96	4.25
15	5.51	91.46	6.02
16	5.02	95.98	5.23

Cuadro 7.4 Separación de la componente temporal

7.4 Variación irregular

Finalmente, una vez separada la componente estacional, procedemos a calcular la componente irregular, lo cual se realiza utilizando nuevamente la ecuación del modelo multiplicativo, relacionándola con el producto de las componentes conocidas hasta ahora, es decir obteniendo la relación:

$$\frac{(T)(C)(E)(I)}{(T)(C)(E)} = I$$

Los valores obtenidos se expresan en porcentaje, el cálculo de esta componente se muestra en la tabla siguiente:

Trimestre	Rendimiento Real	Componentes			
		tendencia Yc	cíclica C	temporal E	Irregular I
1	14.03	10.41	134.78	96.52	103.61
2	10.69	9.96	107.29	100.96	99.05
3	8.63	9.52	90.68	91.46	109.34
4	9.58	9.07	105.60	95.98	104.19
5	7.48	8.63	86.72	96.52	103.61
6	5.98	8.18	73.11	100.96	99.05
7	5.82	7.73	75.26	91.46	109.34
8	6.69	7.29	91.80	95.98	104.19
9	8.12	6.84	118.68	96.52	103.61
10	7.51	6.40	117.42	100.96	99.05
11	5.42	5.95	91.09	91.46	109.34
12	3.45	5.50	62.68	95.98	104.19
13	3.02	5.06	59.71	96.52	103.61
14	4.29	4.61	93.02	100.96	99.05
15	5.51	4.17	132.26	91.46	109.34
16	5.02	3.72	134.94	95.98	104.19
17	5.07	3.27	154.85		

Cuadro 7.5 Cálculo de la componente irregular

En la tabla se presentan los valores de cada una de las componentes, los correspondientes a la cíclica, estacional e irregular se expresan como un porcentaje del valor de la tendencia, la gráfica (7.4) que relaciona todos los valores se presenta enseguida.

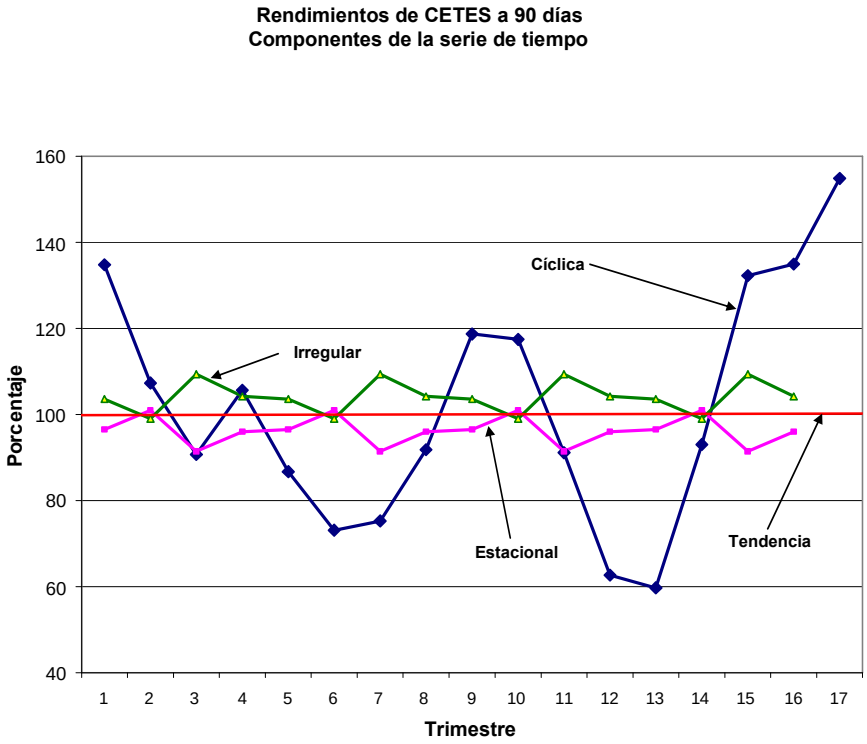


Figura 7.4Gráfica de los componentes de la serie de tiempo para nuestro ejemplo del rendimiento de los CETES a 90 días.

Una vez separadas cada una de los componentes es posible conocer la influencia que cada una de ellas tiene sobre el valor del rendimiento, y tomar una decisión sobre las consideraciones que deban realizarse para llevar a cabo una predicción, en este caso deberá analizarse con mucho atención relación que cada una de ellas haya tenido con los fenómenos económicos y hacer la consideración de las probabilidades que tiene de ocurrir de la misma manera, para considerar o no su participación en la predicción sobre el comportamiento del rendimiento de los CETES.

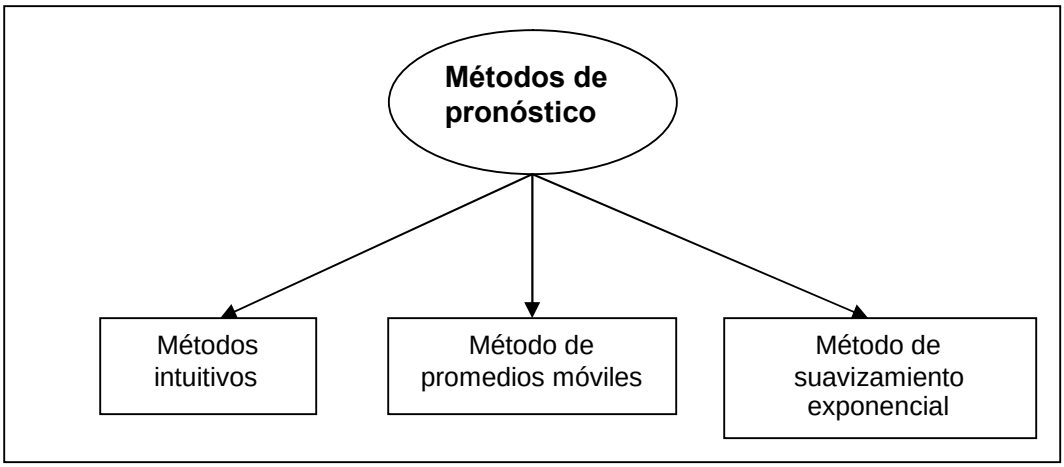
7.5 Análisis de predicciones

Realizar un pronóstico significa estimar los valores futuros de una variable, considerando que el comportamiento de ella en el pasado se repetirá de la misma manera, en la medida que esta hipótesis no se cumpla se presentarán diferencias

entre el valor real y el pronosticado, minimizar esta diferencia debe ser el objetivo del responsable del pronóstico.

Existen muchas maneras de realizar un pronóstico, algunos muy formales utilizan modelos matemáticos completos, como el de la descomposición de una serie de tiempo, analizada en subtemas anteriores, otros emplean modelos numéricos muy sencillos, e incluso la intuición o la experiencia lograda al realizar pronósticos frecuentemente.

Entre los métodos sencillos de pronóstico podemos mencionar los siguientes:



Cuadro 7.6 Métodos de pronóstico

Métodos intuitivos, son muy utilizados, de alguna manera todos los hemos utilizado de manera conciente o no, tiene la enorme desventaja de que al soportarse por la intuición, la cual tiene un aspecto completamente individual, la asignación de valores a una determinada variable puede diferir enormemente si más de una persona participa en el proceso. El método utilizado más sencillo es el suponer que la variable tendrá un comportamiento igual que el presentado en el periodo anterior, numéricamente puede ser expresado como:

$$Y_{i+1} = Y_i$$

En donde:

$\mathbf{Y}_{i+1}$ es el valor de la variable para el siguiente periodo

$\mathbf{Y}_{i}$ es el valor de la variable en el periodo anterior

Los modelos intuitivos de pronósticos resultan muy útiles en situaciones nuevas, en las cuales nos se cuenta con registros sobre el comportamiento de la variable y pueden ser complementados o mejorados en la medida en que se puedan introducir datos sobre tendencia y estacionalidad.

Método de promedios móviles. Estos métodos utilizan el valor promedio de un conjunto de datos anteriores, como el valor de la variable en le futuro. Se denominan móviles porque al disponer de un nuevo dato se elimina el más antiguo, permitiendo ir actualizando el valor pronosticado, el modelo puede expresarse matemáticamente de la manera siguiente:

$$Y_{i+1} = \frac{Y_i + Y_{i-1} + Y_{i-2} + Y_{i-3} + ... + Y_{i-m+1}}{m}$$

En donde:

$\mathbf{Y}_{i+1}$ Es el valor pronosticado para el periodo siguiente.

Y_i Es el valor del periodo anterior

m Es el número de términos, o datos utilizados

Una desventaja de este método estriba en que la participación de cada uno de los valores es considerada igual, perdiéndose algunas características o comportamientos particulares de la variable.

Método de suavizamiento exponencial. Se denomina de esta manera porque el peso considerado a periodos anteriores dentro del pronóstico va disminuyendo exponencialmente, esto significa que los valores mas antiguos participarán cada vez

en menor medida en la estimación, acercándose al valor cero, aunque nunca deja de ser considerado. Para expresar matemáticamente el modelo se emplea la expresión siguiente:

$$\mathbf{P}_{i+1} = \beta Y_i + (1- \beta) P_i$$

En donde:

- $\mathbf{P}_{i+1}$ Es el valor pronosticado para el nuevo periodo
- Y_i Es el dato real más reciente para la variable
- P_i Valor pronosticado para el periodo anterior
- B Constante de suavizamiento, cuyo valor debe estar comprendido entre 0 y 1

La constante β es la clave del suavizamiento, su elección determinará la respuesta del modelo ante cambios de los valores de la variable, su valor depende de la experiencia obtenida en el pasado, valores cercanos a 1 permitirán que el valor más reciente participe mayormente en el pronóstico, mientras que valores cercanos a cero lo evitarán, obteniéndose un pronóstico con un valor similar al del periodo anterior.

Cualquiera que sea el modelo sencillos de pronóstico elegido, es necesario medir el error cometido al comparar su valor con el que realmente ocurrió, el propósito de un pronóstico como hemos anotado antes es el de logra cada vez mejores estimaciones.

Bibliografía del tema 7

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill.

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica.

WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Actividades de aprendizaje

A.7.1. Elabora un glosario conceptual de las definiciones presentadas en el tema por medio de la bibliografía del tema.

A.7.2. Elabora un cuadro comparativo de las variaciones cíclica, temporal e irregular.

A.7.3. Investiga en artículos de revistas especializadas la manera en que se aborda las series de tiempo.

A.7.4. Investiga las aplicaciones de los métodos de pronósticos en los libros citados en la bibliografía y en revistas especializadas.

A.7.5. Elabora un resumen de lo visto en la materia.

Cuestionario de autoevaluación

1. ¿Qué es una serie de tiempo?

2. ¿Cuáles son los elementos de una serie de tiempo?

3. ¿Cuál es el modelo más utilizado para descomponer una serie de tiempo?

4. Explica qué es la tendencia en una serie de tiempo.

5. ¿Cómo se produce la tendencia de una serie de tiempo?

6. Explica qué es la componente cíclica en una serie de tiempo.

7. Explica qué es la componente estacional en una serie de tiempo.

8. Explica qué es la componente irregular en una serie de tiempo.

9. ¿Cómo se produce la componente irregular de una serie de tiempo?

10. ¿Cuál es el objetivo del responsable del pronóstico en el análisis de predicciones?

Examen de autoevaluación

Indica si las siguientes aseveraciones son verdaderas (V) o falsas (F):

_____1. Las series de tiempo resultan especialmente útiles cuando se requiere realizar un pronóstico sobre el comportamiento futuro que puede tener una variable determinada.

_____2. La tendencia es la componente que representa el comportamiento (crecimiento o decrecimiento), en un periodo corto de tiempo.

_____3. El objetivo principal del conocimiento de las series de tiempo es la identificación de los factores que intervienen y la separación de cada uno de ellos.

_____4. Una vez separada la componente estacional, se procede a calcular la componente irregular, la cual se realiza utilizando la ecuación del modelo multiplicativo.

_____5. La separación de la tendencia, utiliza la metodología de la línea de regresión, esta línea solo se puede representar por una recta.

_____6. Las fluctuaciones de los valores de rendimientos alrededor de la línea de tendencia, constituyen la componente estacional.

____7. Realizar un pronóstico significa estimar los valores futuros de una variable, considerando que el comportamiento de ella en el pasado se repetirá de la misma manera.

_____8. El método intuitivo esta basado en procedimientos rigurosos.

_____9. En el método de promedios móviles sus datos son fijos.

_____10. Se denomina método de suavizamiento exponencial porque el peso considerado a periodos anteriores dentro del pronóstico va disminuyendo exponencialmente

Bibliografía básica

ANDERSON R., David, *Estadística para administración y economía*, 8ª edición, Thopmson, 2004.

BERENSON, Mark, David LEVINE y Timothy KREHBIEL, Timothy, *Estadística para administración*, Editorial Pearson-Prentice Hall, 2001.

BLACK, Ken, *Estadística en los negocios*, Editorial CECSA, 2005.

LEVIN Richard I. y Rubin David S., *Estadística para administradores*, México; Alfaomega, 1996, 1017 pp.

LIND, Douglas A., *et al*, *Estadística para administración y economía*, Irwin-McGraw-Hill, 2001.

RAJ, Des, *Teoría del muestreo*, Fondo de Cultura Económica.

WEIMER, Richard, *Estadística,* Editorial CECSA, 2000.

Bibliografía complementaria

ATO, Manuel y Juan J. López, , *Fundamentos de estadística con SYSTAT*, México, Addison Wesley Iberoamericana, 1996.

CHRISTENSEN, H., *Estadística paso a paso*, 2ª edición, México, Trillas,. 1990.

GARZA, Tomás, *Probabilidad y estadística*, México, Iberoamericana, 1996.

GALINDO Caceres Jesús, *Técnicas de investigación en sociedad, cultura y comunicación,* Editorial Addison Wesley Longman, 1998.

HANKE, Jonh E. y Arthur G. Reitsch, *Estadística para negocios*, México, Irwin McGraw–Hill, 1997.

———————, *Pronósticos en los negocios*, México, Prentice Hall, 1996.

HILDEBRAN y Lyman, *Estadística aplicada a la administración y a la economía*, Addison Wesley, México, 1998.

KAZMIER L. Y A. Díaz Mata, *Estadística aplicada a la administración y economía*, México, McGraw–Hill, 1998.

KOHLER Heinz, *Estadística para negocios y economía*, Editorial CECSA, 1996.

KREYSZIG Erwin, *Matemáticas avanzadas para ingeniería*, vol. 2, Editorial Limusa-Wiley, 1990.

MASON D., Robert, Douglas LIND A. y William MARCHAL G, *Estadística para administración y economía*, 11ª edición, Colombia, Alfaomega, , 2004.

MENDENHALL, William, REINMUTH, James, *Estadística para administración y economía*, Grupo Editorial Iberoamericana, 1981.

MENDENHALL, W. Y R.L. Sheaffer, *Estadística matemática con aplicaciones*, México, Iberoamérica, 1986.

MEYER, Paul L., *Probabilidad y aplicaciones estadísticas*, México, Addison Wesley Iberoamericana, 2002.

SCHEAFFER, R. y W. Mndenhall, *Elementos de muestreo*, México, Iberoamericana, 1987.

WEIMER, Richard E., *Estadística*, México, CECSA, 2000.

WILLOUGHBY, Stephen S, *Probabilidad y Estadística*" Editorial Publicaciones cultural, 1974.

RESPUESTAS A LOS EXÁMENES DE AUTOEVALUACIÓN

ESTADÍSTICA II

Tema 1	Tema 2	Tema 3	Tema 4	Tema 5	Tema 6	Tema 7
1. d	1. d	1. c	1. b	1. e	1. c	1. V
2. c	2. b	2. d	2. d	2. b	2. b	2. F
3. a	3. b	3. d	3. a	3. d	3. d	3. V
4. b	4. d	4. a	4. c	4. d	4. b	4. V
5. d	5. c	5. d	5. b	5. a	5. a	5. F
6. d	6. c	6. d	6. d	6. c	6. d	6. F
7. b	7. d	7. b	7. b	7. d	7. a	7. V
8. a	8. b	8. a	8. a	8. e	8. c	8. F
9. c	9. d	9. d	9. b	9. b	9. a	9. F
10. b	10. e	10. b	10. d	10. b	10. c	10. V