28

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS
PROFESIONALES "ACATLAN"

METODOS PARA CALCULAR EL INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES

T E S I N A

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A :

José Xicoténcatl Mondragón Rodríguez

ASESOR DE TESINA:

ACT. MARIA DEL CARMEN GONZALEZ VIDEGARAY

UNAM
CAMPUS ACATLÁN

DICIEMBRE 200

DEDICATORIAS

A mis padres:

Por su cariño, apoyo y comprensión incondicional.

A mis hermanos:

Jesús, Erendira, Erandi, Dulce y Donaji

Que han estado siempre conmigo.

A mis abuelos:

Alfredo y Esperanza
Fernando y Caritina

Por su afecto y cariño.

A mis sobrinos:

María Andrea, Leonardo Daniel y Uriel Benjamín

A la Universidad Nacional Autónoma de México
ENEP Acatlán:

Por darme la oportunidad de formarme como profesionista.

A mi asesor:

Act. María Del Carmen González Videgaray

Por su guía y paciencia en la realización de este trabajo.

ÍNDICE

MÉTODOS PARA CALCULAR EL INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES

CAPITULO I

LA DISTRIBUCIÓN NORMAL

CAPITULO II

ESTIMACIÓN

CAPITULO III

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES

CAPITULO IV

ANÁLISIS Y APLICACIONES

INTRODUCCIÓN

El presente trabajo tiene como objetivo analizar y aplicar los distintos métodos para obtener el intervalo de confianza para la diferencia de medias entre dos poblaciones normales realizando las inferencias respectivas para comparar los parámetros provenientes de poblaciones independientes una de la otra. El objetivo principal es analizar cada uno de los métodos para determinar cuál es el más conveniente para la obtención de un intervalo más eficiente según sea el caso. Otra razón importante para hacer este trabajo es recopilar los métodos, principalmente donde se tiene el supuesto de que la varianza de la población es diferente ya que dichos métodos se encuentran dispersos en libros de distintos autores por lo que son poco conocidos y se espera que sean de gran utilidad para el investigador que los requiera.

El trabajo se ha desarrollado comenzando por los conceptos básicos para hacer más comprensivo las partes que conforman a cada uno de los métodos.

En el capítulo I se describe a la distribución normal, normal estándar y su relación con t de student y el Teorema Límite Central.

En el capítulo II se trata el tema de inferencia estadística, se define estimación, se demuestra la obtención de los estimadores de los parámetros de la distribución normal y su uso; se da una definición de intervalo de confianza y se describe el intervalo para la distribución normal.

En el capítulo III se define el estadístico a utilizar para la diferencia de medias de dos poblaciones normales, se describe la prueba de igualdad de varianzas para ubicar en caso donde estamos, se describen los métodos considerando el caso donde las varianzas se conocen, se desconocen y cuando son iguales y son diferentes.

En el capítulo IV y último se realiza el análisis de los métodos antes mencionados mediante su aplicación a distintos ejemplos prácticos, se habla de las pruebas de hipótesis ya que van muy ligadas a los intervalos de confianza.

En el apéndice se describen las tablas de las distribuciones utilizadas. El apéndice A1 corresponde a la tabla de la distribución normal y el apéndice A2 a la tabla de la distribución t de student.

CAPITULO I

LA DISTRIBUCION NORMAL

1.1 USO DE LA DISTRIBUCIÓN NORMAL

La distribución continua de probabilidad más importante en el área estadística es la distribución normal. En 1733 Abraham De Moivre desarrolla la ecuación matemática de la curva normal, la cual tiene forma acampanada en su gráfica, también es conocida como campana de Gauss en honor a Karl Friedrich Gauss quién derivó su ecuación de un estudio de errores en mediciones repetidas de la misma cantidad obteniendo un grado de regularidad en ellos.

Muchos fenómenos que ocurren en la naturaleza, industria e investigación se aproximan a una distribución normal, por ejemplo las medidas físicas del cuerpo humano, medidas de calidad de procesos industriales, mediciones en balística, biología, física y astronomía, pagos de seguros, etc. La distribución normal es además una forma límite de muchas distribuciones de probabilidad (la binomial, la poisson, la hipergeométrica).

La distribución normal ha sido tabulada ampliamente y con precisión por lo que el investigador tendrá muchas tablas y programas de cómputo que le ahorrarán tiempo. En algunos casos los datos en estudio no se distribuyen en forma normal, pero aún cuando la población original este lejos de distribuirse de esta manera, la distribución de promedios de las muestras tiende a la normalidad bajo una variedad de condiciones. Al utilizar el Teorema Límite Central, que es uno de los teoremas de mayor importancia en probabilidad y estadística, se justifica éste hecho, lo que resulta práctico para el investigador desde el punto de vista analítico.

1.2 DEFINICIÓN DE LA DISTRIBUCION NORMAL

Sea X una variable aleatoria distribuida normalmente, donde su función densidad de probabilidad está dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

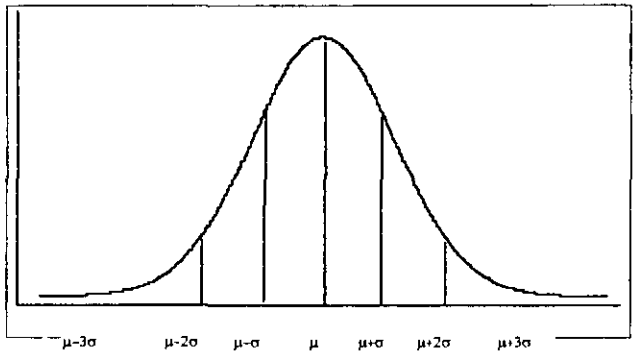
El parámetro μ es la media y σ^2 es la varianza de la población y su raíz es la desviación estándar. La función de densidad es utilizada para calcular probabilidades definidas bajo su área y sus propiedades son:

1. $P(a < X < b) = \int_a^b f(x)dx$

2. $\int_{-\infty}^{\infty} f(x)dx = 1$

3. $f(x) \geq 0 \qquad \forall \quad x \in R$

La distribución normal es perfectamente simétrica respecto a su media y el valor de la desviación estándar determina como se dispersa la gráfica de la distribución sobre el eje horizontal x como se muestra en la siguiente figura:



La función de distribución acumulada correspondiente a la variable aleatoria continua X es:

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad -\infty < x < \infty$$

Esta integral no puede ser valuada algebraicamente por lo que se tiene que hacer

integración numérica o utilizar tablas de la normal para realizar los cálculos que se necesiten de ella. La curva es asintótica con el eje x, y el área bajo la curva es igual a 1.

1.3 MEDIA Y VARIANZA POBLACIONAL

La media de una distribución de probabilidad es el valor promedio de un conjunto de datos y se le conoce como μ, también es conocida como la esperanza matemática o valor esperado de la variable aleatoria X y se expresa como E(X). En el caso de una variable aleatoria discreta se utilizan sumas y en el caso continuo se usan integrales para calcular el valor esperado

$$\mu = E(X) = \sum_{x \in S} xp(x)$$ en el caso discreto.

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$ en el caso continuo.

La varianza de la variable aleatoria X da la variabilidad o dispersión de la distribución de probabilidad, dando junto con la media una descripción adecuada de la forma de la distribución.

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_{x \in S} (x-\mu)^2 f(x)$$ para el caso discreto.

$$\sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)$$ para el caso continuo.

La cantidad x-μ es la desviación de una observación respecto de su media. La suma de las desviaciones de las observaciones es aproximadamente cero debido a que unos valores son positivos y otros negativos, para resolver ésta situación, cada desviación se eleva al cuadrado y después se promedian para así obtener la varianza que será pequeña para un conjunto de observaciones que se encuentren cercanos a la media que para uno que esté alejado de la misma. Los valores esperados de las potencias de una variable aleatoria también se conocen como los momentos de la distribución.

Ahora mostraremos que los parámetros μ y σ son la media y la varianza que son el primer y segundo momento de la distribución respectivamente. Se tiene:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} dx$$

haciendo un cambio de variable y calculando la esperanza tenemos:

$$z = (x - \mu)/\sigma \qquad \text{y} \qquad dx = \sigma dz$$

$$E(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{\frac{-z^2}{2}} dz$$

$$= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{\frac{-z^2}{2}} dz$$

El valor del área bajo la curva de la distribución es igual a uno y la segunda integral es igual a cero por lo que se tiene:

$$E(X) = \mu$$

Para la varianza tenemos que

$$E\left[(X-\mu)^2\right] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(1/2)\left[x-\mu/\sigma\right]^2} dx$$

donde

$$u = z \Rightarrow du = dz \qquad \text{y} \qquad v = e^{-z^2/2} \Rightarrow dv = z e^{-z^2/2} dz$$

entonces

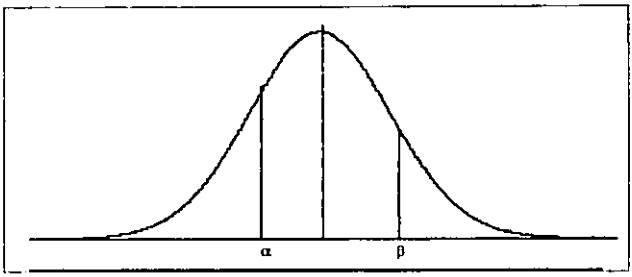
$$E\left[(X-\mu)^2\right] = \frac{1}{\sqrt{2\pi}} \left(-z e^{\frac{-z^2}{2}} \right)_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz = \sigma^2 (0+1) = \sigma^2$$

1.4 AREAS BAJO LA CURVA NORMAL

Para el área bajo la curva de la distribución normal que corresponde a la probabilidad de que X asuma un valor entre x_1 y x_2 se utiliza la función de densidad:

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

como se muestra en la región (a, b) de la siguiente figura:



Ya que la integral de la función de densidad no puede valuarse, como se ha mencionado, se hace necesario el uso de tablas normales (apéndice 1) donde se ha formado una sola tabla para cualquier curva normal. Esta tabla ha sido construida utilizando las áreas de la curva normal como una función de la variable Z que es definida como la distancia entre las observaciones y la media poblacional divididas entre su desviación estándar por lo que Z da la distancia entre una observación y la media en unidades equivalentes a la desviación estándar haciendo la transformación de las observaciones de una variable aleatoria normal X en un nuevo conjunto de observaciones de una variable aleatoria normal Z. Por lo tanto siempre que X asume un valor x, el correspondiente valor de Z será:

$$Z = \frac{X - \mu}{\sigma}$$

lo que nos lleva a deducir:

$$P(X_1 < X < X_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{X1}^{X2} e^{-\left(\frac{1}{2}\right)\left[(x-\mu)/\sigma\right]^2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{X1}^{X2} e^{\frac{-z^2}{2}} dz$$

$$= \int_{Z1}^{Z2} N(Z;0,1) = P(Z_1 < Z < Z_2)$$

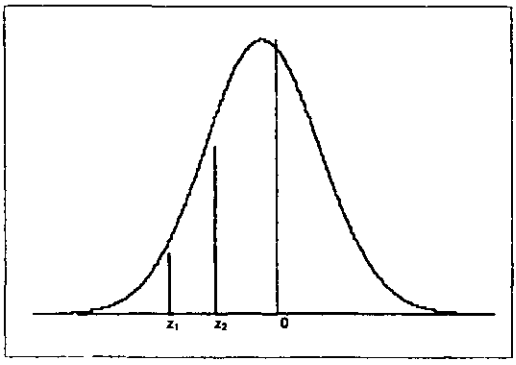
si μ y σ son constantes entonces:

$$E(Z) = \frac{1}{\sigma} E(x-\mu) = \frac{1}{\sigma}(x-\mu) = 0$$

$$Var(Z) = \left(\frac{1}{\sigma}x - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X) = 1$$

Definición: La distribución normal con $\mu = 0$ y $\sigma^2 = 1$ se conoce como distribución normal estándar por lo que $Z \sim N(0, 1)$ y puede ser descrita como el número de desviaciones estándar entre X y μ.

En la siguiente figura se ve la distribución ya transformada en una normal estándar



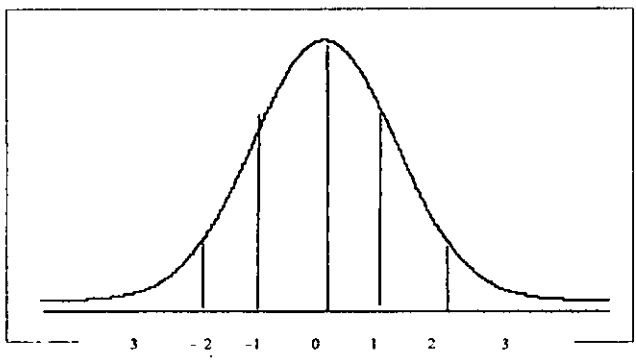
La tabla del apéndice 1 es utilizada para cualquier valor positivo z y da la superficie bajo la curva desde el origen al punto Z dando la probabilidad de que para cualquier observación tomada al azar de la distribución normal estándar se obtenga un valor que caiga entre 0 y Z.

En la siguiente gráfica se ilustra la función de densidad de la distribución normal estándar Indicando las áreas dentro de 1,2, y 3 desviaciones estándar de la media, es decir, entre $Z \pm 1$, $Z \pm 2$ y $Z \pm 3$ donde

$P(-1 \leq Z \leq 1) = 0.6827$ es la probabilidad del 68% que Z caiga en el intervalo [-1,1]

$P(-2 \leq Z \leq 2) = 0.95$ es la probabilidad del 95% de que Z caiga en el intervalo [-2,2]

$P(-3 \leq Z \leq 3) = 0.9973$ es la probabilidad del 99% de que Z caiga en [-3,3]



1.5 MEDIA Y VARIANZA MUESTRAL

Ahora discutiremos acerca de dos resultados de gran importancia en la parte de la estadística llamada muestreo para el estudio de las poblaciones normales.

Una población ha sido definida como el conjunto donde se encuentran la totalidad de las observaciones a estudiar y que son descritas mediante medidas numéricas llamadas parámetros. El estudio de está población nos lleva a la aplicación de la teoría de muestreo para obtener un subconjunto de la misma.

En la teoría del muestreo existen diversos métodos para obtener una muestra

aleatoria que se tomará de la población para realizar el estudio en cuestión pero estos no serán nombrados debido a que no son tema de discusión en este trabajo.

La muestra aleatoria es un extracto de la población, se hace esta partición debido a que resultaría demasiado costoso en dinero y tiempo realizar el estudio sobre todos los componentes de la población, así como obtener datos que no se podrían recolectar de otra forma. De esta manera, los datos obtenidos para el estudio pueden ser más desglosados y mejor analizados, por lo que es muy importante obtener muestras representativas de la población.

Definición: Muestra Aleatoria Si x_1 , x_2 , ... , x_n son variables aleatorias continuas e independientes estadísticamente, cada una con la misma distribución de probabilidad conjunta. Se define x_1 , x_2 , ... , x_n como la muestra aleatoria de tamaño n de la población $f(x)$ y se expresa su densidad de probabilidad conjunta como

$$f(\ x_1,x_2,x_3,...,x_n\ ) = f(x_1)f(x_2)..f(x_n) = \prod_{i=1}^{n} f(x_i)$$

donde la función de densidad de cada x_i es $f(x)$.

El objetivo de seleccionar variables aleatorias es el obtener información acerca de los parámetros desconocidos de la población, así como realizar afirmaciones probabilísticas.

En la sección 1.3 se presentaron los dos parámetros poblacionales μ y σ^2 de la distribución normal, estos no se ven afectados por las observaciones de una muestra aleatoria, lo que nos lleva a analizar las estadísticas de la muestra que son la media y la varianza de la muestra aleatoria.

La media es el promedio aritmético de todos los valores de la muestra y la varianza es la medida de variabilidad de la muestra.

Definición: Sean x_1 , x_2 , ... , x_n que denotan una muestra aleatoria proveniente de una distribución normal $N(\mu, \sigma^2)$, su media y varianza son:

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

y

$$S^2 = \sum \frac{(X_i - \overline{X})^2}{n}$$

De los resultados anteriores se ve que

$$E(\overline{X}) = E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n}E\sum X_i = \frac{1}{n}\sum EX_i = \frac{1}{n}\sum \mu = \frac{1}{n}\mu = \mu$$

y

$$V(\overline{X}) = E\overline{X^2} - E^2\overline{X} = E\left(\frac{\sum X_i}{n}\right)^2 - \mu^2$$

entonces

$$V(\overline{X}) = \frac{1}{n^2}(n\sigma^2 + n^2\mu^2) - \mu^2 = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

Los valores anteriores son estadísticos que se utilizan para estimar los parámetros poblacionales, la media muestral coincide con la poblacional y su varianza es igual a la varianza poblacional dividida entre el tamaño de la muestra. En el caso de la varianza se tiene otro resultado:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} = \hat{\sigma}^2$$

La ecuación anterior es la estadística obtenida por el método de estimación de máxima verosimilitud para la varianza de la muestra aleatoria, sacando su raíz obtenemos la desviación estándar de la muestra. La obtención de estas estadísticas se analizará en el siguiente capítulo. Las ecuaciones de la media y la varianza se corroboran con ayuda de la siguiente definición:

Definición: Sea x_1 , x_2 , ... , x_n variables aleatorias Independientes que tienen distribuciones normales con medias μ_1 , μ_2 , ... , μ_n y varianzas σ^2_1 , σ^2_2 , ... , σ^2_n respectivamente, entonces la variable aleatoria

$$Y = a_1x_1 + a_2x_2 + \ldots + a_n x_n$$

tiene una distribución normal con media

$$\mu = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$$

y varianza

$$\sigma^2 = a_1\sigma^2_1 + a_2\sigma^2_2 + \ldots + a_n\sigma^2_n$$

con este resultado se puede comprobar que la distribución de una suma de variables aleatorias normales independientes se distribuye como una normal[1] con parámetros μ y σ^2 en el caso de la población y como se dijo con anterioridad, en el caso del muestreo de una población con distribución desconocida ya sea finita o infinita, la distribución muestral de $\bar{X}$ será aproximadamente normal con media μ y varianza σ^2/n.

La distribución de probabilidad de las estadísticas es la distribución muestral y describe como varía una estadistica de una a otra muestra (del mismo tamaño); por lo que estas distribuciones ayudan a estudiar el comportamiento de distintas estadísticas. Este resultado es consecuencia del teorema que recibe el nombre de Teorema Límite Central y que se analizará en la siguiente sección.

Por último tenemos el siguiente resultado, $\hat{\sigma} = \frac{\sigma}{\sqrt{n}}$ que es error estándar de la media muestral, esta fórmula nos ayuda a demostrar que la desviación estándar de la distribución de $\bar{X}$ disminuye cuando n, el tamaño de la muestra, se incrementa ya que se tiene más información (más valores de variables aleatorias) con lo que se puede esperar que los valores de $\bar{X}$ se aproximen más a μ que es lo que se intenta estimar.

1.6 TEOREMA LÍMITE CENTRAL

El teorema límite central es uno de los teoremas más importantes en probabilidad y estadística pues nos indica la forma de la distribución de $\bar{X}$.

Teorema: TEOREMA LIMITE CENTRAL. si $\bar{X}$ es la media de una muestra aleatoria de tamaño n que se toma de una de una función de densidad con media μ y varianza σ^2 finita, entonces la variable aleatoria Z definida como

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

---

[1] Kendall "advanced theory of statistics" Vol. 1 y 2

se distribuye aproximadamente como una distribución normal estándar (0,1) conforme n se va incrementando, es decir $n \to \infty$.

La importancia de este teorema es que en aplicaciones prácticas la media de una muestra $\bar{X}$ procedente de cualquier distribución con varianza finita σ^2 y media μ, se distribuye aproximadamente como una normal con media μ y varianza σ^2/n. Algo interesante sobre el teorema es el hecho de que en ningún momento se habla de la función de densidad de la variable aleatoria original; sea cual sea esta distribución lo único que se pide es que se tenga una varianza finita y esto en estadística no restringe fuertemente pues la varianza siempre será necesariamente finita.

La media muestral $\bar{X}$ será aproximadamente normal para muestras grandes. Una muestra de tamaño grande depende de la forma de la función de densidad que se este trabajando. La distribución de $\bar{X}$ será exactamente una normal siempre y cuando se tenga la seguridad de que la población de donde proviene es una normal.

El Teorema Límite Central da una explicación de porque la distribución normal es una forma límite de otras distribuciones además de que se pueden calcular probabilidades aproximadas concernientes a la media de la muestra de que se encuentre en el intervalo (L_1, L_2). En la siguiente sección se analizarán de forma breve dos distribuciones de gran importancia en el muestreo de poblaciones normales.

1.7 LA DISTRIBUCIÓN t DE STUDENT Y SU RELACIÓN CON LA DISTRIBUCIÓN NORMAL

La primera distribución que se verá es la distribución χ^2 cuadrada, esta estadística es la función de los cuadrados de las observaciones de una muestra aleatoria que viene de una población normal estandarizada

$$Z = \frac{X - \mu}{\sigma}$$

la variable Z es la distribución normal estándar, tomando el cuadrado de la misma, es decir, Z^2, se llega a una distribución χ^2 cuadrada con 1 grado de libertad

$$Z^2 = \left(\frac{X - \mu}{\sigma} \right)^2$$

así mismo, si se toma la suma de las variables aleatorias independientes χ^2 cuadrada con un grado de libertad

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tienen una distribución χ^2 con n grados de libertad.

Los grados de libertad son el número de observaciones linealmente independientes o que pueden escogerse libremente. En general, dado el tamaño de la muestra n, los grados de libertad son n-k, donde k es el número de restricciones para calcular una estadística que abarque sumas de cuadrados.

La distribución χ^2 desempeña un papel importante cuando se desea hacer inferencias con respecto a la varianza σ^2 de la población, en pruebas de bondad de ajuste (comparación de una distribución hipótetica con la muestra) y en pruebas de independencia.

Al tener definidas las variables normal estándar y χ^2 cuadrada se ha llegado a un resultado muy importante en los procedimientos para hacer inferencias con respecto a la media de una población normal ya que existe otro caso en el cual la varianza resulta ser desconocida lo que es muy común en la práctica, para tal caso se utilizará a la varianza de la muestra sustituyendo a la varianza de la población. Este resultado es conocido como la distribución t de student que fue descubierto por W. S. Gosset (quien usaba el seudónimo de "student" estudiante) en 1908. Esta distribución ha revolucionado el análisis estadístico cuando se tienen muestras de tamaño pequeño, si la muestra es grande tiende a tener una distribución normal.

Definición: Sea Z una variable aleatoria normal estándar y sea χ^2 una variable aleatoria Ji cuadrada con n grados de libertad. Entonces si Z y χ^2 son independientes

$$t = \frac{Z}{\sqrt{\chi^2 / n}}$$

se dice que tiene una distribución t con n grados de libertad. Si se tiene una χ^2 con n-1 grados de libertad

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

y una normal estándar

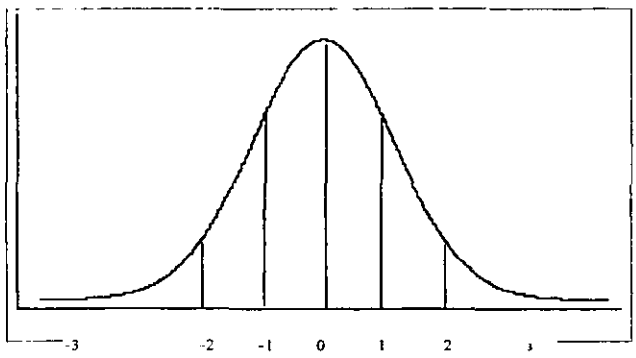
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

entonces se define una distribución t con n-1 grados de libertad como

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

Las distribuciones normal y t tienen forma de campana, pero la distribución t varia más ya que los valores de t dependen de las fluctuaciones de $\bar{X}$ y S^2 mientras que los valores de Z sólo dependen de cambios de $\bar{X}$ de muestra a muestra. La varianza de t difiere de la de Z en que la primera dependen del tamaño de la muestra n y siempre es mayor que 1; solamente cuando $n\to\infty$ las dos distribuciones llegan a ser aproximadamente iguales.

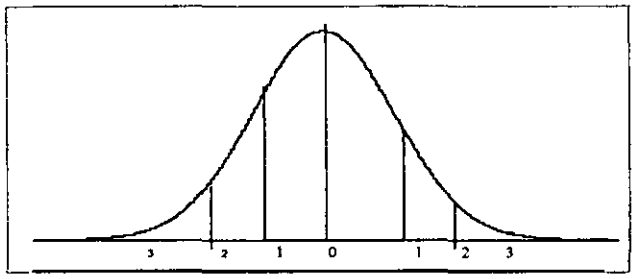
Esto se ve en las siguientes figuras



Distribución normal

En el apéndice 2 se muestra la tabla de la distribución t y contiene los valores t

arriba de los cuales se encuentra un área específica α, donde α es 0.1, 0.05, 0.025, 0.01, y 0.005 siendo estos los encabezados de las columnas y el cuerpo de la tabla los valores de t. La columna de la izquierda son los grados de libertad. Siendo simétrica t, se tiene un área derecha igual a 1- α y a la izquierda es de α igual al valor t negativo. La gráfica anterior representa a la distribución t de student con 4 grados de libertad donde se ve como se asemeja a la distribución normal cuando en tamaño de la muestra crece.



Distribución t

A lo largo de este capítulo se ha descrito a la distribución normal, su importancia en el análisis estadístico y sus componentes, mismos que utilizaremos en la descripción de los intervalos de confianza en capítulos subsecuentes.

CAPÍTULO II

ESTIMACIÓN

2.1 INFERENCIA ESTADISTICA

Uno de los objetivos de la estadística es proporcionar las herramientas necesarias para la toma de decisiones indagando acerca de alguna o algunas de las características de una población.

La inferencia estadística consiste en tomar una muestra aleatoria de la población dando una medida de incertidumbre con la cual se hace esa inferencia y se divide en dos formas: estimación y prueba de hipótesis. La palabra inferir significa sacar u obtener una consecuencia de una cosa o concluir algo de la misma por lo que inferir es generalizar de la muestra hacia toda la población.

Las inferencias buenas dependen de que se haya obtenido una muestra representativa de la población. Una buena muestra es la obtenida por métodos probabilisticos ya que así se garantiza que cada componente de la población tenga la misma probabilidad de ser escogido y se tenga una muestra aleatoria.

La estimación se puede realizar de dos maneras: por medio de estimación puntual donde se obtiene un solo número o punto que estima el parámetro de interés; la otra forma de realizar estimaciones es por el método de estimación por intervalos donde se especifica un "intervalo" de valores posibles que va a incluir el parámetro en estudio.

La estimación puntual es muy importante en la teoría estadística ya que de ella depende la toma de decisiones o llevar a cabo alguna acción respecto al problema en cuestión y consiste en determinar qué función de la muestra de las observaciones se debe usar para estimar los parámetros de la población dada, esa función es

$$h\,(X_1, X_2, \ldots, X_n)$$

que es llamada como una estadística y es usada para estimar el parámetro o parámetros de la población. En el caso de este trabajo se sabe que es una población normal (μ , σ^2) pero desconocemos qué valores de la función son los apropiados para la estimación de dichos parámetros.

A los valores numéricos que toma la estadística se les conoce como estimadores, y son una función de los valores de la muestra. Puede haber muchas estadísticas para un solo parámetro y así mismo puede haber varios estimadores para una

estadística. Por ejemplo, si tenemos un parámetro θ, si éste es conocido no tiene caso hacer la inferencia, pero si no se conoce lo deseable es que sea posible estimarlo a partir de la muestra. En la mayoría de los casos el parámetro desconocido es la media de la población

A los estimadores los denotaremos de la siguiente forma:

$$\hat{\theta} \ , \ \hat{\mu} \ , \ \hat{\sigma^2}$$

se les pone una marca llamada tilde que en estadística se le conoce comúnmente como gorro, acento chapeau o acento circunflejo.

2.2 ESTIMACIÓN DE PARÁMETROS PARA LA MEDIA Y LA VARIANZA DE UNA POBLACIÓN NORMAL

En la estimación puntual existen varios métodos que son el método de máxima verosimilitud, método de momentos, método de la ji cuadrada mínima y el método de mínimos cuadrados. El método de máxima verosimilitud es de los mejores para realizar estimaciones pues los estimadores que se obtienen poseen muchas propiedades estadísticas que los hacen muy eficientes. Lo utilizaremos para obtener los parámetros de la distribución normal y así justificar porqué la media y la varianza de la muestra son las mejores estadísticas para estimar la media y la varianza poblacional.

Lo deseable de un estimador es que las estadísticas sean iguales al parámetro a estimar. Cuando esta propiedad existe se dice que se tiene un estimador insesgado. Se tiene que

$$E\left(\hat{\mu}\right) = \mu \quad y \quad E\left(\hat{S^2}\right) = \sigma^2$$

Con lo anterior se da una idea que lo deseable es encontrar una muestra en la cual su media y su varianza, las estadísticas, fueran iguales a los parámetros a estimar al obtener el valor esperado de las mismas.

Si la propiedad anterior no se diera, se concluye que existe sesgo lo que significa que la estadística se encuentra alejada del parámetro estimado lo que provocaría obtener resultados poco confiables o fuera de la realidad que se está

investigando. El sesgo del estimador puntual es representado por la letra B definiéndola como

$$B = E\left(\hat{\mu}\right) - \mu$$

El método de máxima verosimilitud es uno de los más antiguos e importantes de la teoría de la estimación, fue usado por Gauss y después por Fisher quien lo replanteó en 1912. El método consiste en maximizar la función de verosimilitud que es la distribución conjunta de las variables de la muestra aleatoria encontrando un valor del estimador $\hat{\theta}$ donde se tenga la mayor información para el parámetro θ. Esa función es

$$L(\theta) = g(x_1, x_2, \ldots, x_n; \theta)$$

Ahora describiremos el entorno en el que se llevará a cabo la aplicación del método, para lo que se tiene:

- Sea $x \sim f_x(x;\theta)$ una variable aleatoria. discreta o continua.
- $\theta \in \Omega$
- Ω espacio de parámetros
- $\{f(x;\theta / \theta \in \Omega\}$ familia de distribuciones

Se tienen las observaciones independientes $x_1, x_2, \ldots, x_n$ de una muestra aleatoria que proviene de alguna densidad $f(x;\theta)$ su función de verosimilitud está definida por la densidad conjunta de las mismas que es una función de el parámetro θ.

$$L(\theta) = f(x_1;\theta) f(x_2;\theta) \ldots f(x_n;\theta)$$

La función de verosimilitud da la credibilidad de que las variables aleatorias asuman un valor particular $x_1, x_2, \ldots, x_n$. La verosimilitud es el valor de la función de densidad para variables continuas como lo es la normal y para variables discretas es la función masa de probabilidad. Es necesario conocer la distribución para aplicar el método. Muchas funciones de verosimilitud satisfacen las condiciones de regularidad, así que el estimador de máxima- verosimilitud es la

solución de la ecuación en el caso de un parámetro

$$\frac{dL(\theta)}{d\theta} = 0$$

y en el caso de dos o más parámetros se tiene la

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{k} f(x_i; \theta_1, \theta_2, \ldots, \theta_k)$$

Por lo que se debe dar solución al siguiente sistema de ecuaciones

$$\frac{\partial L(\theta_1,\ldots,\theta_k)}{\partial \theta_1} = 0$$

$$\frac{\partial L(\theta_1,\ldots,\theta_k)}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial L(\theta_1,\ldots,\theta_k)}{\partial \theta_k} = 0$$

Lo que se pretende encontrar es que la función de máxima-verosimilitud encuentre su máximo en el valor cero, es decir, en la distribución normal, se encuentre una recta tangente que tenga una pendiente con valor cero. Es conveniente trabajar con el logaritmo de la función de máxima-verosimilitud pues se facilita el paso anterior por se una función creciente.

$$\log L(\theta) = g(x_1, x_2, \ldots, x_n; \theta)$$

Se tiene una muestra aleatoria de tamaño n que proviene de una distribución normal con la siguiente función de densidad.

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{\left(\frac{1}{\sigma^2}\right)(x_i-\mu)^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{-1}{2\sigma^2}\sum(x_i-\mu)^2\right]$$

donde la función verosimilitud es igual a L, sacando el logaritmo

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2$$

encontrando el máximo tenemos

$$\frac{\partial \log L}{\partial \mu} = -\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2 + \frac{1}{2\sigma^2}2\Sigma(x_i-\mu) = \frac{1}{\sigma^2}\Sigma(x_i-\mu)$$

y

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + D\sigma\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}D\sigma(x_i-\mu) + \Sigma(x_i-\mu)^2 D\sigma\frac{1}{2\sigma^2}$$

$$= -\frac{n}{2\sigma^2} + \frac{\Sigma(x_i-\mu)^2}{2\sigma^4}$$

$$D\sigma\frac{1}{2\sigma^2} = \frac{2\sigma^2(0)+1}{(2\sigma^2)^2} = \frac{1}{2\sigma^4}$$

se llega al sistema de ecuaciones

$$\frac{\Sigma(x_i-\mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0$$

$$\Rightarrow \frac{\Sigma(x_i-\mu)^2 - n}{2\sigma^2} = 0$$

$$\Rightarrow \frac{\Sigma(x_i-\mu)^2}{n} = S^2$$

y tomando la otra ecuación se tiene:

$$\frac{\Sigma(x_i-\mu)}{\sigma^2}=0$$

$$\Rightarrow \frac{1}{\sigma^2}(\Sigma x_i - n\mu)=0$$

$$\Rightarrow (\Sigma x_i - n\mu) = (\sigma^2)\,0$$

$$(\Sigma x_i - n\mu)=0$$

$$\Rightarrow \sum x_i = n\hat{\mu}$$

por lo que se concluye que los estimadores de la media y varianza de la población son la media y la varianza de la muestra.

$$\hat{\mu}=\frac{\sum x_i}{n}=\overline{X}$$

Analicemos si son insesgados. Para la media μ se tiene:

$$E(\overline{X})=\mathrm{E}\frac{\Sigma \mathrm{x_i}}{\mathrm{n}}=\frac{1}{\mathrm{n}}E\,\Sigma \mathrm{x_i}=\frac{1}{n}\Sigma Ex_i=\frac{1}{n}\Sigma\mu=\frac{1}{n}n\mu=\mu$$

$\therefore$ $\overline{X}$ es in estimador insesgado para la media de la población. Para la varianza se tiene:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(\mathrm{x_i}-X)^2}{n}=E\left(\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n}\right)=\sum_{i=1}^{n}E(x^2)-nE(X^2)$$

realizando las operaciones necesarias

$$Var(\overline{X}) = E\overline{X}^2 - E^2\overline{X} = E\left(\frac{\sum x_i}{n}\right)^2 - \mu^2 = \frac{1}{n^2}\left[n(\sigma^2 + \mu^2) + n(n-1)\mu^2\right] - \mu^2 =$$

donde

$$= \frac{1}{n^2}\left[n\sigma^2 + n^2\mu^2\right] - \mu^2 = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

por lo que

$$E\left[\frac{\sum(x_i - \overline{X})^2}{n}\right] = \frac{1}{n}\left[\sum(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \frac{1}{n}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = n\sigma^2 - \sigma^2$$

llegando a

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

Por lo que se concluye que es sesgado es decir, la varianza de la muestra no es igual a la varianza de la población aunque esto es lo que se podría suponer como la elección más lógica para el estimador de la varianza. Para solucionar este problema y obtener un estimador insesgado para la varianza, se divide entre n – 1 que es el tamaño de la muestra reducido en una unidad quedando

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Si se toman muestras aleatorias de cualquier población indefinidamente grande con un valor finito de su varianza, entonces el valor promedio de $\hat{S}^2$ tomado sobre todas las muestras aleatorias será exactamente igual a σ^2 que es la varianza de la población y se le considerará un estimador insesgado. La cantidad n - 1 representa a los grados de libertad que son las piezas de información

independientes o que se pueden escoger libremente en la muestra aleatoria de la distribución. Cuando los datos (valores de la muestra) se utilizan para calcular la media muestral $\bar{X}$, hay un grado de libertad menos en la información utilizada para estimar S^2. Si se tomaran en lugar de n – 1 la función que se obtiene de las observaciones muestrales daría un resultado tendencioso de la varianza de la población ya que en el promedio, las estimaciones serían muy pequeñas. Se verá ahora que este es un estimador insesgado para la varianza de la población.
Aplicando la definición de esperanza esperanza

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

desarrollando $\sum(x_i - \bar{X})^2 = \left[\sum(x_i - \mu) - (\bar{X} - \mu)\right]^2$

$$\sum(x_i - \mu)^2 - 2(\bar{X} - \mu)\sum(x_i - \mu) + n(\bar{X} - \mu)^2 = \sum(x_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$\Rightarrow$

$$= \frac{1}{n-1}\left[\sum E(x_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left(\sum \sigma_{x_i}{}^2 - n\sigma_{\bar{X}}{}^2\right)$$

$$\sigma_{\bar{X}}{}^2 = \frac{\sigma^2}{n} \quad \Rightarrow E(\hat{S}^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2$$

por lo que se concluye

$$S^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2\right] = \frac{n}{n-1}\sigma^2$$

$$\Rightarrow \quad E(\hat{S}^2) = \left(\frac{n}{n-1}\right)E(\sigma^2) = \left(\frac{n}{n-1}\right)\left[\left(\frac{n}{n-1}\right)\sigma^2\right] = \sigma^2$$

$\therefore \; E(\hat{S}^2) = \sigma^2$

que es un estimador insesgado para la varianza poblacional.

Existen otras propiedades de los estimadores puntuales como son mínima varianza, consistencia y suficiencia. Un estimador de mínima varianza es aquel que tiene la varianza más pequeña con respecto a los otros estimadores lo que lo hace un estimador eficiente. La propiedad de consistencia nos dice que los estimadores tomarán valores muy próximos a los parámetros respectivos cuando la muestra tiende a ser grande. Un estimador suficiente es aquel donde se utiliza toda la información de una muestra relevante para la estimación del parámetro de la población, es decir, si toda la información acerca del parámetro de la población es por especificación real de los valores de la muestra, su orden puede obtenerse de la misma manera sólo observando el valor de la estadística ya que no depende del parámetro en cuestión. Estas propiedades no se analizarán en este trabajo.

2.3 DEFINICIÓN DE INTERVALO DE CONFIANZA

Un estimador puntual $\hat{\theta}$ de un parámetro θ no es muy significativo sin alguna medida del posible error en la estimación, por lo que $\hat{\theta}$ debe acompañarse alrededor de θ por un intervalo de la forma

$$L < \theta < U$$

donde

$L = \hat{\theta} - d$ límite inferior

y

$U = \hat{\theta} + d$ límite superior

L y U dependen del valor de la estadística y de su distribución muestral.

Lo que se busca al calcular un intervalo de confianza es que este contenga al parámetro que se desea estimar y además que el intervalo sea lo más corto posible. Los intervalos de confianza varían de manera aleatoria de una muestra a otra ya que depende de las funciones de las mediciones de la muestra y del tamaño de la muestra por esta razón se desea encontrar un estimador por intervalo que genere intervalos cortos con una gran probabilidad de contener a θ.

$$(L, U) = (\hat{\theta} - d, \hat{\theta} + d)$$

Esta probabilidad esta definida por la cantidad

$$P (L < \theta < U) = 1 - \alpha$$

donde $1-\alpha$ se conoce como nivel de confianza, confianza o probabilidad de confianza. Aquí α es un número pequeño positivo menor a uno, llamado nivel de significación

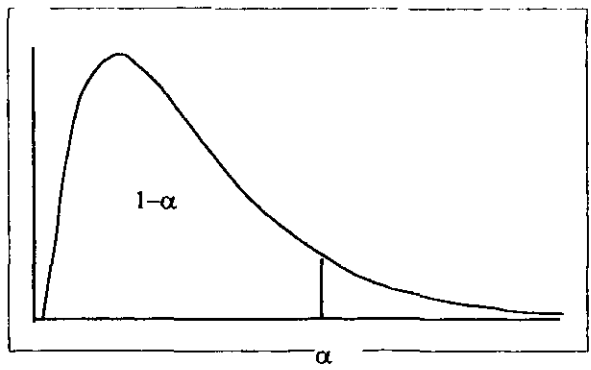
$$0 < \alpha < 1$$

(L , U) es el intervalo de confianza determinado de tal forma que al realizar muestreos repetidos de la misma población este intervalo aleatorio contendrá al parámetro θ, un cierto porcentaje de veces definido como $(1 - \alpha)$ 100%, por ejemplo, si tenemos un nivel de significación de $\alpha = 0.05$ entonces se habla de un intervalo de confianza del 95%. Los niveles de confianza que se utilizan comúnmente son 95%, 80%, 90% y 99%.

La longitud del intervalo depende del nivel de confianza escogido por lo que existe el problema de escoger $1- \alpha$ pues al aumentar la confianza se pierde precisión por lo que el intervalo se hace grande. Si se quiere disminuir el tamaño del intervalo con el mismo nivel de confianza, se debe aumentar el tamaño de la muestra.

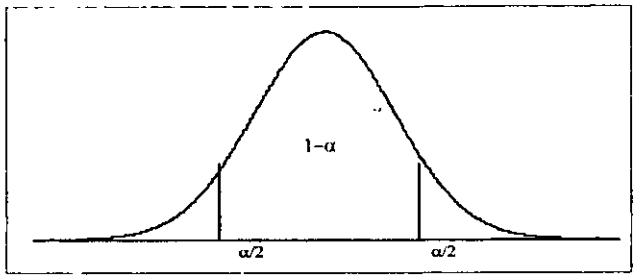
Existen dos tipo de intervalos: de dos colas y de una cola y el uso de cada uno depende del problema a abordar. Un intervalo de una cola requiere que el valor calculado sea sólo mayor o menor que el valor que se especificó con anterioridad tomando un sólo extremo de la gráfica quedando de la siguiente manera

$$P(\theta < U) = 1 - \alpha \text{ o } P(\theta > L) = 1 - \alpha$$



Un intervalo de dos colas es aquel donde se requiere que un valor calculado se encuentre dentro de un intervalo que ha sido especificado previamente quedando

$P(L< \theta <U) = 1-\alpha$



Un intervalo de confianza nos dice la precisión con la que se está trabajando; si el intervalo es de tamaño grande la información obtenida nos da poca precisión sobre el problema dado, ocurriendo lo contrario con un intervalo pequeño. Los niveles de confianza nos dan un grado de confianza de que la media en estudio se encuentra en el intervalo de confianza obtenido.

2 .3.1 INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACIÓN NORMAL

Ya han sido calculados los estimadores de los parámetros de la población normal Se tiene el estadístico Z

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

Al definir el intervalo de confianza

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

sustituyendo Z

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

multiplicando la desigualdad en cada uno de sus términos por $\sigma/\sqrt{n}$. y por –1 para invertir el sentido de las desigualdades y despejando a μ , se tiene:

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

se tiene el intervalo de confianza de (1 - α)100% para μ donde la varianza es conocida. Para el caso donde la varianza resulta desconocida se deberá utilizar la variable aleatoria.

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

que es la distribución t de Student con n – 1 grados de libertad, donde $\hat{S}$ es la desviación estándar muestral que sustituye a la desviación estándar de la población. Para obtener el intervalo de confianza se procede como en el caso anterior donde

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1 - \alpha$$

sustituyendo T,

$$P\left(-t_{\alpha/2} < \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$$

Realizando los mismos paso que en el caso anterior, multiplicando la desigualdad en cada uno de sus términos por $\hat{S}/\sqrt{n}$. y por –1 se tiene que

$$P\left(\overline{X} - t_{\alpha/2}\frac{\hat{S}}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{\hat{S}}{\sqrt{n}}\right) = 1 - \alpha$$

para así obtener un intervalo de confianza de (1 -α)100% cuando se desconoce la varianza.

Ya que el intervalo de confianza nos dice la precisión con la que se realiza la inferencia, ésta también depende de la calidad de los datos y la manera en la que han sido recolectados pues algunas mediciones contienen gran información y otras poca o ninguna con respecto del parámetro por lo que es importante que se seleccione un tamaño de muestra adecuado para el experimento que se realiza. La cantidad de observaciones que se deben incluir en la muestra depende de la exactitud que requiera el investigador. Esta exactitud se puede establecer fijando un límite de error de estimación que se define como la distancia entre el parámetro y el estimador donde

$$P\left(\left|\hat{\theta}-\theta\right| < d\right) = 1-\alpha$$

En el caso de la población normal con varianza conocida se tiene que

$$d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

donde d es la precisión que se desea; cuando la varianza es desconocida se tiene

$$d = t_{\alpha/2}\frac{\hat{S}}{\sqrt{n}}$$

La varianza es sinónimo de precisión ya que da la dispersión de la variable respecto de la media, se eleva al cuadrado para que sea positiva. Si se tiene poca varianza significa que hay mucha precisión que es lo que se busca. Dependiendo de lo que se quiera estimar se necesita cierta precisión y podemos calcular el tamaño de muestra que se requiere.

En la práctica se requiere de intervalos de confianza para medias y varianzas, de aquí se puede obtener el caso de que las muestras sean grandes o pequeñas. Para el caso de muestras grandes se puede usar indistintamente la distribución normal o la distribución t de student teniendo ésta un infinito grado de libertad debido al Teorema Límite Central. Para el caso de muestras pequeñas será necesario utilizar a la distribución t con n – 1 grados de libertad como se explicó con anterioridad. De esta forma se construirán intervalos de confianza para la diferencia de medias de dos distribuciones normales.

CAPÍTULO III

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES

3.1 DEFINICIÓN DE LAS POBLACIONES

Entre los muchos problemas en estadística es comparar las medias de dos poblaciones aplicando la inferencia estadística para saber qué diferencias existen entre las mismas y porqué causas, por ejemplo, al comparar dos tipos de drogas, dos métodos de producción, dos tratamientos médicos, dos tipos de medicina etc. realizando un análisis objetivo para llegar a una buena decisión. Estas comparaciones se realizan de la siguiente forma: el primero es cuando se tienen muestras independientes y el segundo es cuando se tienen muestras apareadas en donde se seleccionan pares de individuos u objetos similares, por ejemplo es cuando se aplica un tratamiento a un miembro de cada par y otro tratamiento al segundo miembro. La diferencia que surga entre las mediciones realizadas a los dos miembros es una estimación de la diferencia en los efectos de los dos tratamientos o procedimientos en cada uno de los casos.

Lo que nos interesa estimar es la diferencia de dos medias poblacionales $\mu_1 - \mu_2$, analizando que tan diferentes o iguales son. Se sabe que se trata de dos poblaciones normales con medias μ_1 y μ_2, y varianzas σ^2_1 y σ^2_2, se toma de ellas una muestra aleatoria de tamaño n y m respectivamente siendo independientes una de la otra. Sea

$x_1, x_2, \ldots, x_n$ una muestra aleatoria normal con media $\overline{X}$ y varianza $\hat{S}_1^2$

$y_1, y_2, \ldots, y_n$ una muestra aleatoria normal con media $\overline{Y}$ y varianza $\hat{S}_2^2$

donde

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \hat{\mu}_1 \qquad \text{y} \qquad \overline{Y} = \sum_{i=1}^{m} \frac{Y_i}{m} = \hat{\mu}_2 ,$$

y

$$\hat{S}_1^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1} = \hat{\sigma}_1^2 \quad \text{y} \quad \hat{S}_2^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \overline{Y}\right)^2}{m-1} = \hat{\sigma}_2^2$$

Definidas las muestras, por el teorema límite central se sabe que la diferencia de medias $\mu_1 - \mu_2$ se distribuye normalmente. La suma de las desviaciones $X_i - \mu$ de la media es igual a cero, esto se debe a que unas desviaciones son positivas y otras negativas por lo que no sirven como medida de dispersión. Para solucionar este problema se omite el signo de estas desviaciones considerando su valor absoluto.

$$\sigma^2_{\bar{X}-\bar{Y}} = \text{Prom}\left[(X-Y) - (\mu_1 - \mu_2)\right]^2$$

donde

$$(X-Y) - (\mu_1 - \mu_2) = (X - \mu_1) - (Y - \mu_2)$$

elevando al cuadrado

$$\left[\ (X-Y) - (\mu_1 - \mu_2)\ \right]^2 = \left[\ (X-\mu_1) - (Y-\mu_2)\ \right]^2 = (X-\mu_1) + (Y-\mu_2) - 2(X-\mu_1)(Y-\mu_2)$$

puesto que

$$\text{Prom}(X-\mu_1)^2 = \sigma^2_{\bar{X}} \qquad \text{y} \qquad \text{Prom}(Y-\mu_2)^2 = \sigma^2_{\bar{Y}}$$

por lo que

$$\text{Prom}(X-\mu_1)(Y-\mu_2) = (X_1 - \mu_1)\,\text{Prom}(X_2 - \mu_2) = 0$$

llegando a

$$\sigma^2_{\bar{X}-\bar{Y}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} - 2\,\text{Prom}(X-\mu_1)(Y-\mu_2)$$

La varianza es el promedio ponderado del cuadrado de las desviaciones de los valores en la población de la media de la población. Se define la varianza de una diferencia de medias como

$$\sigma^2_{\bar{X}\ \bar{Y}} = \sigma^2_X + \sigma^2_Y = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}$$

3.2 PRUEBA DE IGUALDAD DE DOS VARIANZAS

Antes de resolver algún problema lo primero que debe hacerse es comparar las varianzas de las poblaciones en estudio de esta forma sabremos en que caso nos encontramos para así aplicar el método correspondiente.

Para comparar las varianzas se aplica una prueba conocida como la prueba F, desarrollada por Ronald Alymer Fisher. El objetivo de esta prueba es averiguar si las varianzas son iguales o diferentes, si una de ellas es menos del doble de la otra se concluye que las varianzas de las poblaciones son iguales. Las desviaciones ligeras de la suposición de que las varianzas son iguales no altera el grado de confianza del intervalo.

Para realizar esta comparación se deben colocar a las varianzas en forma de razón de la siguiente manera

$$F_c = \frac{\sigma_1^2}{\sigma_2^2} \quad \text{si} \quad \sigma_1^2 \geq \sigma_2^2 \qquad \text{y} \qquad F_c = \frac{\sigma_2^2}{\sigma_1^2} \quad \text{si} \quad \sigma_1^2 < \sigma_2^2$$

cuando las varianzas se conocen y

$$F_c = \frac{\hat{S}_1^2}{\hat{S}_2^2} \quad \text{si} \quad \hat{S}_1^2 \geq \hat{S}_2^2 \qquad \text{y} \qquad F_c = \frac{\hat{S}_2^2}{\hat{S}_1^2} \quad \text{si} \quad \hat{S}_1^2 \leq \hat{S}_2^2$$

cuando las varianzas se desconocen.

3.3 CASO DE VARIANZAS IGUALES

Para el caso donde se sabe que las varianzas son iguales y se conocen se tiene al estadístico

$$Z_c = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

siendo Z_c la variable estandarizada, se le pone una c como subíndice que nos da la referencia de que es el estadístico a calcular. La gama minúscula γ

representará en lo sucesivo al número de grados de libertad.

$$P\left(-z_{\frac{\alpha}{2}} < Z_c < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

sustituyendo Z_c

$$P\left[-Z_{\frac{\alpha}{2}} < \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2_1}{n}+\sqrt{\frac{\sigma^2_1}{m}}}} < Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

despejando $\mu_1 - \mu_2$ obtenemos

$$P\left[\left(\overline{X}-\overline{Y}\right)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_1}{n}+\frac{\sigma^2_2}{m}} < \mu_1-\mu_2 < \left(\overline{X}-\overline{Y}\right)+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_1}{n}+\frac{\sigma^2_2}{m}}\ \right] = 1-\alpha$$

que es el intervalo de confianza para (1-α)100%.

Cuando las varianzas de las poblaciones resultan ser iguales pero desconocidas, se calcula el promedio de las varianzas de las muestras para obtener el estimador puntual de la varianza poblacional y utilizar la distribución t. Lo que se realiza a continuación es sustituir a las varianzas poblacionales por las varianzas de cada muestra y utilizando las variables ji cuadrada con n - 1 y m - 1 grados de libertad se tiene que

$$V = \frac{(n-1)S_1^2}{\sigma^2}+\frac{(m-1)S^2_2}{\sigma^2} = \frac{(n-1)S_1^2+(m-1)S_2^2}{\sigma^2}$$

que tiene una distribución ji cuadrada con n + m - 2 grados de libertad, Combinando la variable aleatoria Z_c y la ji cuadrada anterior se tiene que

$$T_c = \frac{Z_c}{\sqrt{\frac{V}{n+m-2}}}$$

es el estadístico que se distribuye como una t de student con n + m - 2 grados de libertad.

$$\hat{S}_p^2 = \frac{\sum X_i^2 - \left(\sum X_i\right)^2 / n + \sum Y_i - \left(\sum Y_i\right)^2 / m}{n+m-2}$$

llegando a

$$\hat{S}_p^2 = \frac{(n-1)\hat{S}_1^2 + (m-1)\hat{S}_2^2}{n+m-2}$$

que es la estimación combinada del cuadrado medio de la varianza común para la diferencia de medias, que es insesgado y da el mismo peso a toda la información. Para obtener el intervalo de confianza se sustituye la varianza poblacional por la muestral y se llega a

$$P\left[\left(\overline{X}-\overline{Y}\right) - t_{\frac{\alpha}{2},\gamma}\hat{S}_p\sqrt{\frac{1}{n}+\frac{1}{m}} < \mu_1 - \mu_2 < \left(\overline{X}-\overline{Y}\right) + t_{\frac{\alpha}{2},\gamma}\hat{S}_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right] = 1-\alpha$$

que es el intervalo de confianza de (1-α)100%.

3.4 CASO DE VARIANZAS DISTINTAS

En la sección anterior se tiene la suposición de que las varianzas de las poblaciónes son iguales lo que resulta no ser siempre cierto ya que existen varias razones para refutar esta suposición. Algunas de las causas son que las muestras de las poblaciones provienen de poblaciones diferentes, aunque las dos sean normales, a que las medias de las poblaciones sean demasiado diferentes derivando en un cambio en la varianza de las mismas. Para el caso donde las varianzas son distintas varios autores han desarrollado diversas propuestas para el mismo utilizando el siguiente estadístico

$$T_c = \frac{\left(\overline{X}-\overline{Y}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\hat{S}^2_1}{n} + \frac{\hat{S}^2_2}{m}}}$$

donde se han sustituido a las varianzas de la población por las varianzas de las muestras. El problema de este estadístico es que no se distribuye como una t de student con n + m - 2 grados de libertad, para resolver ese problema se utilizan los siguientes métodos.

3.4.1 MÉTODO DE COCHRAN-COX

Este método utiliza un promedio que se obtiene al dividir cada una de las varianzas entre el tamaño de su respectiva muestra y ponderar cada uno de los valores de la tabla t , esto lleva a una disminución en los grados de libertad por lo que el valor de T' es alto lo cual produce que el intervalo aumente su tamaño y en consecuencia se pierda precisión al estimar el intervalo. El intervalo de confianza es

$$\overline{X} - \overline{Y} \pm T' \sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}$$

donde $T' = \frac{w_1 T_{\alpha 1} + w_2 T_{\alpha 2}}{w_1 + w_2}$, $w_2 = \frac{\hat{S}_2^2}{m}$, $w_1 = \frac{\hat{S}_1^2}{n}$ y $T_{\alpha 1}$, $T_{\alpha 2}$ para $\gamma = n - 1$ grados de libertad. El método se ha vuelto bastante popular en su utilización.

3.4.2 MÉTODO DE SMITH - SATTERTHWAITE

En el método sugerido por Smith y expandido por Satterwaite para obtener los grados de libertad las varianzas son divididas entre el tamaño de su respectiva muestra se suman y se elevan al cuadrado, el resultado se divide entre cada una de sus varianzas divididas a su vez entre el tamaño de la muestra cada una elevada al cuadrado y dividida entre el tamaño de la muestra al que se le resta una unidad. Esto conduce a que aumente el valor en el número de grados de libertad y se tenga un intervalo de confianza más pequeño. El intervalo se define de la siguiente forma

$$\overline{X} - \overline{Y} \pm T_{\frac{\alpha}{2},\gamma} \sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}$$

y sus grados de libertad son

$$\gamma = \frac{\left(\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{\hat{S}_2^2}{m}\right)}{m-1}}$$

3.4.3 MÉTODO DE DIXON-MASSEY

Si las suposiciones de normalidad se cumplen al utilizar el estadístico antes mencionado éste tiene una distribución t con γ grados de libertad que en este método se obtienen de forma similar que en el método anterior con la diferencia de que en éste las varianzas divididas entre el tamaño y elevadas cada una al cuadrado son a su vez divididas entre n+1 y m+1 y restándole al valor total un dos que es el número de muestras involucradas en el estudio. El efecto resultante es tener un número mayor de grados de libertad lo que consecuentemente disminuye el tamaño del intervalo.

$$\overline{X} - \overline{Y} \pm T_{\frac{\alpha}{2},\gamma} \sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}$$

donde los grados de libertad están dados por

$$\gamma = \frac{\left(\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n}\right)^2}{n+1} + \frac{\left(\frac{\hat{S}_2^2}{m}\right)}{m+1}} - 2$$

El valor de γ no será un número entero, por lo que si se requiere de mayor precisión se puede realizar una interpolación en la tabla de la distribución t pero el valor más próximo puede ser suficiente.

3.4.4 MÉTODO DE PEÑA SÁNCHEZ DE RIVERA

Para el intervalo de confianza de (1 - α)100% se tiene que

$$\overline{X} - \overline{Y} \pm T_{\frac{\alpha}{2},\gamma} \sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}$$

donde los grados de libertad de la distribución t son $\gamma = n + m - 2 - \Delta$. Se define Δ un número positivo que es utilizado como corrector que se calcula tomando el entero más próximo de la siguiente fórmula

$$\Delta = \frac{\left[(m-1)s_1 - (n-1)s_2\right]^2}{(m-1)s_1^2 + (n-1)s_2^2}$$

siendo

$$s_1 = \frac{\hat{S}_1^2}{n} \quad \text{y} \quad s_2 = \frac{\hat{S}_2^2}{m}$$

Hay que tener cuidado de no confundir s con la $\hat{S}$ de la varianza muestral. Se demuestra que Δ se encuentra entre los valores

$$0 \leq \Delta \leq \max(n-1, m-1)$$

El término corrector se interpreta como sigue: si la varianza de la primera población es mucho mayor que el de la segunda y los tamaños de las muestras son iguales, el valor de s en la primera población será mucho mayor que en la segunda por lo Δ aumentará y los grados de libertad serán menores por lo que éstos dependerán de la precisión con la que se estime la varianza de la primera población. Si las varianzas de las poblaciones son similares así como los tamaños de las muestras, el término corrector se anula quedando el caso de igualdad de varianzas. Si el tamaño de las muestras es muy distinto, el término corrector será alto y los grados de libertad se reducirán.

3.5 MÉTODO DE LAS OBSERVACIONES APAREADAS

Cuando se tiene suficiente evidencia de que las varianzas de las poblaciones no son iguales y hay una gran diferencia entre las poblaciones, existe un método que puede ayudar a disminuir o eliminar las variaciones que se han presentado debido a diferentes factores ajenos al de interés. Este método es conocido como el de observaciones apareadas y consiste en tomar a las observaciones de cada una de las muestras para formar pares de individuos u objetos que tengan características similares (como el peso, la estatura, el tamaño, la capacidad intelectual etc.) exceptuando el que se trata de medir o analizar que es la diferencia de medias.

El procedimiento para formar las parejas es poner a los sujetos a prueba para obtener indicios sobre sus semejanzas y así tener pares que estén positivamente correlacionados. Esta correlación hace que las diferencias entre las dos mitades tengan una tendencia a ser pequeñas provocando que la varianza de la nueva muestra formada al calcular las diferencias de las observaciones de las muestras originales se reduzca considerablemente debido a que existe similitud en la naturaleza de los errores de las observaciones. La elección de los pares es limitada a la disponibilidad de individuos, a las similitudes existentes y a la habilidad del investigador para formarlos.

Lo anterior lleva a tener una baja probabilidad de ocultar alguna diferencia que haya sido impuesta debido a una diferencia en los tratamientos aplicados y de esta forma aumentar la precisión en la inferencia al comparar tratamientos o procedimientos.

La ventaja del método de observaciones apareadas es que las varianzas pueden ser iguales o diferentes, los valores de las observaciones pueden ser o no independientes, se reduce el problema al de una muestra y se eliminan efectos extraños que pueden provocar diferencias significativas en las medias de las poblaciones. Lo único que se necesita es que el tamaño de las muestras sea igual para formar los pares y que se puedan aparear, donde $D_i = X_i - Y_i$ representa la diferencia entre las observaciones que forman una nueva muestra aleatoria obteniendo pares que por el Teorema Límite Central se distribuyen de forma normal. El estadístico a calcular es

$$T_C = \frac{\overline{D} - \mu_D}{\hat{S}_D / \sqrt{n}}$$

que se distribuye como una t de student con $\gamma = n - 1$ grados de libertad y sus componentes se definen como

CAPITULO IV

ANÁLISIS Y APLICACIONES

4.1 PRUEBAS DE HIPÓTESIS

El uso de la Estadística se hace presente en muchos aspectos de la vida cotidiana. Acontecimientos simples como el llevar estadísticas sobre la temporada en algún deporte, registros de los cambios de clima, obtener promedios sobre las calificaciones de un alumno etc. La Estadística cobra gran importancia cuando se aplica a la investigación de una manera formal.

Uno de los problemas que trata la Inferencia Estadística y que va muy ligado al problema de estimación de intervalos de confianza es la Prueba de Hipótesis que es la formulación de un procedimiento de decisión basado en los datos obtenidos para dar como resultado una conclusión acerca de algún sistema, método o procedimiento científico, industrial, a la eficacia de algún aparato o producto con respecto de otro etc.

Las pruebas de hipótesis permiten verificar la veracidad o falsedad de alguna hipótesis establecida acerca de una población para determinar si los valores difieren significativamente de los esperados por la hipótesis y con respecto de ella tomar una decisión que puede ser rechazada o no aportando suficiente evidencia para ello.

Una hipótesis estadística es una afirmación respecto a un parámetro o parámetros de una distribución de probabilidad. Si la hipótesis estadística especifica completamente los valores del o de los parámetros se le llama hipótesis simple que es una prueba de dos colas, por tanto un intervalo; si los valores no se especifican completamente se le llama hipótesis compuesta que es una prueba de una cola. Para el parámetro μ

$\mu = 5$ hipótesis simple

$\mu > 5$ o < 5 hipótesis compuesta

Una prueba de hipótesis se compone de dos tipos de hipótesis, una denominada hipótesis nula denotada por H_0 que se establece con el único propósito de rechazarla y que se especifica en forma opuesta a la que se supone cierta y otra denotada por H_A llamada hipótesis alternativa que es cualquier suposición que difiera de la nula. La hipótesis nula y alternativa deben ser mutuamente

exclusivas. La Prueba de una cola se define cuando

Cola derecha	Cola izquierda
H_0: $\mu \leq \mu_0$	H_0: $\mu \geq \mu_0$
vs	vs
H_A: $\mu > \mu_0$	H_A: $\mu < \mu_0$

La prueba de dos colas esta definida como

$$H_0 : \mu = \mu_0$$

vs

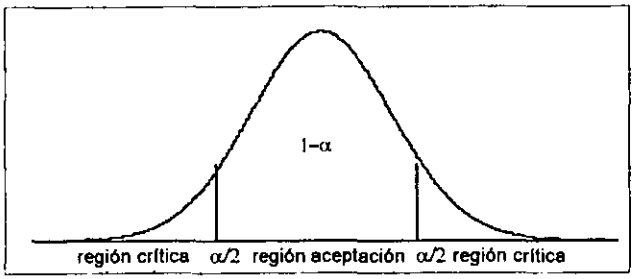
$$H_A : \mu \neq \mu_0$$

De esta forma se habla de probar una hipótesis nula contra una alternativa en el supuesto de que la hipótesis nula es verdadera. Este procedimiento conducirá a dos tipos de errores posibles debido a las fluctuaciones al azar en la obtención de la muestra de la población; el error tipo I que se presenta cuando la hipótesis nula se rechaza siendo ésta verdadera y el error tipo II que se presenta cuando no se rechaza la hipótesis nula siendo falsa.

La probabilidad de cometer alguno de los errores anteriores se consideran como riesgos de decisiones incorrectas. La probabilidad de cometer un error tipo I se designa como α que es el nivel de significación anteriormente definido y la probabilidad de cometer un error tipo II se designará como β. Cuando α disminuye, β aumenta lo cual es considerado como más grave.

En las pruebas de hipótesis se requiere de un intervalo de confianza para satisfacer el objetivo de la inferencia estadística. Con este intervalo se garantiza que la diferencia de medias caerá en el mismo. El valor de la estadística calculada debe caer en la región de aceptación para no rechazar a la hipótesis nula en el caso de que sea verdadera ya que esta contiene los resultados que más favorecen a H_0 . Si H_0 cae en la región de rechazo o región crítica donde se encuentran los resultados que la favorecen menos, la hipótesis nula será rechazada.

En el siguiente gráfico se ve claramente la definición anterior



El procedimiento para llevar a cabo una Prueba de Hipótesis es:

1) Obtener los datos y suposiciones.

2) Formular una hipótesis acerca del o de los parámetros en estudio (H_0, H_A).

3) Escoger el nivel de significación o de riesgo α.

4) Escoger la estadística de las observaciones que se va a calcular.

5) Determinar la región crítica que depende de la estadística que se calcula y del nivel de significación, así como definir el intervalo de confianza que será la región de no rechazo o aceptación.

6) Al obtener una aceptación o rechazo de la hipótesis nula, dar una conclusión al problema planteado.

Las pruebas de hipótesis que se realizarán en este capítulo, son las concernientes a la diferencia de medias entre dos poblaciones normales por lo que la prueba queda definida como

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0$$

4.2 CASO DE VARIANZAS IGUALES

Veamos algunos ejemplos donde la varianza se conoce y se desconoce.

4.2.1 UN ESTUDIO SOBRE CALIFICACIONES

En un estudio realizado a 11 niños y 10 niñas con cierto retraso mental, después de un año de educación especial combinado con terapia, se les aplicó un exámen de conocimientos. La media para los niños fue de 67.0 y para las niñas de 61.5.

Se considera la suposición de que las calificaciones obtenidas por niños y niñas bajo estas circunstancias están normalmente distribuidas con una desviación estándar de 11 y 10 respectivamente, ¿ Es la diferencia de las medias significativo para un nivel de confianza del 90%?

H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$

$Z_c = 1.2001$ donde $Z_{\alpha/2} = \pm 1.645$

El intervalo de confianza para $\mu_1 - \mu_2$ es

$$P(-2.0383 < \mu_1 - \mu_2 < 13.0383) = 0.90$$

Por lo tanto se concluye que no existe diferencia significativa en las medias de las poblaciones y se acepta la hipótesis nula.

4.2.2 UNA APLICACIÓN A LA ASTRONOMÍA

Dos astrónomos anotaron observaciones de una cierta estrella. Las 12 observaciones obtenidas por el primer astrónomo tienen una media de 1.20 y las obtenidas por el segundo tienen una media de 1.15. La experiencia anterior indica que los astrónomos obtuvieron lecturas con una varianza de .40. ¿ Parece razonable la diferencia entre los dos resultados ?

H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$, con $\sigma_1 = \sigma_2 = \sqrt{0.40} = 0.635$

Se elige $\alpha = 0.01$, donde $Z_c = 0.17$ ya que se distribuye como una normal (0,1).

$\therefore$ El intervalo de confianza es $(-2.58 < \mu_1 - \mu_2 < 2.58)$

La hipótesis nula es aceptada ya que que no hay diferencia entre las medias.

4.2.3 UN ESTUDO SOBRE ALIMENTOS

Se usan dos clases de comida para cerdos y se desea comparar cuál de éstas es mejor. Una muestra de 12 cerdos se alimenta de una ración del alimento A y la otra muestra, también de 12 cerdos, con el alimento tipo B. Las ganancias en peso de los cerdos son las siguientes:

Tipo A	31	34	29	26	32	35	38	34	30	29	32	31
Tipo B	26	24	28	29	30	29	32	26	31	29	32	28

H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$ con $\alpha = 0.05$.

Ya que se desconoce la varianza se usa la prueba t donde $\overline{X} = 31700$ y $\overline{Y} = 28$, $\hat{S}^2_p = 8.131 \Rightarrow \hat{S}_p = 2.85$ para $T_{\alpha/2,22} = \pm 2.07 \Rightarrow$ Tc = 2.65. Por lo que rechazamos la hipótesis nula ya que el valor calculado no cae en el intervalo por lo que concluimos que el alimento tipo A es mejor. Su intervalo de confianza es

(.6748,5.4917)

4.2.4 EFICIENCIA DE ENTRENADORES EN VENTAS

En una compañía de seguros se tienen dos entrenadores de agentes de seguros, se espera el mismo resultado con el primer entrenador con un grupo de 5 agentes, y el segundo con un grupo de 7 agentes. A los agentes entrenados se les lleva un registro individual de las pólizas que han vendido. ¿ Se deben seguir empleando a los 2 entrenadores ¿

Grupo 1	Grupo 2
12	6
14	10
18	9
20	15
9	16
	5
	10

Se debe probar la siguiente hipótesis nula

H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$

$\overline{X} = 14.16$ y $\overline{Y} = 10.14 \Rightarrow \hat{S}^2_X = 19.8$ y $\hat{S}^2_Y = 17.14$

$T_c = 1.7842$ y $T_{\frac{\alpha}{2},\gamma=10} = \pm 1.812$

$\hat{S}^2_p = 18.2057 \Rightarrow \hat{S}^2_{X-Y} = 6.2419 \therefore \hat{S}_{X-Y} = 2.4928$

y el intervalo de confianza es(-15.2885,23.3285)

$\therefore$ No se rechaza la hipótesis nula ya que cae dentro del intervalo del 95% de confianza, por lo que se debe seguir empleando a los 2 entrenadores.

4.2.5 UN EXPERIMENTO SOBRE UN MEDICAMENTO

Una compañía está interesada sobre el lapso de tiempo en que un medicamento retiene su potencia. Una muestra aleatoria de 10 botes del medicamento fue tomada de la línea de producción y analizada para obtener su potencia media, una segunda muestra de 10 botes fue obtenida del estante de una farmacia donde ya tenían un año de haber sido colocadas, se analizaron y se obtuvo su potencia media también. Construya un intervalo de confianza del 95% para la diferencia en la potencia media de las muestras. Los datos se describen en la siguiente tabla

Muestra 1	Muestra 2
10.2	9.8
10.5	9.6
10.3	10.1
10.8	10.2
9.8	10.1
10.6	9.7
10.7	9.5
10.2	9.6
10.0	9.8
10.6	9.9

Los resultados obtenidos son $\overline{X} = 10.37$ y $\overline{Y} = 9.83 \Rightarrow \hat{S}^2_X = .105$ y $\hat{S}^2_Y = .058$ $\hat{S}^2_p = .81225 \Rightarrow \hat{S}_p = .285$ y el intervalo de confianza del 95% es (.54$\pm$ .268) por lo que se estima que la diferencia de medias cae en (.272, .808), lo que indica que la potencia media del medicamento de la línea de producción es más elevado.

4.3 CASO DE VARIANZAS DISTINTAS

Para el caso donde las varianzas de las poblaciones son distintas se tienen los siguientes ejemplos a los cuales se les aplicarán los distintos métodos antes descritos.

4.3.1 LA DIFERENCIA ENTRE DOS PEGAMENTOS

Para probar la efectividad de dos cementos dentales para pegar coronas aisladas, se usan 41 moldes de dientes diferentes con cada uno. El valor medio de la fuerza necesaria para quitar cada corona cementada para el cemento 1 fue de 45 pies/libra con una desviación de 6.2 pies/libra y para el cemento dos fue de 42 pies/libra con una desviación de 4.3 pies/libra. Probar si existe diferencia en la fuerza media para quitar las coronas.

Se desea probar H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$

donde

$$\sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}} = 1.1783$$

a) Con el Método de Dixon- Massey escogiendo $\alpha = .05$, se tiene que $T_c = 2.5459$, $T_{\alpha/2,\gamma} = 1.9932$, que se obtiene interpolando los valores de la tabla, donde los grados de libertad son $\gamma = 72.8129$, el intervalo de confianza es

$$.6512 < \mu_1 - \mu_2 < 5.3487$$

b) Aplicando Cochran- Cox donde $w_1 = .9375$, $w_2 = .4509$, $T'_{\alpha/2} = 2.0211$ ya que los tamaños de las muestras son iguales. $T_c = 2.5459$ y el intervalo de confianza es

$$.6184 < \mu_1 - \mu_2 < 5.3815$$

c) Aplicando Smith-Satterthwaite se obtiene $\gamma = 71.2505$, interpolando se tiene $T_{\alpha/2,\gamma} = 1.9939$, por lo que el intervalo de confianza es

$$.6504 < \mu_1 - \mu_2 < 5.3495$$

d) Por el método de Peña Sánchez el término corrector $\Delta = 8.7496$, por lo que el valor de $\gamma = 71.2503$ y $T_{\alpha/2,\gamma} = 1.9939$; por lo que el intervalo que resulta es

$$.6504 < \mu_1 - \mu_2 < 5.3495$$

por lo que se rechaza la hipótesis nula. Las fuerzas medias necesarias para quitar las coronas son significativamente diferentes.

4.3.2 MEDICIONES EN UN RÍO

El departamento de zoología de la UNAM dirigió un estudio para estimar la diferencia en la cantidad de ortofósforo químico medido en dos estaciones diferentes del río Bravo. Se sacaron 15 muestras de la estación 1 con una media de 3.84 miligramos por litro y una desviación estándar de 3.07 mientras en la segunda estación se sacaron 12 muestras con una media de 1.49 y una desviación estándar de 0.80 miligramos/litro. Encontrar el intervalo de confianza del 95% para la diferencia media en los contenidos promedios reales de ortofósforo en las estaciones.

Se tiene

$$\sqrt{\frac{\hat{S}_1^2}{n}+\frac{\hat{S}_2^2}{m}} = .8256$$

a) Aplicando el Método de Dixon-Massey, con $\alpha = .10$ se tiene que $T_c = 2.8463$, $T_{\alpha/2,\gamma} = 1.7417$, que se obtiene interpolando los valores de la tabla, donde los grados de libertad son $\gamma = 16.666$, el intervalo de confianza es

$$.9119 < \mu_1 - \mu_2 < 3.7880$$

b) Con Cochran- Cox donde $w_1 = .6283$, $w_2 = .0533$, ${}_1T_{\alpha/2,\gamma} = 1.7613$ ${}_2T_{\alpha/2,\gamma} = 1.7939$ se tiene $T' = 1.7638$. $T_c = 2.8463$ y el intervalo de confianza es

$$.8937 < \mu_1 - \mu_2 < 3.8062$$

c) Aplicando Smith-Satterthwaite se obtiene $\gamma = 16.3279$, interpolando se tiene $T_{\alpha/2,\gamma} = 1.7438$, por lo que el intervalo de confianza es

$$.9102 < \mu_1 - \mu_2 < 3.7897$$

d) Por el método de Peña Sánchez el término corrector $\Delta = 8.6737$, por lo que el valor de $\gamma = 16.3278$ y $T_{\alpha/2,\gamma} = 1.7438$; por lo que el intervalo que resulta es

$$.9102 < \mu_1 - \mu_2 < 3.7897$$

por lo que la media de ortofósforo es mayor en la estación 1.

4.3.3 LA EFECTIVIDAD DE DOS TIPOS DE DIETA

Se desea comparar 2 tipos de dieta. De una población de 80 músicos bastante gordos se aplica la dieta tipo A a 45 de ellos y la dieta tipo B al grupo restante. Las pérdidas de peso en libras durante el período de una semana son las siguientes

	Muestra	Media muestral	Varianza muestral
Dieta A	45	10.3	7
Dieta B	35	7.3	3.25

¿ Permiten los datos concluir que la pérdida de peso bajo la dieta A es mayor que la pérdida de peso bajo la dieta B? Use $\alpha = .10$.

Se quiere contrastar H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$,

$$\sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}} = .49841$$

a) Aplicando el Método de Dixon- Massey se tiene que $T_c = 6.0191$, $T_{\alpha/2,\gamma} = 1.6644$, que se obtiene interpolando los valores de la tabla, donde los grados de libertad son $\gamma = 78.6124$, el intervalo de confianza es

$$2.17040 < \mu_1 - \mu_2 < 3.82959$$

b) Aplicando Cochran- Cox donde $w_1 = .6283$, $w_2 = ..0533$, ${}_1T_{\alpha/2,\gamma} = 1.6803$, ${}_2T_{\alpha/2\gamma} = 1.69114$, ya que los tamaños de las muestras son diferentes y el intervalo de confianza es

$$2.1605 < \mu_1 - \mu_2 < 3.8394$$

c) Aplicando Smith-Satterthwaite se obtiene $\gamma = 76.8$, interpolando se tiene que $T_{\alpha/2,\gamma} = 1.6649$, por lo que el intervalo de confianza es

$$2.17014 < \mu_1 - \mu_2 < 3.8298$$

d) Por el método de Peña Sánchez el término corrector $\Delta = 1.2042$, por lo que el valor de $\gamma = 76.7957$ y $T_{\alpha/2,\gamma} = 1.7438$; por lo que el intervalo que resulta es

$$2.17014 < \mu_1 - \mu_2 < 3.8298$$

por lo que la hipótesis nula se rechaza, con la dieta A se baja más de peso que con la dieta B.

4.3.4 LA EFICACIA ENTRE DOS MÉTODOS

Se desea probar un método rápido, pero no preciso, para estimar la concentración de un compuesto químico en un proceso de fabricación. Se analizan 8 muestras por el método rápido obtiene una media de 21 y una varianza de17.71 y 4 por el método estándar con una media de 25 y una varianza de 0.67. ¿ Es más efectivo el método rápido ?

Se quiere contrastar H_0: $\mu_1 = \mu_2$ vs H_A: $\mu_1 \neq \mu_2$

$$\sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}} = 1.5427$$

a) Aplicando Dixon- Masseycon $\alpha = .05$ se tiene que $T_c = 2.60$, $T_{\alpha/2,\gamma} = 2.2916$, que se obtiene interpolando los valores de la tabla donde los grados de libertad son $\gamma = 8.3278$, el intervalo de confianza es

$$.46463 < \mu_1 - \mu_2 < 7.53536$$

b) Aplicando Cochran- Cox donde $w_1 = .17$, $w_2 = 2.21$, $T'_{\alpha/2} = 2.42$. $T_c = 2.69$ con y el intervalo de confianza es

$$.2666 < \mu_1 - \mu_2 < 7.7333$$

c) Aplicando Smith-Satterthwaite $\gamma = 8.0077$, interpolando se tiene $T_{\alpha/2,\gamma} = 2.3056$, por lo que el intervalo de confianza es

$$.44300 < \mu_1 - \mu_2 < 7.5569$$

d) Por el método de Peña Sánchez el término corrector $\Delta = 4.1735$, por lo que el valor de $\gamma = 5.8264$ y $T_{\alpha/2,\gamma} = 2.4683$; por lo que el intervalo que resulta es

$$.19198 < \mu_1 - \mu_2 < 7.80801$$

se concluye que el método rápido subestima al método estándar. Se debe seguir usando el método estándar.

4.4 OBSERVACIONES APAREADAS

Cuando se tiene evidencia de que las varianzas de la poblaciones no son iguales, el aparear las observaciones de las muestras resulta de gran utilidad como se

explicó con anterioridad, esto se ve en los siguientes ejemplos.

4.4.1 DIFERENCIA ENTRE DOS PRODUCTOS

Un fabricante desea comparar la resistencia al desgaste de dos tipos de llantas A y B. Para hacer la comparación asignó al azar una llanta A y B a las ruedas posteriores de 5 autos. Los autos recorrieron un número específico de kilómetros y se observó el desgaste de cada llanta registrándolos en la siguiente tabla

Auto	llanta A	llanta B	Diferencia
1	10.6	10.2	0.4
2	9.8	9.4	0.4
3	12.3	11.8	0.5
4	9.7	9.1	0.6
5	8.8	8.3	0.5

¿ Existe diferencia significativa entre el desgaste medio de los dos tipos de llantas? Use un nivel de significación de 5%.

La hipótesis a probar es H_0: $\mu_1 - \mu_2 = 0$ vs H_A: $\mu_1 - \mu_2 = 0$

Los cálculos obtenidos son $\overline{D} = 0.48$ y $\hat{S}_D = 0.0837$ y el intervalo es

(.3760,.5839)

T_c = 12.8285 y la $T_{\alpha/2,4} = \pm 2.7764$ por lo que se rechaza la hipótesis nula y se concluye que si hay una diferencia significativa en los tipos de llanta.

4.4.2 EFECTOS DE UN ESTIMULANTE

Un cierto estimulante va a ser contrastado para comprobar sus efectos en la presión sanguínea. Se midió la presión sanguínea de 12 hombres antes y después del estimulante. ¿ Hay razón suficiente para creer que el estimulante subiría por término medio la presión sanguínea 5 puntos ?

Utilizamos el apareamiento ya que se hacen dos pruebas a un mismo individuo. La muestra consta de 12 individuos con dos mediciones cada uno.

Los resultados obtenidos se indican en la siguiente tabla:

Hombre	Antes	Después	Aumento
1	120	128	8
2	124	134	7
3	130	131	1
4	118	127	9
5	140	132	-8
6	128	125	-3
7	140	141	1
8	135	137	2
9	126	118	-8
10	130	132	2
11	126	129	3
12	127	135	8

H_0: $\mu_1-\mu_2 \leq 5$ que es una prueba de una cola, con un nivel de significación de 5% se tiene que

$\overline{D} = 1.8333$ y $\hat{S}_D = 5.83$ donde (-1.1960,4.8626)

T_c = -1.9 y la $T_{\alpha/2,11}$ = 1.80 por lo que se acepta la hipótesis ya que $T_c > 1.80$.

4.4.3 UN ESTUDIO ESTADÍSTICO

En unas prácticas de formación estadística se eligieron aleatoriamente en una clase de primer grado a 10 niños y 10 niñas. Las calificaciones fueron obtenidas midiendo su habilidad para aprender sílabas sin sentido. Ya que el experimentador sospechó que la variación para niños y niñas sería diferente decidió aparear los datos. El análisis en este experimento es apropiado para realizar un apareamiento ya que se pueden formar pares de individuos pues tienen características similares, por lo que el apareamiento con respecto al CI puede producir una reducción en la varianza de la población, haciendo más fácil descubrir cualquier diferencia en la habilidad de aprendizaje. Las observaciones y el análisis se presentan a continuación:

Tabla para la diferencia entre 10 pares

Niños	28	18	22	27	25	30	21	21	20	27
Niñas	19	38	42	25	15	31	22	37	30	24
Diferencia	9	-20	-20	2	10	-1	-1	-16	-10	3

H_0: $\mu_1-\mu_2 = 0$ vs H_A: $\mu_1-\mu_2 = 0$, es decir, que la calificación media de aprendizaje de los niños es igual a la de las niñas.

Los cálculos realizados dan los siguientes resultados $\overline{D} = -4.4$ y $\hat{S}_D = 11.34$ T_c = -1.2 y la $T_{\alpha/2,9} = \pm 3.25$ por lo que se acepta la hipótesis. Este experimento no indica una diferencia en la habilidad entre los niños y las niñas para aprender sílabas sin sentido. El intervalo de confianza es

(-16.0545,7.2545)

CONCLUSIONES

Durante el desarrollo del presente trabajo se ha mencionado la importancia de comparar parámetros provenientes de dos poblaciones mediante la Inferencia Estadística y de esta forma llegar a una decisión respecto a las mismas siendo está la única forma científica de sacar este tipo de conclusiones y avanzar en el campo de la ciencia.

Se ha dado una sinopsis sobre intervalos de confianza para la diferencia de medias entre dos poblaciones normales, que es muy útil pues con de esta forma garantizamos que la diferencia entre los parámetros inferidos se encuentre contenida en el intervalo con cierto grado de probabilidad. Los métodos expuestos abarcan los casos cuando las varianzas de las poblaciones son iguales y diferentes. Para saber en qué caso nos encontramos se recomienda realizar la prueba F para así poder escoger el método correspondiente.

En el caso donde las varianzas son iguales se puede aplicar el método normal o la distribución t de student sin mayor problema, únicamente se debe revisar si se conoce o no la varianza.

Cuando las varianzas son diferentes éstas no se pueden combinar como en el caso de varianzas iguales por lo que ya se han desarrollado distintos métodos han ayudado a resolver el problema. Estos métodos utilizan también la prueba t que es muy importante en estadística. La diferencia entre cada método radica en el número de grados de libertad que utilizan ya que a mayor número de grados de libertad disminuye el valor de las columnas en la tabla de la distribución t para los distintos valores de α, lo que da como resultado obtener intervalos de confianza más cortos y por lo tanto más eficientes.

Por lo general cualquiera de los métodos anteriores puede ser utilizado sin problema alguno pues los resultados obtenidos en las aplicaciones dan evidencia de que los métodos no difieren de manera significativa para la obtención de los intervalos. Si se requiere de mayor precisión en el cálculo de los intervalos para obtener el más eficiente manejando más de cuatro dígitos después del punto en los valores, se recomienda utilizar el método de Dixon-Massey.

El método de observaciones apareadas tiene la ventaja de toda la cantidad de información puede reducirse al de una sola muestra lo que facilita su manejo y nos ahorra tiempo, además tiene suposiciones que lo hacen bastante eficiente y beneficioso para el investigador permitiendo aumentar la precisión en la inferencia. Lo único que se necesita es que el tamaño de las muestras sea igual y que los pares formados tengan las semejanzas que se requieren para el estudio o investigación.

La experiencia al realizar este trabajo ha sido el comprobar que la estadística es una herramienta fundamental en el desarrollo científico y tecnológico pues tiene una gran variedad de aplicaciones como lo es el diseño de experimentos, la toma de decisiones cuando existen condiciones de incertidumbre, el análisis de datos para así convertirlos en información etc. Durante el desarrollo del trabajo se han aplicado los conocimientos adquiridos en la carrera, específicamente en la materia de estadística para desarrollar los temas expuestos.

APÉNDICE

TABLAS ESTADÍSTICAS

A1 TABLA DE LA DISTRIBUCIÓN NORMAL

A2 TABLA DE LA DISTRIBUCIÓN T DE STUDENT

Probabilidad Normal Estándar de cola superior
(para valores negativos de z, las áreas se obtienen por simetría)

	Segundo decimal de z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013									
3.5	0.00023267									
4.0	0.00003169									
4.5	0.00000340									
5.0	0.00000029									

$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	g.l.
3.0777	6.3137	12.7062	31.8210	63.6559	**1**
1.8856	2.9200	4.3027	6.9645	9.9250	**2**
1.6377	2.3534	3.1824	4.5407	5.8408	**3**
1.5332	2.1318	2.7765	3.7469	4.6041	**4**
1.4759	2.0150	2.5706	3.3649	4.0321	**5**
1.4398	1.9432	2.4469	3.1427	3.7074	**6**
1.4149	1.8946	2.3646	2.9979	3.4995	**7**
1.3968	1.8595	2.3060	2.8965	3.3554	**8**
1.3830	1.8331	2.2622	2.8214	3.2498	**9**
1.3722	1.8125	2.2281	2.7638	3.1693	**10**
1.3634	1.7959	2.2010	2.7181	3.1058	**11**
1.3562	1.7823	2.1788	2.6810	3.0545	**12**
1.3502	1.7709	2.1604	2.6503	3.0123	**13**
1.3450	1.7613	2.1448	2.6245	2.9768	**14**
1.3406	1.7531	2.1315	2.6025	2.9467	**15**
1.3368	1.7459	2.1199	2.5835	2.9208	**16**
1.3334	1.7396	2.1098	2.5669	2.8982	**17**
1.3304	1.7341	2.1009	2.5524	2.8784	**18**
1.3277	1.7291	2.0930	2.5395	2.8609	**19**
1.3253	1.7247	2.0860	2.5280	2.8453	**20**
1.3232	1.7207	2.0796	2.5176	2.8314	**21**
1.3212	1.7171	2.0739	2.5083	2.8188	**22**
1.3195	1.7139	2.0687	2.4999	2.8073	**23**
1.3178	1.7109	2.0639	2.4922	2.7970	**24**
1.3163	1.7081	2.0595	2.4851	2.7874	**25**
1.3150	1.7056	2.0555	2.4786	2.7787	**26**
1.3137	1.7033	2.0518	2.4727	2.7707	**27**
1.3125	1.7011	2.0484	2.4671	2.7633	**28**
1.3114	1.6991	2.0452	2.4620	2.7564	**29**
1.3104	1.6973	2.0423	2.4573	2.7500	**30**
1.3062	1.6896	2.0301	2.4377	2.7238	**35**
1.3031	1.6839	2.0211	2.4233	2.7045	**40**
1.3007	1.6794	2.0141	2.4121	2.6896	**45**
1.2987	1.6759	2.0086	2.4033	2.6778	**50**
1.2971	1.6730	2.0040	2.3961	2.6682	**55**
1.2958	1.6706	2.0003	2.3901	2.6603	**60**
1.2947	1.6686	1.9971	2.3851	2.6536	**65**
1.2938	1.6669	1.9944	2.3808	2.6479	**70**
1.2929	1.6654	1.9921	2.3771	2.6430	**75**
1.2922	1.6641	1.9901	2.3739	2.6387	**80**
1.2916	1.6630	1.9883	2.3710	2.6349	**85**
1.2910	1.6620	1.9867	2.3685	2.6316	**90**
1.2905	1.6611	1.9852	2.3662	2.6286	**95**
1.2901	1.6602	1.9840	2.3642	2.6259	**100**

BIBLIOGRAFÍA

Lyman Ott. *An introduction to Statistical Methods and Data Analysis* (1977). Duxbury Press. North Scituate, Massachusetts.

Ostle B. y Mensing R. W. *Statistics in Research* (1975, 3ra edición)Edición) The Iowa University Press.

Anderson R. L. y T. A. Bancroft. *Statistical Theory in Research* (1952). Mcgraw-Hill, Nueva York.

Dixon W. J. y F. J. Massey. *Introduction to Statistical Analysis* (1969, 3ra. Edición) Mcgraw-Hill, Nueva York.

Peña Sánchez de Rivera D. *Estadística Modelos y Métodos. Fundamentos*, Vol. Y (1986) Alianza Editorial, Madrid.

Snedecor George W. y William G. Cochran. *Métodos Estadísticos* (8va. Impresión 1981) Compañía Editorial Continental S. A. México.

Mood A. Mc. y F. A. Graybill. *Introduction to the Theory of Statistics* (1974, 3ra. Edición) Mcgraw-Hill, Nueva York.

Hogg Robert V. y Allen T. Craig. *Introduction to Mathematical Statistics.* (2a. Edición. 1965) The Mcmillan Company.

Freund John E. y Ronald E. Walpole. *Estadística Matemática con Aplicaciones.* (4a. Edición 1990) Prentice Hall Hispanoamericana S.A.

Larson Harold J. *Introducción a la Teoría de Probabilidades e inferencia Estadística.*(9a. Reimpresión. 1992) Editorial Limusa. Grupo Noriega Editores.

Walpole Ronald E. y Raymond H. Myers. *Probabilidad y Estadística.* (3ra. Edición en Español. 1992)