

I DISTRIBUCIONES DE PROBABILIDAD

1.1 Variable aleatoria.- Una variable aleatoria X es una función de valor numérico que asigna un número real a cada punto del espacio muestral de un experimento.

Se dice que X es aleatoria por que está asociada a la probabilidad de los resultados del espacio muestral.

1.1.1 Variable aleatoria discreta.- Una variable aleatoria es discreta si la cantidad de valores que puede tomar es un número finito o infinito numerable de valores.

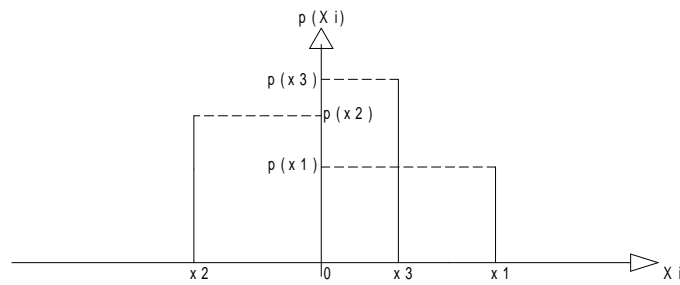
1.1.2 Variable aleatoria continua.- Se dice que X es una variable aleatoria continua cuando los valores que toma ésta son de carácter fraccionario.

1.2 Distribución de probabilidad.- Una distribución de probabilidad de variable aleatoria es el resultado de asignar valores de probabilidad a todos los valores numéricos posibles de dicha variable aleatoria, ya sea, mediante un listado o a través de una función matemática.

1.2.1 Función de cuantía.- Es aquella distribución de probabilidad de una variable aleatoria discreta, función que se representa generalmente mediante un listado de todos los valores numéricos posibles de la variable aleatoria con sus probabilidades correspondientes, tal como se observa en el Cuadro (1.1) y en la Gráfica (1.1).

Cuadro (1.1)
DISTRIBUCIÓN DE PROBABILIDAD
DE LA VARIABLE X

x_i	$p(x_i)$
x_1	$p(x_1)$
x_2	$p(x_2)$
x_3	$p(x_3)$
...	...
x_n	$p(x_n)$



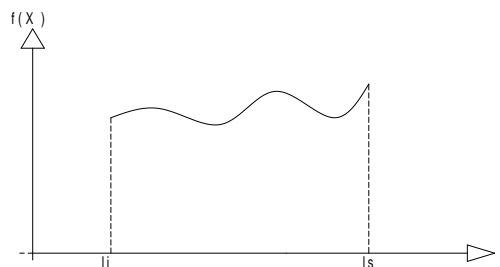
Si x_i es el valor de una variable aleatoria discreta y $p(x_i)$ la probabilidad de que x_i tome un valor en particular, todos los valores de $p(x_i)$ deben satisfacer las siguientes propiedades:

- $p(x_i) \geq 0 \quad \forall x_i \in N_x$
- $\sum_{i=1}^{i=n} p(x_i) = 1 \quad \forall x_i \in N_x$

Por otra parte, la Función de distribución acumulada de x_0 , es decir, la probabilidad de que x sea menor o igual a un valor específico x_0 se calcula con la ecuación (1.1).

$$F(x_0) = p(x \leq x_0) = \sum_{i=1}^{\forall i \rightarrow x_0} p(x_i) \quad \forall x_i \in N_x \quad (1.1)$$

1.2.2 Función de densidad.- Es aquella función en la que la probabilidad de los valores posibles de una variable aleatoria continua se determinan a través de una función matemática y se ilustra en forma gráfica por medio de una curva de probabilidad.



Si existe $f(x)$, se debe cumplir:

- $f(x) \geq 0 \quad \forall \text{ l.i.} < x < \text{l.s.}$
- $\int_{\text{l.i.}}^{\text{l.s.}} f(x) dx = 1$

Es importante recordar que $f(x)$ no representa ninguna probabilidad como tal y que solamente cuando la función se integra entre dos puntos produce una probabilidad, es decir:

$$p(a < x < b) = p(a \leq x \leq b) = \int_a^b f(x) dx \quad \forall x \in R_x \quad (1.2)$$

La Función de distribución acumulada de x_0 se define como:

$$F(x_0) = p(x \leq x_0) = \int_{\text{l.i.}}^{x_0} f(x) dx \quad \forall x \in R_x \quad (1.3)$$

1.3 Valor esperado y varianza de una variable aleatoria.-

1.3.1 Valor esperado.- El valor esperado de una variable aleatoria es el valor que se espera obtener después de repetir muchas veces el experimento. Es llamado también valor a la larga y esperanza matemática. Se define como:

$$E(x) = \mu = \sum_{i=1}^{i=n} x_i p(x_i) \quad \forall x \quad \text{v.a. discreta} \quad (1.4)$$

$$E(x) = \mu = \int_{\text{l.i.}}^{\text{l.s.}} x f(x) dx \quad \forall x \quad \text{v.a. continua} \quad (1.5)$$

La esperanza de una variable aleatoria x tiene las siguientes propiedades:

- a) $E(k) = k \quad \forall k \text{ constante}$
- b) $E(kx) = k E(x) \quad \forall k \text{ constante}$
- c) $E(k \pm x) = k \pm E(x) \quad \forall k \text{ constante}$
- d) $E(x \pm y) = E(x) \pm E(y) \quad \forall x, y \text{ var. aleatorias independientes}$
- e) $E(x y) = E(x) E(y) \quad \forall x, y \text{ var. aleatorias independientes}$

1.3.2 Varianza.- La varianza de una variable aleatoria x se define como:

$$V(x) = \sigma^2 = E(x - \mu)^2 \quad (1.6)$$

Para el caso de variables discretas y continuas, se tiene:

$$\sigma^2 = \sum_{i=1}^{i=n} (x_i - \mu)^2 p(x_i) \quad \forall x \quad \text{v.a. discreta} \quad (1.7)$$

$$\sigma^2 = \int_{li}^{ls} (x - \mu)^2 f(x) dx \quad \forall x \quad \text{v.a. continua} \quad (1.8)$$

Otras formas alternativas de cálculo son:

$$\sigma^2 = \sum_{i=1}^{i=n} x_i^2 p(x_i) - \mu^2 \quad \forall x \quad \text{v.a. discreta} \quad (1.9)$$

$$\sigma^2 = \int_{li}^{ls} x^2 f(x) dx - \mu^2 \quad \forall x \quad \text{v.a. continua} \quad (1.10)$$

A partir de las expresiones anteriores, la varianza también puede expresarse con la ecuación (1.11).

$$\sigma^2 = E(x^2) - [E(x)]^2 \quad (1.11)$$

La varianza tiene las siguientes propiedades:

- a) $V(k) = 0 \quad \forall k \text{ constante}$
- b) $V(kx) = k^2 V(x) \quad \forall k \text{ constante}$
- c) $V(k \pm x) = V(x) \quad \forall k \text{ constante}$
- d) $V(x \pm y) = V(x) + V(y) \quad \forall x, y \text{ var. aleatorias independientes}$

La raíz cuadrada de la varianza de una variable aleatoria se denomina desviación standard (σ) y al igual que la varianza es una medida de dispersión, es decir:

$$\sigma = +\sqrt{\sigma^2} \quad (1.12)$$

1.4 Distribuciones teóricas de probabilidad.-

1.4.1 Distribuciones teóricas de probabilidad de variable aleatoria discreta.- Las principales distribuciones teóricas de probabilidad de variable aleatoria discreta son:

- Distribución Bernoulli.
- Distribución Binomial.

- Distribución Poisson.
- Distribución Hipergeométrica.
- Distribución Uniforme discreta.
- Distribución Polinomial.
- Distribución Geométrica.

1.4.2 Distribuciones teóricas de probabilidad de variable aleatoria continua.- Las principales distribuciones teóricas de probabilidad de variable aleatoria continua se desarrollan a continuación.

1.4.2.1 Distribución Normal.- La distribución Normal o distribución de Gauss es fundamental en la aplicación de la inferencia estadística, ya que las distribuciones de muchos estadígrafos muestrales tienden a la distribución Normal conforme crece el tamaño de la muestra.

Se dice que una variable aleatoria x está normalmente distribuida si su función de densidad está dada por:

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}} \quad -\infty < x < +\infty \quad (1.13)$$

En la que:

μ = valor esperado

$-\infty < \mu < +\infty$

σ^2 = varianza

$\sigma^2 > 0$

La gráfica de la distribución Normal es una curva simétrica con forma de campana, que se extiende sin límites tanto en la dirección positiva como en la negativa.

1.4.2.1.1.- Cálculo de probabilidades.- La probabilidad de que una variable aleatoria normalmente distribuida sea menor o igual a un valor específico, está dada por la función de distribución acumulada de la ecuación (1.14).

$$p(x \leq a) = \int_{-\infty}^a \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma^2}} dx \quad (1.14)$$

La función $f(x)$ no es una función simple y su integración no puede realizarse en forma sencilla, además, si se tabulara la función de densidad de la distribución Normal, la tabla que se elaboraría sería para un par de valores de μ y σ^2 , tarea virtualmente imposible.

Por tanto, para reducir el problema anterior, es necesario standarizar la variable, de tal forma que permita presentar los resultados en una sola tabla, es decir:

$$z = \frac{x - \mu}{\sigma} \quad (1.15)$$

Luego:

$$f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi\sigma^2}} \quad (1.16)$$

con:

$$\begin{aligned} E(z) &= 0 \\ V(z) &= 1 \end{aligned}$$

De manera que:

$$p(x \leq a) = p\left(z \leq \frac{a - \mu}{\sigma}\right) = \int_{-\infty}^{\frac{a - \mu}{\sigma}} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz \quad (1.17)$$

Con métodos de cálculo integral, todavía sigue siendo difícil integrar la función de densidad acumulada de la distribución Normal standarizada, sin embargo, por medio del análisis numérico se han obtenido tablas para dicha función.

1.4.2.1.2.- Propiedad reproductiva de la Distribución Normal.- Una propiedad muy importante de la Distribución Normal es la llamada Propiedad Reproductiva de la Distribución Normal, la cual indica:

“Si $x_1, x_2, x_3, \dots, x_k$, son variables aleatorias normalmente distribuidas cada una con media y varianza: $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$, ..., $N(\mu_k, \sigma_k^2)$, respectivamente, además, si:

$$y = x_1 + x_2 + x_3 + \dots + x_k \quad (1.18)$$

Entonces se cumple:

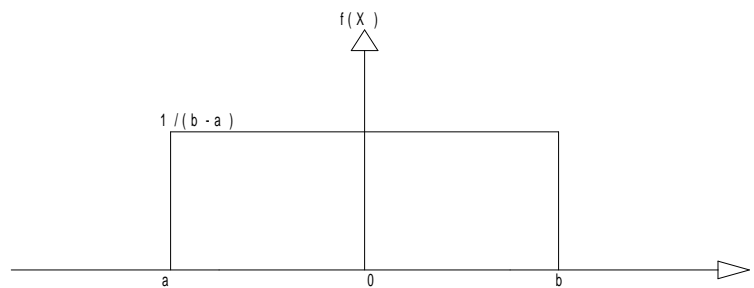
$$E(y) = \mu_y = \mu_1 + \mu_2 + \mu_3 + \dots + \mu_k \quad (1.19)$$

$$V(y) = \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \dots + \sigma_k^2 \quad (1.20)$$

1.4.2.2 Distribución Uniforme continua.- Se dice que una variable aleatoria x está distribuida uniformemente en el intervalo $(,)$ si su función de densidad es:

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha < x < \beta \\ 0 & \text{en otros casos} \end{cases} \quad (1.21)$$

La Distribución Uniforme es llamada también Distribución Rectangular, por la forma de su figura:



La Distribución Uniforme continua presenta en los experimentos en los que ocurre un evento en que la variable aleatoria toma valores de un intervalo finito, de manera que estos se encuentran distribuidos igualmente sobre el intervalo; es decir, la probabilidad de que la variable aleatoria tome un valor en cada subintervalo de igual longitud (contenido en el intervalo $(,)$) es la misma, sin importar la localización exacta del subintervalo.

La esperanza y la varianza de la distribución Uniforme son:

$$E(x) = \mu = \frac{\alpha + \beta}{2} \quad (1.22)$$

$$V(x) = \sigma^2 = \frac{(\beta - \alpha)^2}{12} \quad (1.23)$$

La distribución Uniforme es simétrica y su mediana es igual a la media.

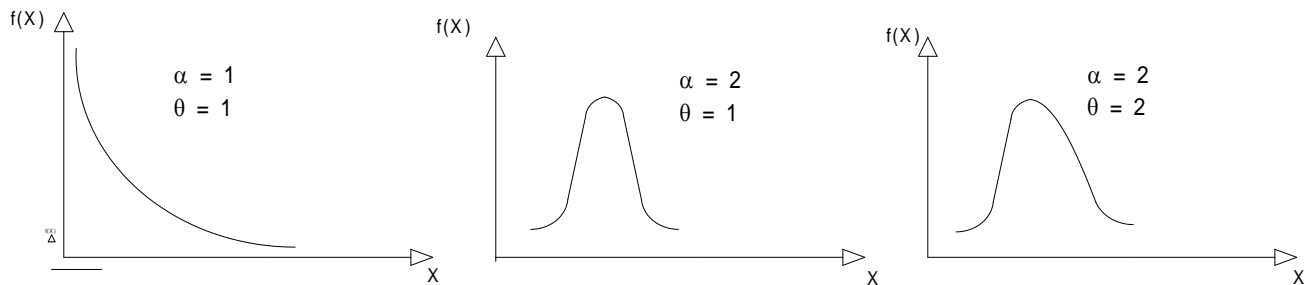
1.4.2.3 Distribución Gamma.- La variable aleatoria x tiene una distribución Gamma si su función de densidad está dada por:

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\theta}}}{\Gamma(\alpha) \theta^{\alpha}} \quad x \geq 0, \alpha \geq 0, \theta \geq 0 \quad (1.24)$$

En la que se define a la función gamma de α a la expresión (1.25).

$$\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1} e^{-z} dz \quad (1.25)$$

Para distintos valores de α y θ se tienen los siguientes gráficos:



La esperanza y la varianza son:

$$E(x) = \alpha\theta \quad (1.26)$$

$$V(x) = \alpha\theta^2 \quad (1.27)$$

Cuando α es un número entero se origina la distribución Erlang, cuya función de densidad es:

$$f(x) = \frac{e^{-\frac{x}{\theta}}}{\alpha(\alpha-1)! \theta^{\alpha}} \quad x \geq 0, \theta \geq 0, \alpha \geq 1 \quad (1.28)$$

1.4.2.4 Distribución Exponencial.- La Distribución Exponencial, llamada también Distribución Exponencial Negativa, es un caso especial de la distribución Gamma con $\alpha = 1$, es decir:

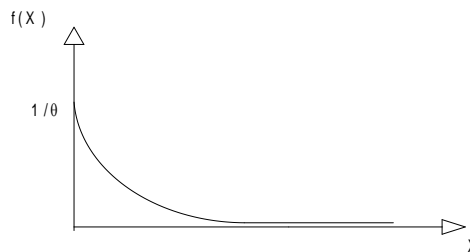
$$f(x) = \begin{cases} \frac{e^{-x/\theta}}{\theta} & \forall x \geq 0, \theta \geq 0 \\ 0 & \text{en otros casos} \end{cases} \quad (1.29)$$

La esperanza y la varianza son:

$$E(x) = \theta \quad (1.30)$$

$$V(x) = \theta^2 \quad (1.31)$$

Gráficamente:



La variable aleatoria Exponencial representa el tiempo que transcurre hasta que se presenta el primer evento Poisson, es decir, la Distribución Exponencial puede modelar el lapso entre dos eventos consecutivos Poisson que ocurren de manera independiente y a una frecuencia constante (el parámetro θ representa el tiempo promedio entre dos eventos Poisson).

Por ejemplo, el tiempo que transcurre entre llegadas de un cliente a una tienda ó un paciente a un servicio de emergencia de un hospital, la duración de una llamada telefónica, la duración de un componente eléctrico, etc..

Esta distribución sirve para modelar problemas del tipo tiempo-falla y problemas de líneas de espera.

1.4.2.5 Distribución Chi Cuadrado.- Un caso especial de la Distribución Gamma, con $\alpha = \nu/2$ y $\theta = 2$, es la Distribución Chi Cuadrado cuya función de densidad es:

$$f(x) = \frac{x^{\frac{\nu}{2}-1} e^{-x/2}}{\Gamma(\frac{\nu}{2}) 2^{\frac{\nu}{2}}} \quad x \geq 0 \quad (1.32)$$

En la que:

ν = grados de libertad (entero positivo)

Los grados de libertad representan la cantidad de valores que se asignan de manera arbitraria en una ecuación, tal que de esa manera se pueda conocer una de esas variables.

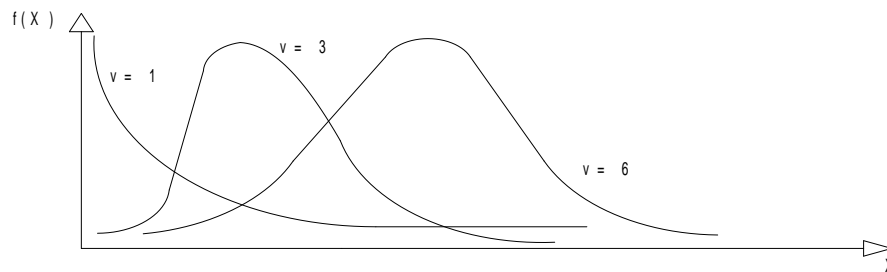
La esperanza y la varianza se muestran en las ecuaciones (1.33) y (1.34).

$$E(x) = \nu \quad (1.33)$$

$$V(x) = 2\nu \quad (1.34)$$

Esta distribución se emplea bastante en la inferencia estadística (pruebas de hipótesis) y de forma especial al hacer inferencias con respecto a las varianzas.

De acuerdo al valor de ν , la gráfica puede ser:



Para calcular probabilidades se aplica la ecuación (1.35), aunque para ello existen tablas con la integral ya desarrollada (ver anexos).

$$p(x > a) = \int_a^{\infty} f(x) dx \quad (1.35)$$

1.4.2.6 Distribución "t" o de Student.- Suponiendo que se realiza un experimento en el que se obtienen dos variables aleatorias independientes: w con Distribución Chi Cuadrado con ν grados de libertad y z con Distribución Normal con media 0 y varianza 1, entonces la variable aleatoria " x " definida como:

$$x = \frac{z}{\sqrt{\frac{w}{v}}} \quad -\infty < z < +\infty \quad (1.36)$$

w>0, v>0 y entero

$-\infty < x < +\infty$

tiene una distribución "t" de Student, cuya función de densidad es:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v}} \left(1 + \frac{x^2}{v}\right)^{-\frac{(v+1)}{2}} \quad -\infty < x < \infty \quad (1.37)$$

v>0, entero positivo

En la que:

v = grados de libertad

Gráficamente, la Distribución "t" tiene forma de campana y es simétrica respecto al origen, además se puede observar que esta Distribución se asemeja a la Distribución Normal puesto que ambas varían en el intervalo $(-\infty; +\infty)$, son unimodales y centradas alrededor de 0.

La esperanza y la varianza es:

$$E(x) = 0 \quad (1.38)$$

$$V(x) = \frac{v}{v-2} \quad v>2 \quad (1.39)$$

Para calcular probabilidades se aplica la ecuación (1.40).

$$p(x \leq a) = \int_{-\infty}^a f(x) dx \quad (1.40)$$

Los valores de probabilidad se encuentran tabulados para ciertos valores especiales (ver anexos).

La importancia de la Distribución "t" radica en el hecho de que es útil al efectuar inferencias respecto a la media aritmética cuando el valor de la desviación standard es desconocido y la población tiene una Distribución Normal sin importar el tamaño de la muestra.

1.4.2.7 Distribución "F" de Fisher.- Esta distribución es también muy utilizada en la inferencia estadística y se define de la siguiente manera:

Sea un experimento en el que se generan dos variables aleatorias independientes w y z , cada una con una Distribución Chi Cuadrado con v_1 y v_2 grados de libertad respectivamente, se define la variable "x" a la relación (1.41).

$$x = \frac{\frac{w}{v_1}}{\frac{z}{v_2}} \quad x > 0 \quad (1.41)$$

y se dice que tiene una distribución "F" con v_1 y v_2 grados de libertad con función de densidad:

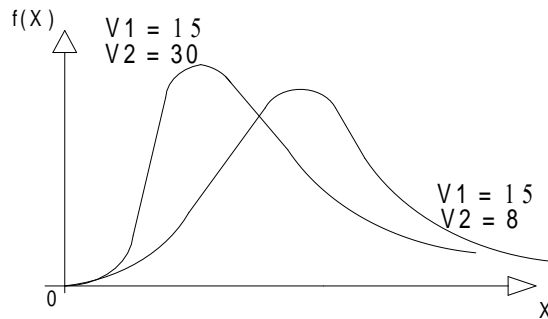
$$f(x) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) v_1^{\frac{v_1}{2}} v_2^{\frac{v_2}{2}} x^{\frac{v_1-2}{2}} (v_1 x + v_2)^{-\frac{v_1+v_2}{2}}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right)} \quad (1.42)$$

La esperanza y la varianza son:

$$E(x) = \frac{v_2}{v_2 - 2} \quad v_2 > 2 \quad (1.43)$$

$$V(x) = \frac{v_2^2(2v_2 + 2v_1 - 4)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad v_2 > 4 \quad (1.44)$$

Gráficamente, para distintos valores de v_1 y v_2 se tiene:



La Distribución “F” es asimétrica hacia la derecha para cualquier par de valores de v_1 y v_2 , pero ésta va disminuyendo conforme v_1 y v_2 se vuelven más grandes.

Para el cálculo de probabilidades se emplea la ecuación (1.45).

$$p(x \leq a) = \int_{-\infty}^a f(x) dx \quad (1.45)$$

La ecuación (1.45) se encuentra tabulada, existiendo tablas para 90%, 95%, y 99%.

La importancia de esta distribución radica en que es útil para efectuar inferencias sobre las varianzas de 2 distribuciones Normales.

BIBLIOGRAFÍA:

- (1) **LEVIN** Richard (1996): “*Estadística para Administración y Economía*”. México
- (2) **MOYA** Rufino y **SARAVIA** Gregorio (1988): “*Probabilidad e Inferencia Estadística*”. Perú.
- (3) **MOYA** Rufino (1991): “*Estadística descriptiva. Conceptos y aplicaciones*”. Perú.

=====

ÍNDICE**Página**

1.1 Variable aleatoria.....	1
1.1.1 Variable aleatoria continua.....	1
1.1.2 Variable aleatoria discreta.....	1
1.2 Distribución de probabilidad de una variable aleatoria.....	1
1.2.1 Función de cuantía	1
1.2.2 Función de densidad	2
1.3 Valor esperado y varianza de una variable aleatoria.....	3
1.3.1 Valor esperado.....	3
1.3.2 Varianza.....	3
1.4 Momento de una variable aleatoria.....	4
1.5 Distribuciones teóricas de probabilidad	5
1.5.1 Distribuciones teóricas de probabilidad de variable aleatoria discreta...	5
1.5.2 Distribuciones teóricas de probabilidad de variable aleatoria continua.	5
1.5.2.1 Distribución Normal.....	5
1.5.2.2 Distribución Uniforme Continúa.....	7
1.5.2.3 Distribución Gamma.....	8
1.5.2.4 Distribución Exponencial.....	9
1.5.2.5 Distribución Chi cuadrado.....	10
1.5.2.6 Distribución “t” de Student.....	11
1.5.2.7 Distribución “F” de Fisher.....	12

II TEORÍA GENERAL DE MUESTREO

2.1. Introducción.- En todo estudio que se realice se desea conocer con absoluta verdad y certeza toda la información requerida para tales fines.

Es natural que se busque conocer de manera exhaustiva las características de una población objeto de estudio y que para ello se requiera efectuar un censo. El censo tiene sus ventajas y desventajas; estas últimas ocasionan que se recurra a otro procedimiento que vendría a ser el muestreo.

El muestreo posee características especiales que la hacen favorable en su uso frecuente por parte de instituciones con recursos económicos y tiempo limitados, aunque también el factor que se debe controlar es el error presente en este procedimiento.

El muestreo tiene distintas etapas, siendo dos de las más importantes, la forma de elegir los elementos de la muestra y el tamaño de la misma. En este capítulo se efectuará el estudio de dichas etapas, en base a las cuales se realizarán inferencias referentes a los parámetros de estudio desconocidos.

2.2 Censo.-

2.2.1 Población objetivo.-

2.2.1.1 Definición.- Es la totalidad de los elementos en discusión y acerca de los cuales se desea obtener alguna información, dichos elementos tienen características comunes que son de interés para el estudio. Ej.: todos los Centros Hospitalarios ubicados en el departamento de Cochabamba, todos los Proyectos de Grado presentados en la Carrera de Ing. Mecánica de la Facultad de Ciencias y Tecnología de la UMSS, etc..

Para garantizar el censo es necesario acotar el universo y conocer las unidades que lo componen; acotar el universo significa concretar la población que va a ser objeto del estudio. Por ejemplo: número de bolsas de cemento producidas por COBOCE el día 22 de julio de 2009 en la planta ubicada en Cochabamba.

2.2.1.2. Tipos de población.- De acuerdo a la magnitud de la población se definen dos tipos de población.

2.2.1.2.1. Población finita.- Una población es finita si tiene un número limitado de sucesos o unidades elementales, numéricamente es aquella que tiene menos de 500.000 unidades, por ejemplo: todos los estudiantes de la Carrera de Biología, número de clientes diarios que llegan a un autobanco, etc..

2.2.1.2.2. Población infinita.- Es aquella que consiste en un número infinitamente grande de observaciones. Se considera infinita a una población que posee más de 500.000 unidades. Ejemplo: el conjunto de estrellas del Universo, toda la población de Bolivia, etc..

2.2.2. Parámetro.- Es posible definir este concepto de dos formas:

a) El parámetro es una caracterización numérica de la distribución de la población, es decir, describe parcial o completamente, la función de probabilidad de la población de la variable de interés. Por ejemplo, cuando se especifica λ de la distribución Poisson, se está definiendo su función de probabilidad:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (2.1)$$

Si se conoce el parámetro se puede calcular cualquier proposición probabilística. Por tanto, puesto que los parámetros son inherentes a todos los modelos de probabilidad, es imposible calcular las probabilidades deseadas sin un conocimiento del valor de éstos.

b) El parámetro es una característica de tipo descriptivo de una determinada población. Ello se refiere a que una población con determinadas características, pueden ser descritas por ciertas medidas descriptivas, como por ejemplo, la media aritmética, índices poblacionales, tasas, etc..

2.2.3. Definición de censo.- Cuando es necesario conocer uno o más parámetros de una población o universo se recurre a la realización de un censo.

El censo constituye un examen completo de todos los elementos de una población. En la mayoría de los casos la realización de censos para conocer las características de una determinada población resulta muy costosa, exige la movilización de muchos recursos humanos, su duración suele ser muy larga y en algunos casos el proceso es destructivo.

Existen muchos tipos de censo, siendo el más importante el censo de Población y Vivienda, en el cual es necesario recabar la información de todos los habitantes del país, por ser marco obligado de referencia para multitud de trabajos e investigaciones y, por razones meramente administrativas.

2.3 Muestreo.-

2.3.1.- Definición.- Para el conocimiento de las características de la población existen métodos opcionales cuyo costo y tiempo de realización se reducen considerablemente. Estos métodos están constituidos en lo que se denomina muestreo, cuyo objetivo es reconstruir modelos reducidos de la población total, con resultados que pueden extrapolarse al universo del que se extraen.

Todo ello quiere decir que a través de muestras se puede obtener en muchos casos, la información requerida, con un ahorro sustantivo de recursos humanos, económicos y de tiempo, sin que ello implique un alejamiento de la realidad que se desea conocer.

Para que el proceso de muestreo sea una reconstrucción reducida pero real del universo que se desea investigar es necesario que el tamaño de las muestras y la metodología utilizada en su elaboración respondan a determinados principios, deducidos del cálculo de probabilidades.

2.3.2. Muestra aleatoria.- La muestra aleatoria es aquella en la que cada unidad elemental para la observación tiene la misma probabilidad de ser incluida en la muestra.

O de una forma más específica: $x_1, x_2, x_3, \dots, x_n$, es una muestra aleatoria de tamaño n , si cumple:

- a)** Cada x_i es una variable aleatoria independiente.
- b)** Cada x_i tiene la misma distribución de probabilidad.

2.3.3. Inferencia estadística.- La inferencia estadística es el proceso mediante el cual se utiliza la información de los datos de una muestra para extraer conclusiones acerca de la población de la que se seleccionó.

La inferencia estadística se basa en la inferencia inductiva, la cual constituye una generalización de los resultados particulares a resultados generales.

Por ejemplo, si se tiene una florería que cuenta con 100.000 semillas de que se desean comercializar, de la cual se sabe que pueden producir flores blancas o rojas. El objetivo para la gerencia es averiguar cuántas de estas 100.000 semillas producirán flores rojas.

Por tanto, lo más lógico sería proceder de la siguiente manera:

1° Para dar una respuesta correcta, se debería sembrar todas las semillas y observar el número de las que producen flores rojas.

2° Como ello es imposible, puesto que se desea vender todas las semillas y aunque no se quisiera venderlas, el obtener una respuesta requerirá invertir mucho esfuerzo y dinero. Por lo que:

3° La solución será emplear unas cuantas semillas y basados en los resultados aparecidos, hacer una afirmación sobre el número de flores rojas que se tendrán del total restante de semillas.

Toda inferencia inductiva constituye un proceso arriesgado, es decir, la inferencia inductiva exacta es imposible, existiendo un grado de incertidumbre susceptible de medición a través de la probabilidad.

La importancia de la inferencia estadística radica en que por medio de ella se hallan nuevos conocimientos.

2.3.4. Estadígrafo.- El estadígrafo es cualquier función de las variables que se observaron en la muestra, de manera que, esta función no contiene cantidades desconocidas. Por ejemplo: si x_1, x_2, \dots, x_n son variables aleatorias obtenidas de una muestra, entonces: $\bar{x} = \left(\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \right)$ es un estadígrafo.

Un parámetro es una constante, pero un estadígrafo es una variable aleatoria. Además, un parámetro describe un modelo de probabilidad, ningún valor de estadígrafo puede desempeñar este papel, porque depende de las observaciones de la muestra.

2.3.5. Diseño de una muestra.

2.3.5.1 Definición.- Por diseño de una muestra se entiende la planificación o metodología para tomar muestras.

2.3.5.2. Criterios para evaluar el diseño de una muestra.- Existen dos criterios para evaluar el diseño de una muestra: su fiabilidad y su efectividad.

2.3.5.2.1. Fiabilidad.- Es de esperar que en el muestreo existan errores. El error de muestreo es la diferencia entre el valor de un estadígrafo y el valor del correspondiente parámetro de población, ello debido a variaciones fortuitas en la selección de las unidades elementales.

Por otra parte, el error de muestreo es cuantificable mediante la fiabilidad, la cual está estrechamente relacionada con la varianza del estadígrafo; por lo cual, cuanto menor la varianza, mayor será la fiabilidad del resultado de la muestra.

2.3.5.2.2. Efectividad.- El diseño de una muestra se considera efectivo si se obtiene cierto grado de fiabilidad al menor costo posible. Un diseño muestral se considera más efectivo que otro, si el primero tiene menor costo que el segundo, dentro del mismo grado de fiabilidad.

2.3.6. Tipos de muestreo.- Para la selección de la muestra se pueden utilizar distintos métodos o combinación de métodos, todos estos divididos en dos grandes grupos:

- Muestreo aleatorio.
- Muestreo no aleatorio.

2.3.6.1. Muestreo aleatorio.- Comprende:

2.3.6.1.1. Muestreo aleatorio simple.- El muestreo aleatorio simple se aplica en casos en que:

- Las unidades elementales son fáciles de identificar.
- Cuando la población es pequeña.
- Cuando la población es homogénea respecto a la característica de interés.

El procedimiento consiste en numerar a toda la población del estudio y extraer al azar una muestra de n unidades. En el muestreo aleatorio simple la

selección de los elementos se efectúa en una sola etapa y en forma directa, pudiendo ser con o sin reemplazo.

Para la selección aleatoria de los números se utilizan tablas de números aleatorios, programas de computación, bolillos numerados, etc..

a) Muestreo aleatorio con reemplazo. En este caso cada elemento de la muestra posee la misma probabilidad de ser elegida, puesto que cada uno es reintegrado a la población de la cual fue extraída.

b) Muestreo aleatorio sin reemplazo. En este caso cada unidad de la población posee la misma probabilidad de ser escogida que las restantes para formar parte de la muestra, considerando que la probabilidad de que un elemento sea extraído dependerá de los que anteriormente hayan sido elegidos.

La clave de este procedimiento es naturalmente la técnica del azar, aunque el lograr dicho "azar" o aleatoriedad no es cosa sencilla. Por ejemplo, si se desea averiguar cuál es la mejor Universidad de Cochabamba, no es aleatoria una muestra de personas, si nos dirigimos al campus de la Universidad Católica Boliviana y se procede a entrevistar a las personas que ingresan a la misma.

Para poblaciones grandes el método es costoso y requiere mucho tiempo, siendo difícil y tediosa la elaboración de listas con toda la población. Cuando el universo no es homogéneo se produce mucho error.

2.3.6.1.2. Muestreo aleatorio sistemático.- El muestreo sistemático se emplea cuando existe heterogeneidad respecto a algún rasgo de los elementos de la población y el tamaño de ésta es pequeño. Para tal efecto es aconsejable disponer de una lista de las unidades de la población, como ser una guía telefónica.

En el control de calidad se emplea frecuentemente el muestreo sistemático tomando muestras de artículos de la corriente de producción.

Este procedimiento consiste en obtener una muestra tomando cada **k**-ésima unidad de la población, tras numerar las unidades de la población u ordenarlas de

alguna manera. La letra **k** representa un número entero llamado razón de muestreo, coeficiente de elevación ó salto y es igual a:

$$k = \frac{N}{n} \quad (2.2)$$

En la que:

N = tamaño de la población.

n = tamaño de la muestra.

Para que toda unidad de la población tenga igual probabilidad de salir, el procedimiento debe empezar al azar; para ello se elige un número al azar, número no superior a **k**, a partir del cual se suma sucesivamente la razón de muestreo.

Ahora bien, la muestra sistemática es menos representativa que el muestreo aleatorio simple, en situaciones en que existe periodicidad oculta en la población, es decir, cuando existe un movimiento cíclico o periódico de los datos con la longitud del ciclo aproximándose a la razón de muestreo **k**; por ejemplo, la venta de entradas en una empresa cinematográfica, el elegir sábado o domingo para tomar una muestra, no siempre es representativo. Este problema se puede solucionar parcialmente si se procede a "desordenar" la lista.

La desventaja principal del muestreo sistemático es numerar u ordenar los elementos de una población grande, lo cual es físicamente imposible si se abarca todo un país o zona geográficamente grande.

2.3.6.1.3. Muestreo aleatorio estratificado.- El proceso de estratificación consiste en dividir la población en clases o grupos llamados **estratos**. Dentro de cada uno de tales estratos se encuentran los elementos situados de manera más homogénea con respecto a las características en estudio. Para cada estrato se toma una submuestra mediante el muestreo aleatorio simple y la muestra global se obtiene combinando las submuestras de todos los estratos.

El muestreo por estratos es efectivo cuando se trata de poblaciones heterogéneas, por que al efectuarse la estratificación, los grupos se establecen de modo que las unidades de muestreo tienden a ser uniformes dentro de cada clase y

los grupos tienden a ser diferentes entre sí. Así se puede controlar la proporción de cada estrato en la muestra global y no dejarla al azar, quedando asegurado el carácter representativo de la muestra.

Si la varianza de la característica observada de cada estrato es menor que de toda la población, que es lo más usual debido a la uniformidad dentro del estrato, resultará aumentada la fiabilidad para un tamaño de muestra.

El aumento de fiabilidad y efectividad se puede incrementar clasificando todavía los estratos en subestratos llamando a este procedimiento estratificación doble.

Para definir los estratos se emplean:

- Datos anteriores.
- Resultados preliminares de otros estudios.

2.3.6.1.4. Muestreo aleatorio por conglomerados.- Llamado también muestreo por áreas, consiste en seleccionar al azar grupos, llamados **conglomerados**, de elementos individuales de la población, y tomar luego todos los elementos o una submuestra de ellos dentro de cada conglomerado para constituir así la muestra total. Como ejemplo de conglomerados se tiene:

- Urbanizaciones.
- Centros hospitalarios.
- Ciudades universitarias.

Con este tipo de muestreo se desea que las diferencias entre conglomerados sean lo más pequeñas posibles, es decir, que exista homogeneidad entre conglomerados; por otro lado, se busca que dentro de los conglomerados, las diferencias entre los elementos individuales sean lo más grandes posibles, es decir, que exista heterogeneidad dentro de los conglomerados. En ello radica la diferencia, diametralmente opuesta, al muestreo por estratos.

El objetivo en el muestreo por conglomerados es que cada conglomerado sea una representación, a escala reducida, del universo. Además, sólo algunos de éstos

forman parte de la muestra, mientras que en el muestreo estratificado existe en la muestra algún elemento de cada uno de los estratos.

Si todos los elementos de cada uno de los conglomerados se incluyen en la muestra, se denomina muestreo de una etapa. Si se extrae una submuestra aleatoria de elementos de cada conglomerado seleccionado, se tiene un muestreo en dos etapas. Si se obtienen más de dos etapas en la obtención de la muestra, se dice que es un muestreo de etapas múltiples o polietápico.

Este tipo de muestreo se emplea a menudo en el control de calidad estadístico, seleccionando lotes o "tandas" de producción al azar como conglomerados.

2.3.6.2. Muestreo no aleatorio.- Frente a los distintos tipos de muestreo aleatorio, se suelen utilizar otros sistemas de selección de la muestra, englobados en lo que también se denomina muestreo dirigido. El recurrir a uno u otro método se encuentra en función no sólo de los costos, sino también de la precisión que se desea obtener de la estimación y la posibilidad de cuantificar los errores de muestreo.

Generalmente, las instituciones oficiales tienden a emplear muestreos aleatorios y las instituciones de opinión, mayormente privadas, emplean el muestreo no aleatorio; ello en virtud a la disponibilidad de información y el costo que ello representa.

2.3.6.2.1 Muestreo opinático.- En este caso el investigador, según su criterio, selecciona la muestra de manera que sea lo más representativa a los efectos de la investigación que se pretende realizar, por ejemplo: estudios sobre el consumo de droga en una determinada ciudad. Sin embargo, está sujeto a la subjetividad del investigador y los resultados carecen de fiabilidad en términos estadísticos.

2.3.6.2.2. Muestreo por cuotas.- Consiste en facilitar al entrevistador el perfil de las personas que tiene que entrevistar de acuerdo a los objetivos del estudio.

2.4. Distribución muestral.- La distribución muestral de un estadígrafo es la distribución de probabilidad que expresa la relación funcional entre cada uno de los

valores del estadígrafo y su correspondiente probabilidad, como resultado de un número infinito de muestras aleatorias independientes, cada una de tamaño n , provenientes de la misma población.

De la distribución muestral los elementos más importantes son el valor esperado y la varianza. Por otro lado, la distribución muestral de un estadígrafo no tiene la misma forma que la función de probabilidad de la población de la cual proviene la muestra.

Por ejemplo, suponga que se tiene interés en el número de clientes que llegan a los bancos de la ciudad, entre las 9:00 y las 10:00 de la mañana, teniendo certeza que cada una de las llegadas es independiente entre sí, se decide seleccionar en forma aleatoria cinco bancos durante 8 días. Para cada muestra diaria, se procede a contar la cantidad de personas que ingresan durante el intervalo de una hora en los cinco bancos. Con tales consideraciones se obtienen los resultados del cuadro (2.1).

Cuadro (2.1)

NUMERO DE LLEGADAS A LOS BANCOS EN UNA HORA								
DIA BANCO	1	2	3	4	5	6	7	8
BISA	63	59	50	36	36	38	55	58
MERCANTIL-STA. CRUZ	32	44	25	57	46	45	45	50
UNION	54	39	39	68	58	50	51	53
GANADERO	52	46	34	58	54	38	54	51
DE CREDITO	48	44	56	67	56	58	41	38
Promedio (\bar{X})	50	46	40	57	50	46	49	50

Fuente: Elaboración propia.

En este caso, el estadístico es el promedio o media muestral y todos los valores obtenidos, conforman la distribución muestral de \bar{x} .

2.4.1. Distribución muestral de \bar{x} (promedio muestral).- Uno de los estadígrafos más importante es el promedio de un conjunto de variables aleatorias e independientemente distribuidas, llamado también promedio o media muestral. Este estadígrafo tiene un papel muy importante en problemas de decisiones para medias poblacionales desconocidas.

Por tanto, si: $x_1, x_2, x_3, \dots, x_n$, es una muestra aleatoria de n variables aleatorias independientes e igualmente distribuidas con $E(x_i) = \mu$ y varianza $VAR(x_i) = \sigma^2$, para $i = 1, 2, 3, \dots, n$; se define a la media muestral como:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \sum_{i=1}^n \frac{X_i}{n} \quad (2.3)$$

Si se aplica muestreo con reemplazo, entonces se cumple que:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu \quad (2.4)$$

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = \frac{\sigma^2}{n} \quad (2.5)$$

de lo que se deduce:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (2.6)$$

que se denomina error típico de la media muestral o desviación standard de la distribución muestral de la media muestral.

Este resultado es válido sin importar la distribución de probabilidad de la población de interés, siempre y cuando la varianza tenga un valor finito.

Lo expuesto anteriormente hace posible encontrar el error típico de la media sin conocer la distribución de \bar{X} .

Para el caso del muestreo sin reemplazo, se tiene:

$$\mu_{\bar{X}} = \mu \quad (2.7)$$

$$\sigma_{\bar{X}}^2 = \frac{(N - n)}{(N - 1)} \frac{\sigma^2}{n} \quad (2.8)$$

En la que:

N = Número de elementos de la población.

(N-n)/(N-1) = corrección finita de la población

Cuando **N** tiende a infinito la ecuación **(2.8)** se transforma en la ecuación **(2.6)**.

El error típico de la media varía proporcionalmente a la desviación standard de la población, pero varía inversamente proporcional a la raíz cuadrada del tamaño de la muestra, es decir, dado el tamaño de la muestra, cuanto mayor sea el valor de σ tanto mayor será el valor de $\sigma_{\bar{x}}$, y dado σ , cuanto mayor sea el valor de n menor será el valor de $\sigma_{\bar{x}}$. Por tanto, se deduce que cuanto mayor sea la muestra, se tendrá más certeza de que la media muestral es una buena estimación de la media poblacional.

2.4.2. Teorema central del límite.- Sean: $x_1, x_2, x_3, \dots, x_n$ un conjunto de n variables aleatorias independientes e igualmente distribuidas, tal que $E(x_i) = \mu$ y $VAR(x_i) = \sigma^2$, tienen un valor finito para $i = 1, 2, 3, \dots, n$.

$$\text{Si:} \quad Y_n = x_1 + x_2 + x_3 + \dots + x_n \quad (2.9)$$

con valor esperado y varianza:

$$E(Y_n) = n\mu \quad (2.10)$$

$$VAR(Y_n) = n\sigma^2 \quad (2.11)$$

entonces la variable aleatoria z , estandarizada de la siguiente manera:

$$z = \frac{y - n\mu}{\sigma\sqrt{n}} \quad (2.12)$$

se aproxima a una Distribución Normal con media igual a cero y varianza igual a 1, siempre y cuando "n" tienda al infinito. Esto significa que la suma de un número grande ($n \geq 30$) de variables aleatorias tendrá una Distribución Normal Standard, independiente de la distribución de probabilidad de la variable aleatoria original.

Ahora bien, efectuando operaciones algebraicas se tiene también que la ecuación (2.12) se puede expresar como:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (2.13)$$

que también se ajusta a una distribución Normal standarizada.

En otras palabras, para n grande ($n \geq 30$), la variable aleatoria $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ se

aproxima a una Distribución Normal con media 0 y varianza 1, sin importar el modelo de probabilidad a partir del cual se obtuvo la muestra.

2.5. 2.5 Cálculo del tamaño de la muestra.- Dependiendo del tamaño de la población objetivo, el cálculo de n , se distingue si la población es finita o infinita.

2.5.1 Cálculo del tamaño de muestra para poblaciones infinitas.-

2.5.1.1. Teorema o desigualdad de Tchebycheff.- Si una variable aleatoria x tiene una distribución de probabilidad conocida, se podrá conocer la media (μ) y la varianza (σ^2). Pero, si se conoce μ y σ^2 no se puede determinar la distribución de probabilidad de x , sin embargo, se puede calcular un límite superior (o inferior) para la probabilidad del tipo $p(|x - \mu| < k\sigma)$.

La desigualdad de Tchebycheff indica: Si la variable aleatoria x con función de probabilidad $f(x)$ (generalmente desconocida) tiene media y varianza conocidos, entonces para cualquier $k > 1$, se cumple que:

$$p(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad (2.14)$$

La ecuación (2.14) indica que la probabilidad de que x tome un valor dentro del intervalo $(\mu - k\sigma; \mu + k\sigma)$ es por lo menos $1 - \frac{1}{k^2}$.

Puesto que $(|x - \mu| \geq k\sigma)$ y $(|x - \mu| < k\sigma)$ son eventos complementarios también se cumple:

$$p(|x - \mu| \geq k\sigma) < \frac{1}{k^2} \quad (2.15)$$

Lo anterior significa que la probabilidad de que x tome algún valor fuera del intervalo $(\mu - k\sigma; \mu + k\sigma)$ es a lo más $1/k^2$.

La ventaja más importante de este teorema es que se aplica a todo tipo de distribución y su desventaja es que sólo proporciona un límite superior (o inferior, según sea el caso) de probabilidad.

2.5.1.2. Ley de los grandes números.- El teorema de Tchebycheff se aplica a la variable x , pero si este Teorema se aplicase a la variable \bar{x} , esta aplicación se denomina Ley de los grandes números, la cual indica:

Sean: $x_1, x_2, x_3, \dots, x_n$, n variables aleatorias independientes e igualmente distribuidas, tales que $E(x_i) = \mu$ y $VAR(x_i) = \sigma^2$, tienen un valor finito para $i = 1, 2, 3, \dots, n$, y considerando que $\bar{x} = \sum_{i=1}^n x_i / n$ es un buen estimador de μ (media poblacional).

A partir del Teorema de Tchebycheff para población:

$$p(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad (2.16)$$

Aplicando a la variable aleatoria \bar{x} , se tiene:

$$p(|\bar{x} - \mu_{\bar{x}}| < k\sigma_{\bar{x}}) \geq 1 - \frac{1}{k^2} \quad (2.17)$$

Puesto que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, se tiene:

$$p(|\bar{x} - \mu| < k \frac{\sigma}{\sqrt{n}}) \geq 1 - \frac{1}{k^2} \quad (2.18)$$

o también:

$$p(|\bar{x} - \mu| \geq k \frac{\sigma}{\sqrt{n}}) < \frac{1}{k^2} \quad (2.19)$$

O expresado de otra forma, haciendo que:

$$e = \frac{k\sigma}{\sqrt{n}} \quad (2.20)$$

$$p(|\bar{x} - \mu| < e) \geq 1 - \frac{\sigma^2}{ne^2} \quad (2.21)$$

de lo que se deduce que:

$$n \geq \left(\frac{k\sigma}{e} \right)^2 \quad (2.22)$$

La Ley de los grandes números indica que se puede determinar una muestra aleatoria de tamaño n de una población con función de probabilidad $f(x)$, tal que la probabilidad de que \bar{x} difiera de μ en menos de una cantidad arbitrariamente pequeña e , llegue a ser tan próxima a 1 cuanto más grande sea n . Es decir, si n crece, la probabilidad de que \bar{x} valga μ se acerca a 1.

2.5.2. Cálculo del tamaño de muestra para poblaciones finitas.-

Adicionalmente, a la ecuación presentada con anterioridad, deducida de la Ley de los Grandes Números, la cuales es aplicada para poblaciones infinitas, ya sea el caso en el que se traten de caracteres cualitativos o cuantitativos, existen otras dos, las cuales son utilizadas con bastante frecuencia en los estudios de mercado para cuando las poblaciones sean finitas. Dichas ecuaciones, deducidas empíricamente, son:

- Carácter cuantitativo:

$$n = \frac{Z_{\text{tablas}}^2 N \sigma^2}{Z_{\text{tablas}}^2 \sigma^2 + Ne^2} \quad (2.23)$$

- Carácter cualitativo:

$$n = \frac{Z_{\text{tablas}}^2 N P Q}{Z_{\text{tablas}}^2 P Q + Ne^2} \quad (2.24)$$

En la que:

N = tamaño de la población.

σ^2 = varianza poblacional (en caso de no disponer de ella, se utiliza su estimador respectivo)

e = error absoluto

P = proporción poblacional correspondiente al atributo de interés (en caso de no disponer de ella se estima a partir de una muestra piloto)

$Q = 1 - P$

Z_{tablas} = valor perteneciente a la Distribución Normal Standard correspondiente a un nivel de confianza $(1-\alpha)\%$, siendo los más frecuentes:

$Z_{\text{tablas}} = 2.575 \vee 1-\alpha = 99\%$

$Z_{\text{tablas}} = 1.96 \vee 1-\alpha = 95\%$

$Z_{\text{tablas}} = 1.645 \vee 1-\alpha = 90\%$

2.6. Determinación del tamaño de muestra en el caso del muestreo aleatorio estratificado.- El problema de conceder a cada estrato la adecuada representación en la muestra (conociendo de antemano el tamaño de la muestra n) se conoce con el nombre de afijación. Para tal efecto se conocen tres criterios:

2.6.1. Afijación igual.- Siendo L el número de estratos y n el tamaño de la muestra, ambos conocidos de antemano, entonces:

$$n_1 = n_2 = \dots = n_L = \frac{n}{L} \quad (2.25)$$

2.6.2. Afijación proporcional.- Considerando N_i el tamaño de la población en el estrato i -ésimo y denominando fracción de muestreo al cociente n/N , este criterio consiste en que, en cada estrato, la fracción de muestreo permanezca constante, por tanto:

$$n_i = \frac{n}{N} N_i \quad i = 1, 2, \dots, L \quad (2.26)$$

Para calcular el estimador de la media poblacional \bar{x} , se emplea la relación:

$$\bar{x} = \sum_{i=1}^n \frac{N_i x_i}{N} \quad (2.27)$$

2.6.3. Afijación óptima.- Consiste en que cada tamaño de la muestra por estrato depende del tamaño de la población en el mismo (N_i) y de la dispersión de la variable que se estudia, tomándose como medida de dicha dispersión a la desviación standard σ_i en el correspondiente estrato i -ésimo, por lo que, los valores de n_i serán:

$$n_i = \frac{N_i \sigma_i n}{\sum_{i=1}^L N_i \sigma_i} \quad (2.28)$$

El estimador de la media poblacional se calcula empleando la ecuación (2.27).

BIBLIOGRAFÍA:

- (1) **CANAVOS** George. "Probabilidad y estadística. Aplicaciones y métodos", México, 1994.
- (2) **HINES** Walter y **MONTGOMERY** David. "Probabilidad y Estadística para Ingeniería y Administración". McGraw-Hill, Mexico, 1996.
- (3) **KAZMIER** Leonard. "Estadística aplicada a Administración y Economía", McGraw-Hill, México, 1991.
- (4) **LEVIN** Richard y **RUBIN** David. "Estadística para Administradores", Prentice Hall, México, 1996
- (5) **MILLER** Irwin, **FREUND** John y **JOHNSON** Richard. "Probabilidad y estadística para ingenieros", México, 1994.
- (6) **MOYA** Rufino y **SARAVIA** Rufino. "Probabilidad e Inferencia Estadística". Perú, 1988.

=====

III TEORÍA DE LA ESTIMACIÓN ESTADÍSTICA

3.1 Introducción.- La estimación estadística consiste en el proceso de aproximar un parámetro de población desconocido, mediante un estadígrafo obtenido a partir de observaciones efectuadas en una muestra.

El proceso de estimación, básicamente, consiste en los siguientes pasos:

- a) Seleccionar un estimador para inferir el parámetro deseado del conjunto o universo bajo estudio.
- b) Seleccionar una muestra de este conjunto.
- c) Valorar al estimador de la muestra seleccionada.
- d) Inferir, de este valor, el parámetro buscado de ese universo.

La estimación estadística se divide en estimación puntual y estimación por intervalos.

3.2 Estimación puntual.- La estimación puntual consiste en estimar un sólo valor como estimación de un parámetro de población desconocido, se denomina puntual porque se utiliza un sólo punto del conjunto de todos los valores posibles.

En el caso general, si θ es el parámetro desconocido de una variable aleatoria x con distribución de probabilidad $f(x, \theta)$, y sean $x_1, x_2, x_3, \dots, x_n$, una muestra aleatoria de n valores de x tomados de esta distribución; se denominará $\hat{\theta}$ (theta circunflejo) a la estimación de θ calculada a partir de dicha muestra de n observaciones; de esta manera, $\hat{\theta}$ es un estadígrafo muestral con una distribución muestral teórica.

De todas maneras, en toda muestra existen errores, puesto que la muestra es una parte pequeña de todo el conjunto de observaciones posibles, por lo que, es muy arriesgado afirmar que el valor de un estimador obtenido a partir de una muestra es el correspondiente al valor del parámetro poblacional.

3.2.1 Propiedades que debe tener un buen estimador.- Para determinar un buen estimador se aplican cuatro propiedades: consistencia, ausencia de sesgo, eficiencia y suficiencia.

A lo largo de todo el análisis se supondrá la existencia de un sólo parámetro desconocido, sin embargo, en condiciones generales estos conceptos pueden extenderse a un número mayor de parámetros desconocidos.

3.2.1.1 Consistencia.- Es razonable esperar que un buen estimador de un parámetro θ sea cada vez mejor conforme crece el tamaño de la muestra. Esto es, a medida que

III-1

la información en una muestra aleatoria se vuelve más completa, la distribución muestral de un buen estimador se encuentra cada más concentrada alrededor del parámetro θ . Se tendrá un mejor estimador de θ si se basa en 30 observaciones que si se lo hace en 18.

Un estimador consistente es el que tiende a tener una probabilidad de acercarse al parámetro de la población a medida que el tamaño de la muestra crece, es decir, si $\hat{\theta}$ es un estadígrafo muestral calculado a partir de una muestra de tamaño n y θ es el parámetro de la población que se va a estimar, entonces, $\hat{\theta}$ es un estimador consistente de θ si, para todo número positivo arbitrariamente pequeño ϵ , se cumple la ecuación (3.1).

$$\lim_{n \rightarrow \infty} (P|\hat{\theta} - \theta| \leq \epsilon) = 1 \quad (3.1)$$

La ecuación (3.1) se denomina convergencia en probabilidad, es decir, si un estimador es consistente converge en probabilidad al valor del parámetro que está intentando estimar conforme el tamaño de la muestra crece.

3.2.1.2 Ausencia de sesgo.- Para comprender mejor esta propiedad, se define el Error Cuadrático Medio de $\hat{\theta}$ como $E[(\hat{\theta} - \theta)^2]$, es decir, el Error Cuadrático Medio es el valor esperado del cuadrado de la diferencia entre θ y $\hat{\theta}$.

Desarrollando la expresión anterior y efectuando operaciones se tiene:

$$E[(\hat{\theta} - \theta)^2] = \sigma_{\hat{\theta}}^2 + [\theta - E(\hat{\theta})]^2 \quad (3.2)$$

La ecuación (3.2) significa que el error cuadrático medio es la suma de 2 cantidades no negativas: $\sigma_{\hat{\theta}}^2$ es la varianza del estimador y el término $[\theta - E(\hat{\theta})]$, el cual se denomina sesgo del estimador, elevado al cuadrado.

Es deseable que el error cuadrático medio sea lo más pequeño posible, para lo cual la varianza del estimador ($\sigma_{\hat{\theta}}^2$) debe ser lo más pequeña posible y el sesgo próximo a cero o cero.

En vista de que la varianza del estimador ($\sigma_{\hat{\theta}}^2$) no es posible controlar, lo deseable será tener un estimador cuyo sesgo sea cero, estimador al que se denominará insesgado.

Puesto que $\hat{\theta}$, estimador de θ , es una variable aleatoria, como tal tiene una distribución de probabilidad con media y varianza, se dice, que $\hat{\theta}$ es un estimador

insesgado de θ , si el valor esperado de $\hat{\theta}$ es igual a θ , es decir, si:

$$E(\hat{\theta}) = \theta \quad (3.3)$$

Dicho de otra forma, es de esperar que si se toman muchas muestras de tamaño dado partiendo de la misma distribución, y si de cada una se obtiene un valor de $\hat{\theta}$, la media aritmética de todos los valores de $\hat{\theta}$ han de estar muy cerca de θ .

3.2.1.3 Eficiencia.- Un estimador $\hat{\theta}$ es eficiente, si entre todos los estimadores insesgados, tiene varianza más pequeña. Dicho estimador también se llama estimador insesgado de varianza mínima.

En otras palabras, suponiendo que de la misma muestra se obtienen 2 estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ y, ambos son estimadores insesgados de θ ; además, si por ejemplo la varianza de $\hat{\theta}_1$ es menor que la varianza de $\hat{\theta}_2$, se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$, por que sus valores están más cerca de θ que los de $\hat{\theta}_2$.

3.2.1.4 Suficiencia.- Un estimador suficiente del parámetro θ , es aquel que utiliza toda la información pertinente sobre θ que se puede disponer de la muestra.

Por ejemplo, si se toma una muestra de 30 observaciones con el fin de estimar μ , y si \hat{x}_1 es el promedio de la primera y última observaciones, \hat{x}_2 es el promedio de las 10 primeras observaciones y \hat{x}_3 es el promedio de las 5 observaciones centrales, se concluye que \hat{x}_2 es el estimador suficiente entre los 3 estimadores calculados.

3.2.2 Estimación por el método de máxima verosimilitud.- Aunque un experimentador decide sobre qué propiedades desea que posea un estimador, tiene que enfrentarse con el problema de cómo obtener dichos estimadores. Uno de los más utilizados es el método de máxima verosimilitud.

Básicamente, el método de estimación por máxima verosimilitud selecciona como estimador a aquel valor del parámetro que tiene la propiedad de maximizar el valor de la probabilidad de la muestra aleatoria observada.

El procedimiento consiste en considerar todos los valores imaginables del parámetro de población, que se encuentran en la muestra, y calcular la probabilidad de que se hubiera obtenido el estadígrafo muestral particular, dados todos los valores imaginables del parámetro.

Sea una variable aleatoria cuya función de cuantía o densidad $f(\mathbf{x})$, y con un sólo parámetro θ ; suponiendo que se efectúa n veces el experimento correspondiente, con lo que se obtiene una muestra de n números: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n$.

Además, si existe independencia de los n ensayos, entonces la probabilidad de que una muestra de tamaño n conste precisamente de estos n valores está expresada por una función $L(\theta)$, función que se denomina función de verosimilitud y que se muestra en la ecuación (3.4).

$$L(\theta) = f(\mathbf{x}_1; \theta) * f(\mathbf{x}_2; \theta) * f(\mathbf{x}_3; \theta) * \dots * f(\mathbf{x}_n; \theta) \quad (3.4)$$

Los valores $L(\theta) = f(\mathbf{x}_1; \theta) * f(\mathbf{x}_2; \theta) * f(\mathbf{x}_3; \theta) * \dots * f(\mathbf{x}_n; \theta)$ dependen del parámetro θ , luego, L depende de $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n$ y θ . Si $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n$ son constantes y conocidos, L será función sólo de θ .

La estimación por la máxima verosimilitud consiste en hallar el valor de θ de manera que L tenga un valor máximo, para lo que será necesario derivar L respecto de θ , es decir:

$$\frac{\partial L}{\partial \theta} = 0 \quad (3.5)$$

obteniendo el estimador $\hat{\theta}$, llamado estimador máximo verosímil de θ .

En virtud a que $L(\theta)$, $\ln L(\theta)$ y $\log L(\theta)$ tienen su máximo para el mismo valor de θ , en la mayor parte de los casos es posible utilizar esta propiedad para facilitar los cálculos. Por lo que se tiene:

$$\frac{\partial \ln L}{\partial \theta} = 0 \quad (3.6)$$

$$\frac{\partial \log L}{\partial \theta} = 0 \quad (3.7)$$

Para los casos en que existen varios parámetros, la función de máxima verosimilitud es:

$$L(\theta_1, \theta_2, \dots, \theta_k) = f(\mathbf{x}_1; \theta_1, \theta_2, \dots, \theta_k) * f(\mathbf{x}_2; \theta_1, \theta_2, \dots, \theta_k) * \dots * f(\mathbf{x}_n; \theta_1, \theta_2, \dots, \theta_k) \quad (3.8)$$

Si se satisfacen ciertas condiciones de regularidad, el punto en que la

verosimilitud es máxima es una solución del sistema de **k** ecuaciones compuesta por:

$$\frac{\partial L}{\partial \theta_1} = 0 \quad (3.9)$$

$$\frac{\partial L}{\partial \theta_1} = 0 \quad (3.10)$$

.....
.....

$$\frac{\partial L}{\partial \theta_k} = 0 \quad (3.11)$$

También en este caso puede ser más fácil trabajar con el logaritmo (natural o decimal) de la función de verosimilitud.

Este método tiene la propiedad de proporcionar estimadores que son funciones de estadísticas suficientes, siempre y cuando el estimador máximo verosímil sea único. Además, también proporciona un estimador eficiente, si es que existe. Sin embargo, la mayoría de estos estimadores son sesgados.

La desventaja de este método radica en el hecho de que no da medida alguna de la precisión de la estimación y no indica la magnitud del error en que se puede incurrir.

3.3 Estimación por intervalos.- La estimación por intervalos describe un intervalo de valores dentro del cual es posible que se encuentre un parámetro poblacional, más propiamente, consiste en determinar un intervalo **(a,b)** que comprende un parámetro de población θ con cierta probabilidad **(1- α)** , es decir:

$$p(a < \theta < b) = 1 - \alpha \quad (3.12)$$

En esta expresión:

- **a** y **b** son variables aleatorias que dependen del estimador $\hat{\theta}$ y que se denominan: límite de confianza inferior y límite de confianza superior, respectivamente.
- Al intervalo **(a,b)** se denomina intervalo de confianza y es un estimador de intervalo que se construye respecto a $\hat{\theta}$ y que permite especificar el alcance de la estimación que se está efectuando.
- **b-a** es una medida de la precisión.
- **(1- α)** se denomina nivel de confianza y representa la “confianza” ó probabilidad de que en ese intervalo se incluya el parámetro que se estima. Una probabilidad más alta representa más confianza.

Para tal efecto, se puede construir distintos intervalos de confianza, ya sean unilaterales o bilaterales:

- Intervalo de confianza para la media aritmética.
- Intervalo de confianza para la diferencia de dos medias aritméticas.
- Intervalo de confianza para la proporción.
- Intervalo de confianza para la varianza.
- Intervalo de confianza para la razón de dos varianzas.

3.3.1 Intervalo de confianza bilateral para la media aritmética de la

población.- Para estimar un intervalo de confianza para μ , se toma una muestra aleatoria de n observaciones: $x_1, x_2, x_3, \dots, x_n$, y de dicha muestra se calcula el estimador puntual \bar{x} .

En el cuadro (3.1) se muestran los intervalos de confianza para la media poblacional tanto para Distribuciones Normales como para las que no lo son.

Cuadro (3.1)
INTERVALOS DE CONFIANZA PARA ESTIMAR LA
MEDIA ARITMÉTICA DE LA POBLACION

DISTRIBUCION DE LA POBLACIÓN	TAMAÑO DE MUESTRA	σ^2 CONOCIDO	σ^2 DESCONOCIDO
Normal	Grande ($n \geq 30$)	$\bar{X} \pm Z_{\text{tablas}} \sigma_{\bar{x}}$	$\bar{X} \pm Z_{\text{tablas}} S_{\bar{x}}$
Normal	Pequeña ($n < 30$)	$\bar{X} \pm Z_{\text{tablas}} \sigma_{\bar{x}}$	$\bar{X} \pm t_{\text{tablas}} S_{\bar{x}}$
Cualquiera	Grande ($n \geq 30$)	$\bar{X} \pm Z_{\text{tablas}} \sigma_{\bar{x}}$	$\bar{X} \pm Z_{\text{tablas}} S_{\bar{x}}$

FUENTE: Elaboración propia.

En la que:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}} \quad (3.13)$$

- $\hat{\sigma} = s$ = estimador de la desviación standard poblacional = desviación standard muestral.

$$\hat{\sigma}_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (3.14)$$

- $\hat{\sigma}_{\bar{x}} = s_{\bar{x}}$ = estimador de la desviación standard de la distribución muestral del estadígrafo \bar{x} .

- z_{tablas} = valor absoluto de “z” perteneciente a la Distribución Normal Standardizada correspondiente a un valor de $(1-\alpha)$ central.
- t_{tablas} = valor absoluto de “t” perteneciente a la Distribución “t” correspondiente a un valor de $(1-\alpha)$ central con $v = n-1$ grados de libertad.

3.3.2 Intervalo de confianza bilateral para la diferencia de dos medias aritméticas poblacionales.-

En el cuadro (3.2), considerando dos muestras aleatorias de tamaños n_1 y n_2 respectivamente, se presentan los intervalos de confianza para la diferencia entre medias aritméticas de dos distribuciones ($\mu_1 - \mu_2$).

Cuadro (3.2)

INTERVALOS DE CONFIANZA PARA ESTIMAR LA DIFERENCIA ENTRE MEDIAS ARITMÉTICAS DE DOS POBLACIONES

DISTRIBUCION DE POBLACIÓN	TAMAÑO DE MUESTRAS	σ_1^2 y σ_2^2 CONOCIDOS	σ_1^2 y σ_2^2 DESCONOCIDOS
Normal	$(n_1, n_2 \geq 30)$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\text{tablas}} \sigma_{\bar{x}_1 - \bar{x}_2}$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\text{tablas}} s_{\bar{x}_1 - \bar{x}_2}$
Normal	$(n_1, n_2 < 30)$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\text{tablas}} \sigma_{\bar{x}_1 - \bar{x}_2}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\text{tablas}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Cualquiera	$(n_1, n_2 \geq 30)$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\text{tablas}} \sigma_{\bar{x}_1 - \bar{x}_2}$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\text{tablas}} s_{\bar{x}_1 - \bar{x}_2}$

FUENTE: Elaboración propia.

En la que:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (3.15)$$

- $\sigma_{\bar{x}_1 - \bar{x}_2}$ = desviación standard de la distribución muestral de la diferencia de dos medias muestrales.

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (3.16)$$

- $s_{\bar{x}_1 - \bar{x}_2}$ = estimador de la desviación standard de la distribución muestral de la diferencia de dos medias muestrales.

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad (3.17)$$

- S_p = estimador combinado de la desviación standard de la distribución muestral de la diferencia de dos medias muestrales.
- t_{tablas} = valor absoluto de “t” perteneciente a la Distribución “t” correspondiente a un valor de $(1-\alpha)$ central con v grados de libertad.

$$v = n_1 + n_2 - 2 \quad (3.18)$$

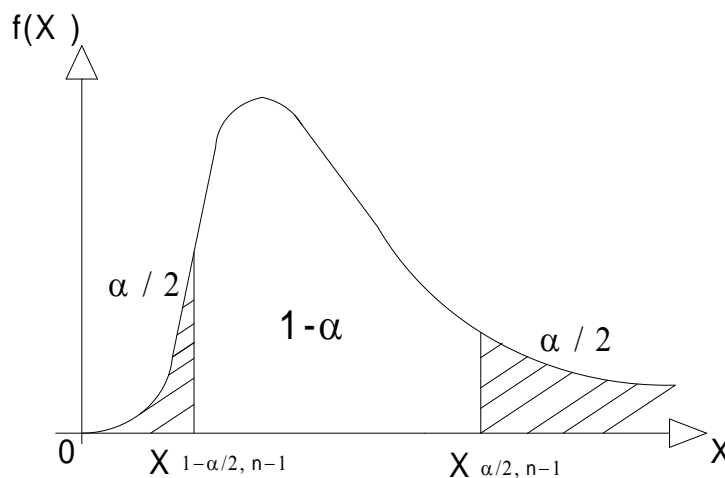
3.3.3 Intervalo de confianza bilateral para la varianza de una

Distribución Normal.- Para estimar un intervalo de confianza para σ^2 que pertenece a una Distribución Normal, se toma una muestra aleatoria de n observaciones: $x_1, x_2, x_3, \dots, x_n$, y de dicha muestra se calcula el estimador puntual S^2 .

Es posible demostrar que la variable:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad (3.19)$$

pertenece a una Distribución Chi cuadrado con $(n-1)$ grados de libertad, tal como se muestra en el siguiente gráfico.



Para desarrollar el intervalo de confianza, se puede observar del gráfico:

$$p\left(\chi_{1-\frac{\alpha}{2}, n-1}^2 \leq \chi^2 \leq \chi_{\frac{\alpha}{2}, n-1}^2\right) = 1 - \alpha \quad (3.20)$$

Efectuando operaciones se tiene:

$$p\left(\chi^2_{1-\frac{\alpha}{2}, n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2}, n-1}\right) = 1 - \alpha \quad (3.21)$$

$$p\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}\right) = 1 - \alpha \quad (3.22)$$

3.3.4 Intervalo de confianza bilateral para la proporción de una

Distribución Binomial.- Considerando que se ha tomado una muestra aleatoria de n observaciones de una población con Distribución Binomial con parámetros n y p ; para estimar el valor de p , se obtiene x' observaciones en esta muestra que pertenecen a la clase de interés y se utiliza el estimador puntual:

$$\hat{p} = \frac{x'}{n} \quad (3.23)$$

Es posible demostrar que:

$$E(\hat{p}) = \mu_p = p \quad (3.24)$$

$$V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n} \quad (3.25)$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (3.26)$$

Solamente para el caso de tener un tamaño de muestra grande ($n \geq 30$), aplicando el Teorema Central del Límite y por analogía con el caso de la estimación de la media aritmética para el caso de una distribución cualquiera (con $n \geq 30$ y varianza conocida), el intervalo de confianza para la proporción es:

$$\hat{p} \pm z_{tab} \sigma_{\hat{p}} \quad (3.27)$$

Ahora bien, puesto que en la expresión (3.26) se desconoce “ p ”, se reemplaza por su estimador \hat{p} , por lo cual se tiene el intervalo (3.28).

$$\hat{p} \pm z_{tablas} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (3.28)$$

BIBLIOGRAFÍA:

- (1) **HINES** Walter y **MONTGOMERY** David (1996): “*Probabilidad y Estadística para Ingeniería y Administración*”. McGraw-Hill, México.

- (2) **KAZMIER** Leonard (1991): "*Estadística aplicada a la administración y economía*". McGraw-Hill, México.
- (3) **LEVIN** Richard y **RUBIN** David (1996): "*Estadística para administradores*". Prentice-Hall, México
- (4) **MILLER** Irwin, **FREUND** John y **JOHNSON** Richard (1994): "*Probabilidad y estadística para ingenieros*", México.
- (5) **MOYA** Rufino (1988): "*Estadística Descriptiva*". Perú.
- (6) **TRIOLA** Mario F. (2000): "*Estadística elemental*". Prentice-Hall, México.

=====

INDICE**Pág.**

4.1	Introducción.....	1
4.2	Estimación puntual.....	1
4.2.1	Propiedades que debe tener un buen estimador.....	1
4.2.1.1	Consistencia.....	2
4.2.1.2	Ausencia de sesgo.....	2
4.2.1.3	Eficiencia.....	3
4.2.1.4	Suficiencia.....	3
4.2.2	Estimación por el método de máxima verosimilitud.....	3
4.3	Estimación por intervalos.....	5
4.3.1	Intervalos de confianza bilaterales para la media de la población.....	6
4.3.2	Intervalos de confianza bilaterales para la diferencia de dos medias poblacionales.....	7
4.3.3	Intervalo de confianza bilateral para la varianza de una Distribución Normal.....	8
4.3.4	Intervalo de confianza bilateral para la proporción de una Distribución Binomial.....	8

IV PRUEBAS DE HIPÓTESIS ESTADÍSTICAS

4.1 Introducción.- La inferencia relativa a un parámetro cualquiera de una población suele hacerse a través de 2 métodos: estimando el parámetro en base de una muestra aleatoria o realizando una prueba sobre la aceptación o refutación del valor del parámetro. En este capítulo se estudiará el segundo método: la prueba o contraste de hipótesis estadística.

4.2 Conceptos básicos.- A continuación se detallan las principales definiciones referidas a las pruebas estadísticas.

4.2.1 Hipótesis estadística.- Una hipótesis estadística es un enunciado que se hace acerca de la distribución de probabilidad de una o más variables aleatorias. Las hipótesis estadísticas a menudo involucran uno ó más parámetros.

Se puede especificar una hipótesis indicando el tipo de distribución y el valor o valores del parámetro que la definen. En la práctica, la distribución de población, generalmente se asume, por tanto, una hipótesis se especifica con el valor o los valores del parámetro.

4.2.2 Hipótesis nula e hipótesis alterna.- La hipótesis nula, denotada por H_0 , es la hipótesis estadística que se desea probar; mientras que, la hipótesis alterna, denotada por H_1 , es una suposición de lo que sería si es que no se cumple la hipótesis nula.

La hipótesis nula suele determinarse de tres maneras:

- Puede resultar de la experiencia o conocimiento pasado del futuro.
- Puede determinarse a partir de alguna teoría o modelo.
- Cuando el valor del parámetro poblacional es resultado de consideraciones experimentales.

Una hipótesis nula debe considerarse como verdadera a menos que existiera suficiente evidencia en contra (evidencia que es proporcionada por la muestra).

4.2.3 Prueba de hipótesis estadística.- La prueba de hipótesis estadística es una metodología que, en base de los valores experimentales observados, conduce a una decisión, ya sea aceptar o rechazar una hipótesis bajo consideración.

Existen dos tipos de pruebas, las pruebas unilaterales y las pruebas bilaterales.

4.2.3.1 Pruebas unilaterales.- Estas pruebas se clasifican en:

4.2.3.1.1 Prueba de la cola inferior o cola izquierda.- En este caso las hipótesis se plantean de la siguiente forma:

$$\begin{array}{lll} H_0: \theta \geq a & H_0: \theta > a & H_0: \theta = a \\ H_1: \theta < a & H_1: \theta < a & H_1: \theta < a \end{array}$$

4.2.3.1.2 Prueba de la cola superior o prueba de la cola derecha.- Para este caso las hipótesis se plantean de la siguiente manera:

$$\begin{array}{lll} H_0: \theta \leq a & H_0: \theta < a & H_0: \theta = a \\ H_1: \theta > a & H_1: \theta > a & H_1: \theta > a \end{array}$$

4.2.3.2 Pruebas bilaterales o prueba de dos colas.- En este caso, las hipótesis se formulan de la siguiente forma:

$$\begin{array}{l} H_0: \theta = a \\ H_1: \theta \neq a \end{array}$$

4.2.4 Tipos de errores.- La decisión para aceptar o rechazar la hipótesis nula (H_0) se basa en los datos de la muestra aleatoria. Cuando se toma una decisión utilizando la información de una muestra aleatoria esta decisión se encuentra sujeta a error. En las pruebas de hipótesis pueden cometerse dos tipos de errores: error del tipo I y error del tipo II.

4.2.4.1 Error tipo I.- El error de tipo I se comete cuando se rechaza la hipótesis nula (H_0) siendo que en realidad es verdadera. La probabilidad de cometer el error de tipo I es igual a α , es decir, es el nivel de significación. Los niveles de significación o significancia más utilizados son: 10%, 5% y 1%.

El nivel de confianza es el complemento del nivel de significación, de tal forma que se cumple la ecuación (4.1).

$$\text{nivel de confianza} + \text{nivel de significación} = 1 = 100\% \quad (4.1)$$

4.2.4.2 Error tipo II.- El error de tipo II se comete cuando se acepta la hipótesis nula (H_0) cuando en realidad es falsa. La probabilidad de cometer el error tipo II se representa por β .

En el cuadro (4.1) se muestran todas las situaciones que se pueden presentar en la toma de decisiones.

CUADRO (4.1)
OPCIONES QUE SE PRESENTAN EN LA TOMA DE DECISIONES

DECISION	H_0 VERDADERA	H_1 VERDADERA
ACEPTAR H_0	Decisión correcta	Error tipo II
RECHAZAR H_0	Error tipo I	Decisión correcta

Por ejemplo, si:

H_0 = el medicamento XYZ no es peligroso.

H_1 = el medicamento XYZ es peligroso.

- Si H_0 es verdadera y se acepta, se toma una decisión correcta.
- Si H_0 es falsa y se rechaza, se toma una decisión correcta.
- Si H_0 es falsa (es decir, el medicamento es peligroso) y se acepta, se lanza al mercado una droga peligrosa. En este caso se comete un error del tipo II.
- Si H_0 es verdadera y se la rechaza, se está eliminado en el sector salud a un medicamento que podría ser benéfico. Se dice que en este caso se comete un error del tipo I.

Se ha demostrado que para cualquier tamaño de muestra, la probabilidad de cometer un error tipo I guarda una proporción inversa a la probabilidad de cometer uno del tipo II (si α disminuye, β aumenta y viceversa). La probabilidad de cometer simultáneamente ambos errores decrece a medida que el tamaño de muestra aumenta; sin embargo, a un aumento en el tamaño de la muestra corresponde un aumento en el costo del procedimiento.

4.2.5 Estadígrafo de prueba.- El estadígrafo de prueba es el estimador insesgado del parámetro que se prueba (obtenido de una muestra), el cual se transforma posteriormente, para comparar con los valores de tablas.

Por ejemplo, para probar el valor hipotético de la media poblacional (μ), se considera la media de una muestra aleatoria (\bar{x}) de dicha población como estadígrafo de prueba, para posteriormente standarizarlo.

4.2.6 Regiones de aceptación y rechazo.- La región de aceptación es la región que contiene los valores de la variable standarizada para las cuales se da por válida la hipótesis nula.

La región de rechazo o región crítica es la región que lleva al rechazo de la hipótesis nula en consideración, lo cual significa aceptar la hipótesis alterna.

4.3 Etapas básicas en una prueba de hipótesis.- En todas las pruebas de hipótesis los pasos que se siguen son:

- 1º Plantear la hipótesis nula y la hipótesis alterna.
- 2º Especificar el nivel de significación a utilizar.
- 3º Elegir el estadígrafo de prueba más adecuado y su correspondiente transformación.
- 4º Establecer la región de aceptación y rechazo, especificando el o los valores críticos.
- 5º Calcular el estadígrafo de prueba empleando una muestra aleatoria de tamaño n y obtener su correspondiente transformación.
- 6º Tomar una decisión, es decir, aceptar o rechazar H_0 .

4.4 Prueba de hipótesis sobre la media aritmética de una Distribución con varianza conocida.- En este caso se utiliza la Distribución Normal considerando la standarización que se muestra en la ecuación (4.2).

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (4.2)$$

En el cuadro (4.2) se muestran las regiones de aceptación para cada caso.

CUADRO (4.2)
REGIONES DE ACEPTACIÓN PARA LA MEDIA ARITMÉTICA DE UNA DISTRIBUCIÓN CON VARIANZA CONOCIDA

Parámetro	Distribución	n	Hipótesis	Región de aceptación
μ	Normal	$n \geq 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-Z_{1-\alpha/2}; +Z_{1-\alpha/2}]$
μ	Normal	$n \geq 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[Z_{\alpha}; +\infty]$
μ	Normal	$n \geq 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; Z_{1-\alpha}]$
μ	Normal	$n < 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-Z_{1-\alpha/2}; +Z_{1-\alpha/2}]$
μ	Normal	$n < 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[Z_{\alpha}; +\infty]$
μ	Normal	$n < 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; Z_{1-\alpha}]$
μ	Cualquiera	$n \geq 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-Z_{1-\alpha/2}; +Z_{1-\alpha/2}]$

μ	Cualquiera	$n \geq 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[z_{\alpha}; +\infty]$
μ	Cualquiera	$n \geq 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; z_{1-\alpha}]$
μ	Cualquiera	$n < 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	Se aplican pruebas no paramétricas
μ	Cualquiera	$n < 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	Se aplican pruebas no paramétricas
μ	Cualquiera	$n < 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	Se aplican pruebas no paramétricas

Fuente: Elaboración propia.

4.5 Prueba de hipótesis sobre la media aritmética de una distribución con varianza desconocida.

- La Distribución "t" es apropiada a aplicar cuando la muestra proviene de una distribución con varianza desconocida y la variable pertenece a una Distribución Normal o se puede aproximar a ésta (cuando el tamaño de muestra es menor a 30). En este caso el estadígrafo de prueba es \bar{x} y su transformación es la que se muestra en la ecuación (4.3).

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \quad (4.3)$$

En el cuadro (4.3) se muestran las regiones de aceptación para cada caso.

CUADRO (4.3)
REGIONES DE ACEPTACIÓN PARA LA MEDIA ARITMÉTICA DE UNA DISTRIBUCIÓN CON VARIANZA DESCONOCIDA

Parámetro	Distribución	n	Hipótesis	Región de aceptación
μ	Normal	$n \geq 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-t_{1-\alpha/2, n-1}; t_{1-\alpha/2, n-1}]$
μ	Normal	$n \geq 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[t_{\alpha, n-1}; +\infty]$
μ	Normal	$n \geq 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; t_{1-\alpha, n-1}]$
μ	Normal	$n < 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-t_{1-\alpha/2, n-1}; t_{1-\alpha/2, n-1}]$
μ	Normal	$n < 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[t_{\alpha, n-1}; +\infty]$
μ	Normal	$n < 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; t_{1-\alpha, n-1}]$
μ	Cualquiera	$n \geq 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	$[-t_{1-\alpha/2, n-1}; t_{1-\alpha/2, n-1}]$

μ	Cualquiera	$n \geq 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	$[t_{\alpha, n-1}; +\infty]$
μ	Cualquiera	$n \geq 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	$[-\infty; t_{1-\alpha, n-1}]$
μ	Cualquiera	$n < 30$	$H_0: \mu = a$ $H_1: \mu \neq a$	Se aplican pruebas no paramétricas
μ	Cualquiera	$n < 30$	$H_0: \mu \geq a$ $H_1: \mu < a$	Se aplican pruebas no paramétricas
μ	Cualquiera	$n < 30$	$H_0: \mu \leq a$ $H_1: \mu > a$	Se aplican pruebas no paramétricas

Fuente: Elaboración propia.

4.6 Prueba de hipótesis sobre la varianza de una Distribución Normal.- En este caso se ha demostrado que la Distribución Chi Cuadrado es la más adecuada para efectuar pruebas de hipótesis sobre la varianza de una Distribución Normal.

Para efectuar la prueba sobre la varianza de una Distribución Normal, el estadígrafo a utilizar será el estimador insesgado de la varianza poblacional:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (4.4)$$

Posteriormente se debe efectuar la siguiente transformación:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad (4.5)$$

En el cuadro (4.4) se muestran las regiones de aceptación para cada caso.

CUADRO (4.4)
REGIONES DE ACEPTACIÓN PARA LA VARIANZA
DE UNA DISTRIBUCIÓN NORMAL

Parámetro	Distribución	Hipótesis	Región de aceptación
σ^2	Normal	$H_0: \sigma^2 = a$ $H_1: \sigma^2 \neq a$	$[\chi^2_{1-\alpha/2, n-1}; \chi^2_{\alpha/2, n-1}]$
σ^2	Normal	$H_0: \sigma^2 \geq a$ $H_1: \sigma^2 < a$	$[\chi^2_{1-\alpha, n-1}; +\infty]$
σ^2	Normal	$H_0: \sigma^2 \leq a$ $H_1: \sigma^2 > a$	$[0; \chi^2_{\alpha, n-1}]$

Fuente: Elaboración propia.

4.7 Prueba de hipótesis sobre la proporción de una Distribución

Binomial.- En este caso se efectuará la prueba solamente para el caso en que $n \geq 30$ (lo que implica que se aproxima a una Distribución Normal). Para efectuar la prueba sobre la proporción, se utilizará el valor de x' (número de elementos de una determinada característica, en una muestra de tamaño n). El estadígrafo de prueba es el que se muestra en la ecuación (4.6).

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} \quad (4.6)$$

Reemplazando las ecuaciones (3.24) y (3.26):

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad (4.7)$$

O también se tiene:

$$z = \frac{x' - np}{\sqrt{np(1-p)}} \quad (4.8)$$

En el cuadro (4.5) se muestran las regiones de aceptación para cada caso.

CUADRO (4.5)
REGIONES DE ACEPTACIÓN PARA LA PROPORCIÓN
DE UNA DISTRIBUCIÓN BINOMIAL

Parámetro	Distribución	n	Hipótesis	Región de aceptación
p	Binomial	$n \geq 30$	$H_0: p = a$ $H_1: p \neq a$	$[-z_{1-\alpha/2}; +z_{1-\alpha/2}]$
p	Binomial	$n \geq 30$	$H_0: p \geq a$ $H_1: p < a$	$[z_{\alpha}; +\infty]$
p	Binomial	$n \geq 30$	$H_0: p \leq a$ $H_1: p > a$	$[-\infty; z_{1-\alpha}]$

Fuente: Elaboración propia.

BIBLIOGRAFÍA:

- (1) **HINES** Walter y **MONTGOMERY** David (1996): "Probabilidad y Estadística para Ingeniería y Administración". McGraw-Hill, México.
- (2) **KAZMIER** Leonard (1991): "Estadística aplicada a la administración y economía". McGraw-Hill, México.
- (3) **LEVIN** Richard y **RUBIN** David (1996): "Estadística para administradores". Prentice-Hall, México.

- (4) **MILLER** Irwin, **FREUND** John y **JOHNSON** Richard (1994): "*Probabilidad y estadística para ingenieros*", México.
- (5) **MOYA** Rufino (1988): "*Probabilidad e inferencia estadística*". Perú.

=====

INDICE

	Pag.
5.1 Introducción.....	1
5.2 Conceptos básicos.....	1
5.2.1 Hipótesis estadística.....	1
5.2.2 Hipótesis nula e hipótesis alterna.....	1
5.2.3 Prueba de hipótesis estadística.....	1
5.2.3.1 Pruebas unilaterales.....	2
5.2.3.1.1 Prueba de la cola inferior o cola izquierda.....	2
5.2.3.1.2 Prueba de la cola superior o cola derecha.....	2
5.2.3.2 Pruebas bilaterales o prueba de dos colas.....	2
5.2.4 Tipos de errores.....	2
5.2.4.1 Error tipo I.....	2
5.2.4.2 Error tipo II.....	2
5.2.5 Estadística de prueba.....	3
5.2.6 Regiones de aceptación y rechazo.....	3
5.3 Etapas básicas en una prueba de hipótesis.....	4
5.4 Prueba de hipótesis sobre la media de una Distribución con varianza conocida.....	4
5.5 Prueba de hipótesis sobre la media de una distribución con varianza desconocida.....	5
5.6 Prueba de hipótesis sobre la proporción de una Distribución Binomial..	5
5.7 Prueba de hipótesis sobre la varianza de Distribución Normal.....	6
5.8 Análisis de varianza.....	7
5.8.1 Análisis de experimentos estadísticos.....	7
5.8.2 Análisis de varianza.....	7
5.8.3 Análisis de varianza con un criterio o factor de clasificación....	8
5.8.3.1 Prueba de hipótesis.....	8
5.8.3.2 Diagnóstico y validación del modelo.....	10
5.8.3.2.1 Distribución de residuos.....	10
5.8.3.2.2 Relación entre el valor de los residuos y el valor esperado de la respuesta.....	11
5.8.3.2.3 Relación entre los residuos y el tiempo.....	11

V ANÁLISIS DE REGRESIÓN Y CORRELACIÓN LINEAL

5.1. Introducción.- En este capítulo se examinarán las asociaciones cuantitativas entre un determinado número de variables, así como el grado de relación existente entre dichas variables, es decir, se examinarán técnicas que permitan ajustar una ecuación de algún tipo al conjunto de datos dado, con el propósito de obtener una ecuación empírica de predicción razonablemente precisa.

5.2. Análisis de regresión.- El objetivo principal del análisis de regresión es estimar el valor de una variable aleatoria (llamada variable dependiente o variable respuesta) conociendo el valor de un grupo de variables asociadas (llamadas variables independientes ó de predicción). La ecuación de regresión es la fórmula algebraica mediante la cual se estima el valor de la variable dependiente.

Dicha ecuación que se obtiene de esta forma puede tener algunas limitaciones con respecto a su interpretación física, sin embargo, en un medio empírico, será muy útil si puede proporcionar una adecuada capacidad de predicción para la respuesta en el interior de una región específica de las variables de predicción. Como ejemplos de variables se tiene: relación entre el peso y la altura de los seres humanos, relación entre la temperatura ambiente y el consumo de energía eléctrica, etc..

Las suposiciones principales en que se basa el modelo de regresión son:

- La variable dependiente es una variable aleatoria, pero no es necesario que las variables independientes sean variables aleatorias.
- La relación entre las diversas variables independientes y la variable dependiente es lineal.
- La variable dependiente tiene una Distribución Normal con varianza constante. Si bien la primera suposición no es crítica, la suposición de varianza constante es crucial. Una estimación insesgada de σ es el error standard de estimación.

El modelo de regresión propuesto debe ser relativamente sencillo y deberá contener pocos parámetros. Un procedimiento muy útil para la selección inicial cuando se tiene sólo una variable de predicción es graficar la variable dependiente contra la variable independiente.

Las ecuaciones que más se utilizan para relacionar 2 ó más variables son:

- Lineal simple:

$$y = a + bx \quad (5.1)$$

- Lineal inversa:

$$y = a + \frac{b}{x} \quad (5.2)$$

- Lineal logarítmica natural:

$$y = a + b \ln(x) \quad (5.3)$$

- Exponencial:

$$y = ab^x \quad (5.4)$$

- Potencial:

$$y = ax^b \quad (5.5)$$

- Lineal múltiple:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k \quad (5.6)$$

- Lineal polinomial:

$$y = a + b_1x_1 + b_2x_2^2 + \dots + b_kx_k^k \quad (5.7)$$

5.2.1 Método de estimación de parámetros por mínimos cuadrados.-

Este método se aplica siempre y cuando *la función sea de carácter lineal o se encuentre linealizada*. El método halla las estimaciones para los parámetros en la ecuación seleccionada mediante la minimización de la suma de los cuadrados de las diferencias entre los valores observados de la variable dependiente y de aquellos proporcionados por la ecuación de regresión.

$$z = \sum e_i^2 \quad (5.8)$$

$$e_i = V_{o,i} - V_{c,i} \quad (5.9)$$

$$\frac{\partial z}{\partial a} = 0$$

$$\frac{\partial z}{\partial b} = 0$$

$$\frac{\partial z}{\partial c} = 0$$

.....

En la que:

e_i = error o residuo de la observación "i".

$V_{o,i}$ = valor observado "i" de la variable dependiente

$V_{c,i}$ = valor calculado "i" de la variable dependiente

a = término independiente

b, c, d, \dots = coeficientes de las variables independientes

La constante "a" en la ecuación de regresión se refiere al valor de la ordenada al origen en el caso lineal con una variable independiente; en el caso de la regresión múltiple y polinomial, es el valor de la variable dependiente cuando todas las variables independientes son iguales a cero.

Cuando se obtiene una ecuación de regresión por el método de mínimos cuadrados, surgen una serie de propiedades, algunas de las cuales son:

$$\sum e_i = 0 \quad (5.10)$$

$$\sum v_{o,i} = \sum v_{c,i} \quad (5.11)$$

$$\sum_{j=1}^k x_{ij} e_j = 0 \quad \forall \quad j = 1, 2, 3, \dots, k. \quad (5.12)$$

5.2.2 Error standard de estimación.- El error standard de estimación o desviación standard residual es una medida de cuan buena es la recta estimada de regresión a las observaciones. Por tanto, cuanto más pequeño sea este valor, el modelo se ajustará mejor a los datos.

El error standard de estimación se calcula con la ecuación (5.13).

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - k - 1}} \quad (5.13)$$

En la que:

n = número de observaciones.

k = número de variables independientes.

El valor de S_{yx} viene expresado en las mismas unidades que la variable dependiente y el cuadrado de dicho valor (S_{yx}^2) se denomina varianza residual.

5.2.3 Prueba de hipótesis para coeficientes de regresión.- La pruebas de hipótesis para coeficientes de regresión se efectúa con el objetivo de conocer si cada una de las variables independientes se debe incluir o no en la ecuación de regresión, es decir, si existe alguna relación entre las dos variables (entre la variable dependiente y la correspondiente variable independiente analizada). Para esta prueba se utiliza la distribución "t" de Student.

Este tipo de prueba es de carácter bilateral, con los siguientes pasos:

1° Plantear las hipótesis:

$H_{0,i}$ = no existe relación entre la variable dependiente y la variable independiente "i".

$H_{1,i}$ = existe relación entre la variable dependiente y la variable independiente "i".

2° Especificar α %.

3° El estadígrafo de la prueba es:

$$t_i = \frac{b_i}{S_i} \quad (5.14)$$

En la que:

b_i = estimador del coeficiente de la variable independiente "i".

S_i = estimación de la desviación standard del coeficiente de la variable independiente "i"

El valor de S_i se calcula con la siguiente ecuación:

$$S_i = \sqrt{q_{ii} S_{yx}} \quad (5.15)$$

En la que:

$[X]$ = matriz de los valores observados de las variables independientes más la columna de "unos" como primera columna.

$[X']$ = matriz transpuesta de $[X]$.

q_{ii} = elemento "i" de la diagonal formada por la matriz $[X'X]^{-1}$

4° Determinar la región de aceptación:

$$\left(t_{\frac{\alpha}{2}, n-k-1}; t_{1-\frac{\alpha}{2}, n-k-1} \right)$$

5° Calcular el valor de t_i .

6° Tomar la decisión:

Si $t_i \in \left(t_{\frac{\alpha}{2}, n-k-1}; t_{1-\frac{\alpha}{2}, n-k-1} \right)$ aceptar $H_{0,i}$, caso contrario, aceptar $H_{1,i}$.

5.2.4 Prueba de hipótesis para la regresión.- La prueba de hipótesis de regresión utiliza la distribución "F" para probar si existe o no relación de todas las variables independientes como grupo con la variable dependiente. Los pasos a seguir son:

1º Plantear las hipótesis:

H₀ = no existe relación entre todas las variables independientes con la variable dependiente.

H₁ = existe relación entre todas las variables independientes y la variable dependiente.

2º Especificar α %.

3º El estadístico a utilizar es **F_c**, que relaciona el cociente entre dos varianzas (cuadrados medios), por lo cual se empleará la Distribución "F".

4º Definir la región de aceptación:

$$[0; F_{\text{tablas}}]$$

El valor de **F_{tablas}** se obtiene en tablas "F", con α %, $v_1 = k$ y $v_2 = n-k-1$ grados de libertad.

5º Se calcula el valor de **F_c** construyendo el cuadro (5.1) que es el cuadro de análisis de varianza, en el que se divide la variabilidad total en dos componentes: la variabilidad explicada (variabilidad debido a la regresión) y la variabilidad no explicada (variabilidad residual o debido al error de muestreo).

La variabilidad explicada (VE) se calcula con la ecuación (5.16)

$$VE = \sum (VD_{c,i} - \overline{VD})^2 \quad (5.16)$$

En la que:

$VD_{c,i}$ = valor calculado "i" de la variable dependiente.

\overline{VD} = media aritmética de los valores de la variable dependiente.

La variabilidad no explicada (VNE) se determina con la ecuación (5.17).

$$VNE = \sum (VD_{o,i} - VD_{c,i})^2 \quad (5.17)$$

En la que:

$VD_{o,i}$ = valor observado "i" de la variable dependiente.

La variabilidad total (VT) es:

$$VT = \sum (VD_{o,i} - \overline{VD})^2 \quad (5.18)$$

Es de hacer notar que, para cualquier caso, se cumple la siguiente identidad:

$$(VD_{o,i} - \overline{VD}) = (VD_{cc,i} - \overline{VD}) + (VD_{o,i} - VD_{c,i}) \quad (5.19)$$

CUADRO (5.1)
ANÁLISIS DE VARIANZA PARA LA PRUEBA DE REGRESIÓN

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	RATIO "F"
Regresión	VE	k	s_e^2	F_c
Error	VNE	n-k-1	s_{ne}^2	
Total	VT	n-1		

En la que:

$$s_e^2 = \frac{VE}{k} \quad (5.20)$$

$$s_{ne}^2 = \frac{VNE}{n - k - 1} \quad (5.21)$$

$$F_c = \frac{s_e^2}{s_{ne}^2} \quad (5.22)$$

6° Tomar la decisión:

Si $F_c \in [0; F_{tablas}]$, se acepta la hipótesis nula H_0 , es decir, no existe relación entre la(s) variable(s) independiente(s) y la variable dependiente, caso contrario se rechaza H_0 .

5.3 Análisis de correlación.- El principal objetivo del análisis de correlación es medir el grado de relación entre todas las variables independientes y la variable dependiente.

Para efectuar el análisis de correlación se calculan dos coeficientes: el coeficiente de determinación y el coeficiente de correlación.

5.3.1 Coeficiente de determinación.- El coeficiente de determinación mide la proporción de variabilidad que ha sido estadísticamente explicada, respecto a la variabilidad total, mediante la ecuación de regresión, es decir:

$$R = \frac{VE}{VT} = 1 - \frac{VNE}{VT} \quad (5.23)$$

Los valores que toma están siempre comprendidos en el intervalo:
 $0 \leq R \leq 1$.

De manera ideal se desea tener un valor de $R = 1$, puesto que entonces la variabilidad no explicada sería igual a cero, y que toda la variación puede explicarse por la presencia de las variables independientes en la ecuación de regresión.

5.3.2 Coeficiente de correlación.- El coeficiente de correlación indica el grado de relación que existe entre las variables independientes con la variable dependiente. Se calcula de la siguiente manera:

$$r = \pm\sqrt{R} \quad (5.24)$$

El valor de r fluctúa entre $0 \leq r \leq 1$, cuando r es igual a 1 la relación es perfecta y cuando el valor de r es igual a cero, se dice que no existe relación entre las variables consideradas.

Para el caso de un modelo lineal con una sola variable independiente, el valor r varía entre -1 y 1, siendo el signo de " r " el mismo que el del coeficiente de la variable independiente.

5.4 Diagnóstico y validación del modelo.- Con el objeto de validar el modelo encontrado se efectúa el diagnóstico de los datos a través del análisis de residuos. Dicho análisis se efectúa mediante la construcción y análisis de ciertos gráficos, los principales son:

- Gráfico: Residuos Vs. Valores calculados.
- Gráfico: Residuos Vs. Valores observados.
- Gráfico: Residuos Vs. Tiempo.

En algunos casos también se recomienda elaborar gráficos de Residuos Vs. Variable(s) independientes(s).

Para todos los gráficos elaborados, los puntos deben estar distribuidos en forma aleatoria, es decir, no deben formar ninguna curva conocida.

5.5 Análisis de autocorrelación.- El análisis de autocorrelación se realiza cuando en el gráfico: Residuos Vs. Tiempo se ha podido detectar algún tipo de relación, lo que significa, la presencia del tiempo como variable de predicción.

Por tal motivo es que se realiza la prueba de Durwin-Watson, cuyo estadígrafo de prueba es:

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2} \quad (5.25)$$

Considerando: **k** (número de variables independientes) y **n** (tamaño de la muestra) se emplean tablas "Durwin-Watson" para obtener los valores **d_L** y **d_U** con los que se efectuará el análisis respectivo.

Con los valores de **d_L** y **d_U** se elaboran los siguientes intervalos y se determina la existencia de autocorrelación, así como la dirección de ésta.

$0 < d < d_L$	Autocorrelación positiva.
$d_L < d < d_U$	Prueba no concluyente.
$d_U < d < 4 - d_U$	No existe autocorrelación.
$4 - d_U < d < 4 - d_L$	Prueba no concluyente.
$4 - d_L < d < 4$	Autocorrelación negativa.

BIBLIOGRAFÍA:

- (1) **HINES** Walter y **MONTGOMERY** David (1996): "*Probabilidad y Estadística para Ingeniería y Administración*". McGraw-Hill, México.
- (2) **LEVIN** Richard y **RUBIN** David (1996): "*Estadística para administradores*". Prentice-Hall, México.
- (3) **MILLER** Irwin, **FREUND** John y **JOHNSON** Richard (1994): "*Probabilidad y estadística para ingenieros*", México.

=====

ÍNDICE

	Página
6.1	Introducción..... 1
6.2	Análisis de regresión..... 1
6.2.1	Método de estimación de parámetros por mínimos cuadrados..... 2
6.2.2	Error standard de estimación..... 3
6.2.3	Prueba de hipótesis para coeficientes de regresión..... 3
6.2.4	Prueba de hipótesis para la regresión..... 4
6.3	Análisis de correlación..... 6
6.3.1	Coeficiente de determinación..... 6
6.3.2	Coeficiente de correlación..... 7
6.4	Diagnóstico y validación del modelo..... 7
6.5	Análisis de autocorrelación..... 7

**UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA
DEPARTAMENTO DE MATEMÁTICAS**

ESTADÍSTICA II

CAPITULO VI

***"ANÁLISIS DE REGRESIÓN Y
CORRELACIÓN LINEAL"***

SEMESTRE: I/2004

DOCENTE: Ing. Roberto Manchego C.

Cochabamba, Mayo de 2004

VI ANÁLISIS DE SERIES CRONOLÓGICAS

6.1 Introducción.- La planificación racional exige prever los sucesos del futuro que probablemente vayan a ocurrir. La previsión suele basarse en lo ocurrido en el pasado, por lo que, estamos en presencia de un nuevo tipo de inferencia estadística que se realiza acerca del futuro de alguna variable basados en sucesos pasados. Esta técnica se basa en el análisis de series cronológicas.

6.2 Serie cronológica.- Una serie cronológica o serie de tiempo es un conjunto de valores observados de cierta variable dispuestos en el orden cronológico de su ocurrencia, por lo general, registrados a intervalos igualmente espaciados.

En virtud a que una serie de tiempo es una descripción del pasado inmediato, el procedimiento más lógico para pronosticar el futuro es utilizar dichos datos históricos. Bajo el supuesto de que la historia ha de repetirse, es decir, si los datos pasados indican lo que se puede esperar en el futuro, es posible postular un modelo matemático que sea representativo del proceso.

En situaciones más reales, la forma exacta del modelo que genera la serie de tiempo no se conoce. Con frecuencia, se elige un modelo mediante la observación de los resultados de la serie de tiempo durante un periodo. Por ejemplo, en el cuadro (6.1) se muestra las ventas anuales de una compañía que comenzó a operar desde 1998.

CUADRO (6.1)
VENTAS ANUALES DE UNA
COMPAÑÍA (EN MILES DE \$)

VENTAS	AÑO
0.3	1998
0.4	1999
0.8	2000
0.9	2001
1.0	2002
1.5	2003
1.2	2004
1.0	2005
1.7	2006
2.1	2007

6.3 Análisis de series cronológicas.- Las variaciones de la serie cronológica se pueden atribuir a varios factores. Dichos factores pueden ser naturales, institucionales y socioeconómicos, algunos presentan una variación a corto plazo y otros lo hacen a largo plazo. Es así que, una serie de tiempo está conformada de variados elementos o componentes, que son los que explican los cambios observados en un período de tiempo.

El análisis de series cronológicas es el procedimiento mediante el cual se identifican y separan los factores relacionados con el tiempo que influyen sobre los valores observados de la serie. Una vez identificados estos valores, se los puede utilizar para mejorar la interpretación de los valores históricos de la serie de tiempo y para pronosticar valores futuros.

El enfoque clásico en el análisis de series de tiempo identifica cuatro factores o componentes básicos en una serie cronológica.

6.4 Componentes de las series cronológicas.- Los componentes de la serie cronológica (Y) son:

6.4.1 Tendencia secular (T).- Es el movimiento global y regular a largo plazo de los valores de la serie de tiempo durante un número prolongado de años, en el que se refleja un crecimiento, un estancamiento o una declinación de los valores de la serie. Se recomienda que en el análisis de series se utilicen cuando menos de 15 a 20 años, para no incluir como señal de tendencia los movimientos cíclicos, los cuales implican pocos años de duración.

El método de mínimos cuadrados es la base más común que se utiliza para identificar la tendencia en una serie de tiempo.

En el cuadro (6.2) se muestra, como ejemplo de tendencia, las ventas de cemento para un determinado país.

CUADRO (6.2)
VENTAS DE CEMENTO
(MILES DE BOLSAS)

VENTAS	AÑO
100	1993
110	1994
90	1995
130	1996
150	1997
180	1998
220	1999
230	2000
240	2001
310	2002
400	2003
390	2004
470	2005
500	2006
510	2007

Fuente: Elaboración propia.

Del cuadro (6.2) se observa que existe una tendencia creciente de las ventas de cemento en dicho país.

6.4.2 Variaciones cíclicas (C).- Las variaciones cíclicas se caracterizan por movimientos recurrentes ascendentes y descendentes, respecto a la tendencia, que se extienden por períodos de tiempo, por lo general, de 2 ó más años.

Si bien se han estudiado por bastante tiempo el origen de las fluctuaciones cíclicas, en general se puede decir que son de naturaleza económica y reflejan el estado de las actividades comerciales de tiempo en tiempo.

En todas las variaciones cíclicas se puede identificar la presencia de picos y depresiones. Los picos, son etapas de prosperidad, mientras que las depresiones son sinónimas de recesión económica.

En el cuadro (6.3) se muestra los datos de producción de soya en el país desde 1997 a 2007.

CUADRO (6.3)
PRODUCCIÓN DE SOYA EN BOLIVIA
(EN MILES DE TM)

PRODUCCIÓN	AÑO
10	1997
18	1998
20	1999
18	2000
16	2001
19	2002
33	2003
32	2004
28	2005
29	2006
38	2007

Fuente: Elaboración propia.

6.4.3 Variaciones estacionales (E).- Por variaciones estacionales se entienden las variaciones periódicas, que retornan con cierta regularidad dentro de un período específico de 2 años o menos.

El término "estacional" se emplea para indicar toda clase de movimiento periódico, diario, semanal o mensual, dentro un año como período de recurrencia máximo.

Los factores que generalmente originan variaciones estacionales son las condiciones climáticas, las costumbres sociales y las festividades religiosas.

Por ejemplo, en el cuadro (6.4) se detalla el consumo de energía eléctrica en Cochabamba, según las cuatro estaciones del año, para un período de tres años.

CUADRO (6.4)
CONSUMO DE ENERGÍA ELÉCTRICA EN COCHABAMBA (kW)

CONSUMO	AÑO	ESTACIÓN
28	2006	Primavera
24	2006	Verano
29	2006	Otoño
32	2006	Invierno
30	2007	Primavera
25	2007	Verano
30	2007	Otoño
34	2007	Invierno
31	2008	Primavera
26	2008	Verano
32	2008	Otoño
37	2008	Invierno

Fuente: Elaboración propia.

Las variaciones estacionales para un mejor entendimiento de éstas, vienen expresadas en proporción o en porcentaje,

6.4.4 Variaciones irregulares (I).- Las variaciones irregulares o variaciones aleatorias se deben a ciertos factores que ocurren de forma inesperada, siendo muy difícil su predicción, tales como confrontaciones bélicas y fenómenos naturales (terremotos, inundaciones, sequías), etc. Dichas variaciones son impredecibles y generalmente se las puede considerar como parte de las variaciones estacionales o cíclicas o ignorarlas por completo.

6.5 Modelos de series cronológicas.- Existen dos modelos de series cronológicas que se aceptan como buena aproximación a las verdaderas relaciones entre los componentes de los datos observados. Dichos modelos son:

6.5.1 Modelo aditivo.- En este modelo se supone que el valor de la serie compuesta es la suma de los cuatro componentes, esto es:

$$Y = T + C + E + I \quad (6.1)$$

6.5.2. Modelo multiplicativo.- En este caso se supone que el valor de la serie compuesta es el producto de los valores de los 4 componentes, es decir:

$$Y = T \times C \times E \times I \quad (6.2)$$

En esta relación, la tendencia siempre maneja valores absolutos y los demás componentes pueden estar expresados en proporción ó en porcentaje.

Generalmente, el modelo multiplicativo, por ser más conservador, se considera el modelo más adecuado para el análisis de las series de tiempo.

=====

ÍNDICE

	Pág.
7.1 Introducción.....	1
7.2 Serie cronológica.....	1
7.3 Análisis de series cronológicas.....	2
7.4 Componentes de las series cronológicas.....	2
7.4.1 Tendencia secular (T).....	2
7.4.2 Variaciones cíclicas (C).....	3
7.4.3 Variaciones estacionales (E).....	4
7.4.4 Variaciones irregulares (I).....	5
7.5 Modelos de series cronológicas.....	5
7.5.1 Modelo aditivo.....	5
7.5.2 Modelo multiplicativo.....	5
7.6 Descomposición de las series cronológicas.....	6

VII PRUEBAS NO PARAMÉTRICAS

7.1 Introducción.- Hasta ahora la mayor parte de las pruebas estadísticas (pruebas de hipótesis estadísticas, análisis de varianza, ajuste de curvas de regresión) e intervalos de confianza se basan en ciertos supuestos, por lo cual, han sido denominados métodos paramétricos.

Las pruebas paramétricas se basan en el análisis de un parámetro poblacional cuyo estimador tiene una distribución conocida (generalmente una distribución Normal) o puede aproximarse a una Distribución Normal.

Pero, en el caso de no cumplirse alguno de este supuesto es necesaria la aplicación de las denominadas *pruebas no paramétricas* ó *pruebas de distribución libre*.

Las pruebas no paramétricas se utilizan:

- Cuando se tiene duda de que las observaciones pertenecen a una Distribución Normal.
- Cuando se tienen muestras pequeñas con distribuciones desconocidas (ya no es posible aplicar el Teorema Central del Límite).
- Para probar hipótesis sobre la forma y posición de las distribuciones.

Las ventajas de las pruebas no paramétricas son:

- Se pueden aplicar a datos de tipo cuantitativo y cualitativo.
- Son rápidas y fáciles de realizar.

Las desventajas de las pruebas no paramétricas son:

- En el caso de tener la posibilidad de realizar pruebas paramétricas y no paramétricas para una determinada situación, es mejor efectuar las pruebas paramétricas por ser más precisas.
- Las pruebas no paramétricas son menos eficientes puesto que no utilizan toda la información proveniente de la muestra (lo cual implica incrementar el tamaño de la muestra).

7.2 Prueba de corridas para la aleatoriedad.- En todas las pruebas anteriores uno de los supuestos fundamentales era el hecho de la existencia de aleatoriedad en la toma de datos. En la práctica, no siempre es posible controlar la forma en la que han sido

tomados los datos, por lo que, es necesario efectuar una prueba para determinar la existencia o no de aleatoriedad en los datos.

La prueba de corridas se utiliza:

- Para analizar la existencia de aleatoriedad en los datos recolectados, considerando el orden en el que han sido obtenidos.
- Para determinar si existe alguna tendencia en los datos.

Una *corrida* se define al conjunto de observaciones similares contenidas dentro de un conjunto de observaciones diferentes.

En esta prueba se puede presentar dos casos:

- Datos cualitativos, para los cual los datos se dividen en dos categorías.
- Datos cuantitativos, para lo cual los datos se dividen en dos categorías, en función a si están por encima o por debajo de la mediana muestral.

El criterio básico radica en indicar que no es aleatoria la muestra que tiene un número muy grande o un número muy pequeño de corridas, por lo cual este tipo de prueba es de carácter bilateral.

Los pasos ha seguir son:

1º Plantear las hipótesis:

H_0 : las observaciones han sido recolectadas en forma aleatoria.

H_1 : las observaciones no han sido recolectadas en forma aleatoria.

2º Especificar α %.

3º Para efectuar la prueba se determina el estadígrafo R que es el número de corridas en la muestra.

4º Establecer la región de aceptación. Para este caso se define lo siguiente:

R = número de corridas en la muestra.

n_1 = número de elementos en la muestra del primer tipo.

n_2 = número de elementos en la muestra del segundo tipo.

Si $n_1 \leq 10$ ó $n_2 \leq 10$ se emplea la prueba C (con tablas específicas para tal efecto). Si $n_1 > 10$ y $n_2 > 10$ se puede aproximar a la Distribución Normal previa estandarización de R y por lo cual la región de aceptación es:

$$\left(-z_{1-\frac{\alpha}{2}}; +z_{1-\frac{\alpha}{2}} \right)$$

5º Calcular el valor de R empleando una muestra de tamaño n y standarizar.

$$z = \frac{(R - \mu_R)}{\sigma_R} \quad (7.1)$$

En la que:

$$\mu_R = \frac{2n_1n_2}{n_1 + n_2} + 1 \quad (7.2)$$

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}} \quad (7.3)$$

6º Tomar la decisión:

Si $z \in \left(-z_{1-\frac{\alpha}{2}}; +z_{1-\frac{\alpha}{2}} \right)$, entonces se acepta H_0 , caso contrario se rechaza H_0 .

7.3 Prueba sobre la mediana de una población.- Para probar la hipótesis respecto al valor de la mediana de una población se tiene prueba denominada prueba del rango con signo o de Wilcoxon. Esta prueba es la equivalente a la prueba paramétrica sobre el valor de la media poblacional.

Se requiere que los valores de la muestra aleatoria se encuentren por lo menos en una escala ordinal y no es necesario efectuar suposiciones respecto a la forma de la distribución de probabilidad.

Puesto que la distribución Normal es simétrica, la media aritmética de una distribución Normal es igual a la mediana, por lo cual, la prueba de Wilcoxon puede

emplearse para probar hipótesis respecto a la media aritmética de una distribución Normal.

La hipótesis nula y alternativa se plantean con respecto a la mediana poblacional y puede ser de una o de dos colas. Para esta prueba se utilizan tablas especiales, llamadas tablas de Wilcoxon.

Los pasos ha seguir (para el caso bilateral) son:

1º

$$H_0: Me = a$$

$$H_1: Me \neq a$$

2º Especificar α %.

3º Se determina el estadígrafo **W**.

4º Definir el criterio de aceptación, considerando: **n** = tamaño de la muestra y α %, se busca un valor de **W** _{α ,n} en las tablas de Wilcoxon.

5º Calcular **W**. Para ello se extrae una muestra de tamaño **n**, luego se determina la diferencia entre cada uno de los valores observados y el valor hipotético de la mediana, y esta diferencia, con signo aritmético, se designa como **d** = (**x** - **a**).

Si alguna de las diferencias es igual a cero, se elimina del análisis la observación correspondiente. Después, se ordenan los valores absolutos de la diferencia, de menor a mayor, asignando el rango 1 a la diferencia absoluta más pequeña. Cuando las diferencias absolutas son iguales, se asigna el rango promedio a los valores iguales. Finalmente, se obtiene por separado la suma de los rangos para las diferencias positivas (**W**⁺) y negativas (**W**⁻) y se calcula el valor de **W** mediante la siguiente expresión:

$$W = \text{Min} (W^+, W^-) \quad (7.4)$$

6º Tomar la decisión:

Si **W** \geq **W** _{α ,n} , entonces se acepta **H**₀, caso contrario se rechaza **H**₀.

En el caso de que **n** \geq **30**, se puede aproximar a una distribución Normal, estandarizando previamente.

7.4 Prueba sobre la diferencia de dos medianas poblacionales.- La prueba para determinar si existe diferencia entre dos medianas de dos poblaciones que tienen la misma varianza es la prueba **U** ó prueba de **Mann - Whitney**. Se requiere que los

valores de las dos muestras aleatorias independientes se encuentren por lo menos en escala ordinal.

El problema consiste en decidir si las dos poblaciones son las mismas o si una probablemente produzca observaciones mayores que la otra.

Para esta prueba se pueden presentar pruebas unilaterales y bilaterales, en este capítulo sólo se efectuará el análisis para el caso bilateral.

Los pasos ha seguir (para el caso bilateral) son:

1º

$$H_0: Me_1 = Me_2$$

$$H_1: Me_1 \neq Me_2$$

que es equivalente a indicar:

H_0 : las dos poblaciones son iguales.

H_1 : las dos poblaciones son distintas.

2º Especificar α %.

3º Se debe utilizar el estadígrafo U con la siguiente ecuación

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad (7.5)$$

En la cual:

n_1 = tamaño de la primera muestra.

n_2 = tamaño de la segunda muestra.

R_1 = suma de los rangos de la primera muestra.

4º Definir la región de aceptación. En el caso de que $n_1 < 10$ o $n_2 < 10$, existen tablas especiales de la estadística **U**. Si se tiene $n_1 \geq 10$ y $n_2 \geq 10$ la distribución muestral de **U** se aproxima a la Distribución Normal, por lo cual la región de aceptación es:

$$\left(-Z_{1-\frac{\alpha}{2}}; +Z_{1-\frac{\alpha}{2}} \right)$$

5º Calcular **U** empleando una muestra de tamaño **n**, para lo que se debe seguir el siguiente procedimiento: se combinan los datos de las dos muestras de n_1 y n_2

VII-5

observaciones, en orden ascendente, identificando los valores muestrales de acuerdo al grupo del cual provienen. Luego, se asigna el rango 1 al valor más pequeño hasta el valor de $n = n_1 + n_2$. Cuando se encuentran valores iguales, se les asigna el promedio de sus rangos. Posteriormente, se obtiene la estadística U y se estandariza:

$$z = \frac{(U - \mu_U)}{\sigma_U} \quad (7.6)$$

En la que:

$$\mu_U = \frac{n_1 n_2}{2} \quad (7.7)$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad (7.8)$$

6º Tomar la decisión:

Si $z \in \left(-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}} \right)$, entonces se acepta H_0 , caso contrario se rechaza H_0 .

7.5 Prueba sobre la diferencia de varias medianas de poblaciones.- La prueba de *Kruskal-Wallis*, llamada también prueba H , es la prueba más adecuada para probar si existe o no diferencia entre medianas de las poblaciones con idéntica distribución. Se supone que las diversas poblaciones tienen la misma forma y dispersión y se requiere que los valores de las diversas muestras aleatorias estén cuando menos en escala ordinal.

Los pasos a seguir son:

1º

$$H_0: Me_1 = Me_2 = Me_3 = \dots = Me_k$$

$$H_1: Me_1 \neq Me_2 \neq Me_3 \neq \dots \neq Me_k$$

que es equivalente a indicar:

H_0 : las poblaciones son iguales.

H_1 : las poblaciones son distintas.

2º Especificar α %.

3º Utilizar el estadígrafo H , el cual se calcula con la ecuación (7.9).

$$H = \left\{ \left[\frac{12}{n(n+1)} \right] \left[\sum_{i=1}^k \left(\frac{R_i^2}{n_i} \right) \right] \right\} - 3(n-1) \quad (7.9)$$

En la que:

n = tamaño de toda la muestra

R_i = suma de los rangos para la i-ésima muestra o tratamiento.

n_i = número de observaciones en la i-ésima muestra.

4º Definir la región de aceptación. En el caso de tener muestras pequeñas existen tablas especiales de valores críticos para la prueba H. Si el tamaño de cada una de las muestras es cuando menos 5 (**n_i ≥ 5**), la estadística **H** se distribuye en forma aproximada con la distribución Chi Cuadrado con **k-1** grados de libertad (**k** = número de tratamientos). Por tanto, la región de aceptación es:

$$(0; \chi_{\alpha, k-1}^2)$$

5º Calcular **H** empleando una muestra de tamaño **n**, con el siguiente procedimiento: se consideran las diversas muestras como un conjunto de valores y se ordenan de menor a mayor. Cuando se tienen valores iguales, se les asigna un rango promedio. Posteriormente, se asigna **R_i** la suma de los rangos ocupados por las **n_i** observaciones de la i-ésima muestra y además:

$$n = n_1 + n_2 + n_3 + \dots + n_k \quad (7.10)$$

con todo ello se calcula el estadístico **H**.

6º Tomar la decisión:

Si **H** ∈ $(0; \chi_{\alpha, k-1}^2)$, entonces se acepta **H₀**, caso contrario se rechaza **H₀**.

7.6 Pruebas de bondad de ajuste.- En las pruebas de bondad de ajuste el objetivo es comparar las frecuencias de la muestra clasificadas en categorías definidas o distribución de frecuencias observadas, teniendo el patrón esperado de frecuencias que se basan en una hipótesis nula específica o distribución de frecuencias esperadas.

La hipótesis nula en una prueba de bondad de ajuste es una afirmación sobre el patrón esperado de las frecuencias en un conjunto de categorías. El patrón esperado

puede ajustarse a la suposición de que la Distribución puede ser Uniforme, Binomial, la Poisson, la Normal, etc. o cualquier distribución empírica

7.6.1 Prueba de Pearson.- Esta prueba se utiliza con preferencia en el caso de:

- Distribuciones discretas.
- Distribuciones continuas con tamaño de muestra grande.

Para aceptar la hipótesis nula, debe ser posible atribuir las diferencias entre las frecuencias observadas y las esperadas a la variabilidad del muestreo y al nivel de significancia. Es así que el estadígrafo de prueba de Pearson se basa en la magnitud de esta diferencia para cada una de las categorías de la distribución de frecuencias. El valor de dicho estadígrafo es:

$$\pi = \sum_{i=1}^k \frac{(f_{o,i} - f_{e,i})^2}{f_{e,i}} \quad (7.11)$$

Por otro lado es posible demostrar que el estadígrafo π sigue una distribución Chi cuadrado.

En la ecuación (7.11) se observa que, si las frecuencias observadas son muy cercanas a las frecuencias esperadas, el valor calculado de π estará cercana a 0. Conforme las frecuencias observadas se alejan de las frecuencias esperadas, el valor de π se vuelve mayor. Por ello, se concluye que la prueba de Pearson implica el uso solamente del extremo superior, con el objeto de determinar si un patrón observado de frecuencias es diferente de un patrón esperado.

Los pasos a seguir son:

1º Establecer las hipótesis:

H₀: Las frecuencias están distribuidas según una distribución determinada.

H₁: Las frecuencias no están distribuidas según una distribución determinada.

2º Especificar el valor de α %.

3º Se debe utilizar el estadístico π .

$$\pi = \sum_{i=1}^k \frac{(f_{o,i} - f_{e,i})^2}{f_{e,i}}$$

En la que:

f_{oi} = frecuencia observada absoluta correspondiente a la categoría "i".

f_{ei} = frecuencia esperada absoluta correspondiente a la categoría "i".

4º Definir la región de aceptación:

$$(0; \chi^2_{\alpha, k-m-1})$$

En la que

k = número de categorías de datos

m = número de parámetros estimados a partir de la muestra.

5º Calcular el valor de π .

6º Tomar la decisión:

Si $\pi \in (0; \chi^2_{\alpha, k-m-1})$, entonces se acepta H_0 , caso contrario se rechaza H_0 .

7.6.2 Prueba de Kolmogorov-Smirnov.- La prueba de Kolmogorov-Smirnov es la prueba de bondad de ajuste que se aplica en casos en cuales se trata de datos que provienen de una distribución continua de probabilidades y el tamaño de muestra es pequeño.

Esta prueba se basa en una comparación entre las funciones de distribución acumulada que se observan en la muestra ordenada y la distribución propuesta bajo la hipótesis nula. Si esta comparación revela una diferencia suficientemente grande entre las funciones de distribución muestral y la distribución propuesta, entonces la hipótesis nula se rechaza.

En este tipo de prueba se utiliza una tabla especial llamada Tabla Kolmogorov-Smirnov.

Los pasos a seguir son:

1º Establecer las hipótesis:

H_0 : Las frecuencias están distribuidas según una distribución determinada.

H_1 : Las frecuencias no están distribuidas según una distribución determinada.

2º Especificar el valor de α %.

3º Se debe utilizar el estadístico D.

$$D = \text{Max}|FO_i - FE_i| \quad (7.12)$$

En la que:

FO_i = Frecuencia observada acumulada relativa correspondiente a la observación "i".

FE_i = Frecuencia esperada acumulada relativa correspondiente a la observación "i".

4º Definir la región de aceptación:

$$(0; D_{1-\alpha, n})$$

En la que **n** representa el número de observaciones.

5º Calcular el valor de D, ordenando previamente los datos en forma ascendente.

6º Tomar la decisión:

Si **D** \in $(0; D_{1-\alpha, n})$, entonces se acepta H_0 , caso contrario se rechaza.

=====

INDICE

	Pág.
8.1 Introducción.....	1
8.2 Pruebas de corridas sobre la aleatoriedad.....	2
8.3 Prueba sobre la mediana de una población.....	3
8.3.1 Prueba del signo.....	3
8.3.2 Prueba de Wilcoxon del rango con signo.....	5
8.4 Prueba sobre la diferencia de dos medianas poblacionales.....	6
8.5 Prueba sobre la diferencia de varias medianas de poblaciones.....	8
8.6 Pruebas de bondad de ajuste.....	9
8.6.1 Prueba de Pearson.....	10
8.6.2 Prueba de Kolmogorov-Smirnov.....	11

**UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA
DEPARTAMENTO DE MATEMÁTICAS**

ESTADÍSTICA II

CAPÍTULO VIII

"PRUEBAS NO PARAMÉTRICAS"

SEMESTRE: II/2003

DOCENTE: Ing. Roberto Manchego C.

Cochabamba, Noviembre de 2003