# FACULTAD DE CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS

# CONTABILIDAD Y AUDITORÍA

## ESTADÍSTICA INFERENCIAL


**INTEGRANTES:**

Dolores Calle.

Patricia Suárez.

Braulio Pozo.

Marco Yunga.

**PROFESOR:**

Ing. Roberto Molina.

**CICLO:**

Cuarto, aula 12.


**Ciclo lectivo:**

Marzo – Julio de 2010.

# Contenido

CAPÍTULO 6: DISTRIBUCIONES MUESTRALES ..... 3
CAPÍTULO 7: ESTIMACIÓN CON INTERVALOS DE CONFIANZA ..... 21
CAPÍTULO 8: PRUEBA DE HIPÓTESIS ..... 35
CAPÍTULO 9: PRUEBAS CUANDO SE TIENEN DOS COLAS ..... 56
CAPÍTULO 11: REGRESIÓN SIMPLE Y CORRELACIÓN ..... 77
Bibliografía. ..... 99

# CAPÍTULO 6: DISTRIBUCIONES MUESTRALES

## Estadística Inferencial

- ✓ La **inferencia estadística** o **estadística Inferencial** es una parte de la Estadística que comprende los métodos y procedimientos para deducir propiedades (hacer inferencias) de una población, a partir de una pequeña parte de la misma (muestra).
- ✓ La estadística Inferencial trata sobre las inferencias con respecto a poblaciones (sus parámetros, μ y σ2) a partir de la información contenida en las muestras (los estadísticos, X y S2).

Para poder llevar a cabo esas inferencias es necesario conocer la relación que se establece entre estadísticos y parámetros. El concepto que permite poner en relación ambas cosas es "*La distribución muestral"*

## Muestreo.

En estadística se conoce como **muestreo** a la técnica para la selección de una muestra a partir de una población.

Al elegir una muestra se espera conseguir que sus propiedades sean extrapolables a la población. Este proceso permite ahorrar recursos, y a la vez obtener resultados parecidos a los que se alcanzarían si se realizase un estudio de toda la población.

## Distribución Muestral.

La **distribución muestral** es lo que resulta de considerar todas las muestras posibles que pueden ser tomadas de una población. Su estudio permite calcular la probabilidad que se tiene, dada una sola muestra, de acercarse al parámetro de la población. Mediante la distribución muestral se puede estimar el error para un tamaño de muestra dado.

La fórmula para la distribución muestral dependerá de la distribución de la población, del estadístico y del tamaño de la muestra.

A partir de las muestras seleccionadas de una población pueden construirse variables aleatorias alternativas, de cuyo análisis se desprenden interesantes propiedades estadísticas. Las dos formas más comunes de estas variables corresponden a las distribuciones muéstrales de las medias y de las proporciones.

**Ejemplo**: Tenemos una población con los siguientes $N$ = 3 elementos: $X$ = {1, 2 y 3}.

Donde μ = 2 σ2 = 0,67.
Se extraen muestras de *n* = 2 elementos:
Con reposición, tenemos 9 posibles muestras:
(1, 1); (1, 2); (1, 3); (2, 1*); (*2, 2); (2, 3); (3, 1); (3, 2); y (3, 3).
Sin reposición, tenemos 6 posibles muestras:
(1, 2); (1, 3); (2, 1); (2, 3); (3, 1); y (3, 2).
En cada una de las muestras pueden calcularse los correspondientes estadísticos descriptivos:
Por ejemplo, con reposición:
Las medias serían: 1; 1,5; 2; 1,5; 2; 2,5; 2; 2,5; y 3, respectivamente.
Las varianzas serían: 0; 0,25; 1; 0,25; 0; 0,25; 1; 0,25; y 0, respectivamente.
Por tanto, los estadísticos descriptivos son variables aleatorias que pueden adoptar diferentes valores y que tienen su propia distribución de probabilidad.
En el ejemplo vemos que X puede tomar 5 posibles valores y que la probabilidad que corresponde a cada uno de ellos (*f* (X *i*), su distribución) es:

X *i* 1 1,5 2 2,5 3 Total:
*f* ( X *i*) 1/ 9 2/ 9 3/ 9 2/ 9 1/ 9 1

Donde E(X) = Σ X*i* · *f* (Xi*)* = (1)(1/ 9) + (1,5)(2/ 9) + … + (3)(1/ 9) = 2
σ 2 ( X ) = Σ [ X*i* 2 · *f* ( X*i* )] – [ E( X )] 2 = [(12)(1/ 9) + … + (32)(1/ 9)] - 22 = 0,33

No es necesario construir la distribución de un estadístico (p.e. de X ) en todos los casos ya que cada estadístico tiene su propia distribución muestral conocida.

En este tema nos ocuparemos de la distribución muestral de la media: X y de la proporción: P.

# Error De Muestreo.

En estadística, **error de muestreo** o **error de la valoración** es error causado observando una muestra en vez de la población entera.
El error de muestreo se puede poner en contraste con error del no-muestreo. el error del No-muestreo es un coger-todo término para las desviaciones del valor verdadero que no son una función de la muestra elegida, incluyendo vario errores sistemáticos y cualquier es errores al azar que no sean debido al muestreo. Los errores del No-Muestreo son mucho más duros de cuantificar que error de muestreo.

*"Es la muestra entre el parámetro poblacional y él estadístico de la muestra utilizado para estimar el parámetro." (X – μ)*

# Distribución muestral de la media.

Dada una población constituida por un número n de elementos, cuya **media aritmética** es ⍰ y donde la **desviación típica** viene dada ⍰, pueden formarse $n^2$ muestras con re emplazamiento distintas, formadas por dos elementos de la población.

Para cada una de estas muestras es posible una **media muestral**, que denotaremos con el símbolo $\bar{x}$. Un ejemplo de la tabla de muestras de tamaño 2, tomada de la población {1, 3, 5}, con sus medias aritméticas reflejadas, sería:

| Muestra | 1,1 | 1,3 | 1,5 | 3,1 | 3,3 | 3,5 | 5,1 | 5,3 | 5,5 |
|---|---|---|---|---|---|---|---|---|---|
| Media $\bar{x}_i$ | 1 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 5 |

A partir de la variable estadística original x de la población se puede construir una nueva variable estadística $\bar{x}$, que tendría como valores las medias de las muestras tomadas de la población. La media aritmética de esta **distribución muestral de las medias** se denota por $\mu\bar{x}$ y su desviación típica por $\sigma\,\bar{x}$ .

Parámetros de la distribución muestral de las medias de tamaño 2

Establecida una distribución muestral de las medias de tamaño 2, su **esperanza matemática** $E[X] = \mu = \mu_{\bar{x}_i} = E[\bar{X}]$ adopta el valor siguiente:

Siendo m la media aritmética de la población $\bar{x}_i$ la media aritmética de cada muestra, U$\bar{x}_i$ la media aritmética de todas las medias, E [x] la esperanza matemática de la variable aleatoria x (para la población) y E [ $\bar{x}$ ] la esperanza matemática de la variable aleatoria $\bar{x}$ (para la distribución muestral de las medias).

Por su parte, los valores de la varianza y la desviación típica de esta distribución muestral de tamaño 2 son: $V[X] = \sigma^2$ ; $V[\bar{X}] = \sigma_{\bar{x}}$ ; $\sigma_{\bar{x}} = s/\sqrt{2}$ donde ⍰ es la desviación típica de la población, $\sigma\bar{x}$ la desviación típica de la distribución muestral, V [x] la varianza de la variable x (población) y V [$\bar{x}$] la varianza de la variable $\bar{x}$ (distribución muestral de las medias).

# Distribución muestral de las medias de tamaño n.

En una distribución muestral de las medias, la variable aleatoria media muestral sigue una ley normal descrita como N (m,s/Ön).

| | | Extracción | |
|---|---|---|---|
| | | Con reemplazo | Sin reemplazo |
| Población | Infinita | $\mu_{\bar{x}} = \mu$<br>$\sigma_{\bar{x}} = \sigma/\sqrt{n}$ | $\mu_{\bar{x}} = \mu$<br>$\sigma_{\bar{x}} = \sigma/\sqrt{n}$ |
| | Finita (N) | $\mu_{\bar{x}} = \mu$<br>$\sigma_{\bar{x}} = \sigma/\sqrt{n}$ | $\mu_{\bar{x}} = \mu$<br>$\sigma_{\bar{x}} = (\sigma/\sqrt{n}) \cdot \sqrt{(N-n)/(N-1)}$ |

Parámetros estadísticos de una distribución muestral de las medias de tamaño n:

## Distribución muestral de las proporciones.

Sea una población formada por n elementos, de los cuales algunos poseen una determinada característica y otros no (llamaremos p a la proporción de los elementos que poseen la característica, y q = 1 - p a la de los restantes elementos). Entonces, es posible extraer muestras de la población de manera que a cada una se asocie como valor la proporción de la característica analizada.

Por ejemplo, en la población {1, 2, 3}, la característica par tiene un valor p = 1 / 3, mientras que la impar es q = 2 / 3. Mediante la tabla siguiente de muestras se construye una nueva **distribución muestral de las proporciones**.

| **Muestra** | 1,1 | 1,2 | 1,3 | 2,1 | 2,2 | 2,3 | 3,1 | 3,2 | 3,3 |
|---|---|---|---|---|---|---|---|---|---|
| **Proporción f/n** | 0 | 0,5 | 0 | 0,5 | 0 | 0,5 | 0 | 0,5 | 0 |

Parámetros estadísticos de una distribución muestral de las proporciones de tamaño n:

| | | Extracción | |
|---|---|---|---|
| | | Con reemplazo | Sin reemplazo |
| Población | Infinita | $\mu_p = \mu$<br>$\sigma_p = \sqrt{pq/n}$ | $\mu_p = \mu$<br>$\sigma_p = \sqrt{pq/n}$ |
| | Finita (N) | $\mu_p = \mu$<br>$\sigma_p = \sqrt{pq/n}$ | $\mu_p = \mu$<br>$\sigma_p = \sqrt{(pq/n)} \cdot \sqrt{(N-n)/(N-1)}$ |

Una distribución muestral de las proporciones se comporta como una distribución normal descrita por los parám$(p, \sqrt{pq/n})$.

**Ejemplo 1**

Se toman muestras de tamaño 2 de una población consistente en tres valores, 2, 4 y 6, para simular una población "grande" de manera que el muestreo pueda realizarse un gran número de veces, supondremos que éste se hace con reemplazo, es decir, el número elegido se reemplaza antes de seleccionar el siguiente, además, se seleccionan muestras ordenadas. En una muestra ordenada, el orden en que se seleccionan las

observaciones es importante, por tanto, la muestra ordenada (2,4) es distinta de la muestra ordenada (4,2). En la muestra (4,2), se seleccionó primero 4 y después 2. La siguiente tabla contiene una lista de todas las muestras ordenadas de tamaño 2 que es posible seleccionar con reemplazo y también contiene las medidas muéstrales y los correspondientes errores muéstrales. La media poblacional es igual a $\mu$= (2+4+6)/3 = 4. Ver la tabla en la siguiente página.

Nótese las interesantes relaciones siguientes contenidas en la tabla:

La media de la colección de medias muéstrales es 4, la media de la población de la que se extraen las muestras. Si $\mu_x$ denota la media de todas las medias muéstrales entonces tenemos:

$\mu_x$ = (3+4+3+4+5+5+2+4+6)/9 = 4

La suma de los errores muéstrales es cero.

$e_1 + e_2 + e_3 + \ldots + e_9$ = (-2) + (-1) + 0 + (-1) + 0 + 1 + 0 + 1 + 2 = 0

| **Muestras ordenadas** | **x** | **Error muestral e = x - $\mu$** |
|---|---|---|
| (2,2) | 2 | 2 – 4 = -2 |
| (2,4) | 3 | 3 – 4 = -1 |
| (2,6) | 4 | 4 – 4 = 0 |
| (4,2) | 3 | 3 – 4 = -1 |
| (4,4) | 4 | 4 – 4 = 0 |
| (4,6) | 5 | 5 – 4 = 1 |
| (6,2) | 4 | 4 – 4 = 0 |
| (6,4) | 5 | 5 – 4 = 1 |
| (6,6) | 6 | 6 – 4 = 2 |

En consecuencia, si x se usa para medir, estimar, la media poblacional $\mu$, el promedio de todos los errores muéstrales es cero.

# Teorema del límite central.

El **teorema del límite central** o **teorema central del límite** indica que, en condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución normal (también llamada *distribución gaussiana* o *curva de Gauss* o *campana de Gauss*) cuando la cantidad de variables es muy grande.

**Teorema**: Sea $X_1$, $X_2$, ..., $X_n$ una muestra aleatoria de una distribución con media **μ** y varianza $\boldsymbol{\sigma^2}$. Entonces, si **n** es suficientemente grande, la variable aleatoria

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

Tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$.

También se cumple que si

$$T_0 = \sum_{i=1}^{n} X_i$$

Tiene aproximadamente una distribución normal con $\mu_{T_0} = n\mu$ y $\sigma^2_{T_0} = n\sigma^2$, cuanto más grande sea el valor de **n**, mejor será la aproximación.

El teorema del límite central garantiza una distribución normal cuando n es suficientemente grande.

Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.

La aproximación entre las dos distribuciones es, en general, mayor en el centro de las mismas que en sus extremos o colas, motivo por el cual se prefiere el nombre "teorema del límite central" ("central" califica al límite, más que al teorema).

Este teorema, perteneciente a la teoría de la probabilidad, encuentra aplicación en muchos campos relacionados, tales como la inferencia estadística o la teoría de renovación.

El **Teorema Central del Límite** también nos dice que si tenemos un grupo numeroso de variables independientes y todas ellas siguen el mismo modelo de distribución (cualquiera que éste sea), la suma de ellas se distribuye según una **distribución normal**.

**Media**: **n * m** (media de la variable individual multiplicada por el número de variables independientes)

**Varianza**: **n * $s^2$** (varianza de la variable individual multiplicada por el número de variables individuales)

**Ejercicio 1.**

La renta media de los habitantes de un país se distribuye uniformemente entre 4,0 millones ptas. y 10,0 millones ptas. Calcular la probabilidad de que al seleccionar al azar a 100 personas la suma de sus rentas supere los 725 millones ptas.

Cada renta personal es una variable independiente que se ditribuye según una función uniforme. Por ello, a la suma de las rentas de 100 personas se le puede aplicar el **Teorema Central del Límite.**

La **media** y **varianza** de cada variable individual es:

$$\mathbf{m} = (4 + 10) / 2 = 7$$

$$\mathbf{s^2} = (10 - 4)^2 / 12 = 3$$

Por tanto, la suma de las 100 variables se distribuye según una normal cuya **media** y **varianza** son:

**Media**: **n * m** = 100 * 7 = 700

**Varianza** : **n * $s^2$** = 100 * 3 = 300

Para calcular la probabilidad de que la suma de las rentas sea superior a 725 millones ptas, comenzamos por calcular el valor equivalente de la variable normal tipificada:

$$Y = \frac{725 - 700}{17,3} = 1,44$$

Luego:

$$P(X > 725) = P(Y > 1,44) = 1 - P(Y < 1,44) = 1 - 0,9251 = 0,0749$$

Es decir, la probabilidad de que la suma de las rentas de 100 personas seleccionadas al azar supere los 725 millones de pesetas es tan sólo del 7,49%

**Ejercicio 2.**

En una asignatura del colegio la probabilidad de que te saquen a la pizarra en cada clase es del 10%. A lo largo del año tienes 100 clases de esa asignatura. ¿Cuál es la probabilidad de tener que salir a la pizarra más de 15 veces?

Se vuelve a aplicar el **Teorema Central del Límite**.

Salir a la pizarra es una variable independiente que sigue el modelo de distribución de Bernouilli:

**"Salir a la pizarra"**, le damos el valor 1 y tiene una probabilidad del 0,10

**"No salir a la pizarra"**, le damos el valor 0 y tiene una probabilidad del 0,9

La **media** y la **varianza** de cada variable independientes es:

$$\mathbf{m} = 0{,}10$$

$$\mathbf{s^2} = 0{,}10 * 0{,}90 = 0{,}09$$

Por tanto, la suma de las 100 variables se distribuye según una normal cuya **media** y **varianza** son:

**Media**: $\mathbf{n * m} = 100 * 0{,}10 = 10$

**Varianza**: $\mathbf{n * s^2} = 100 * 0{,}09 = 9$

Para calcular la probabilidad de salir a la pizarra más de 15 veces, calculamos el valor equivalente de la variable normal tipificada:

$$Y = \frac{15 - 10}{3{,}0} = 1{,}67$$

Luego:

$$P(X > 15) = P(Y > 1{,}67) = 1 - P(Y < 1{,}67) = 1 - 0{,}9525 = 0{,}0475$$

Es decir, la probabilidad de tener que salir más de 15 veces a la pizarra a lo largo del curso es tan sólo del 4,75%.

# Medias muestrales.

La distribución de muestreo de medias muéstrales es la distribución de probabilidad de todas la medias muéstrales posibles de un tamaño de muestra dado, seleccionadas de una población, y la probabilidad de ocurrencia asociada con cada media muestral.

Ejemplo:
El despacho de abogados Hoya & Asociados tiene cinco socios.
En su junta de socios semanal cada uno informa el número de horas que cobraron a los clientes por sus servicios la semana anterior.

| Socio | Horas |
|---|---|
| Giordanino | 22 |
| Barnabá | 26 |
| Torres | 30 |
| Carrizo | 26 |
| Salgado | 22 |

Si se seleccionan al azar dos socios, ¿cuántas muestras diferentes son posibles?.

Organice las medias muéstrales en una distribución de muestreo.

| Media Muestral | Frecuencia | Frecuencia relativa |
|---|---|---|
| 22 | 1 | 1/10 |
| 24 | 4 | 4/10 |
| 26 | 3 | 3/10 |
| 28 | 2 | 2/10 |

Calcule la media de las medias muéstrales y compárela con la media poblacional:

## Media de las medias muéstrales.

= [(22)(1) + (24)(4) + (26)(3) + (28)(2)]/10=25.2

Media poblacional = (22+26+30+26+22)/5 = 25.2

Observe que la media de las medias muéstrales es igual a la media poblacional.

- Sea una población compuesta por cinco números: ***2, 3, 6, 8, 11***. En este caso, es fácil, calcular la media y la varianza de la población:

$$(a) \quad \mu = \frac{2 + 3 + 6 + 8 + 11}{5} = 6.00$$

$$(b) \quad \sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} = 10{,}8$$

| Muestras ; n=2 | Media muestra | Varianza muestra | Cuasi varianza muestra |
|---|---|---|---|
| (2,2) | 2.00 | 0.00 | 0.00 |
| (3,2) | 2.50 | 0.25 | 0.50 |

| | | | |
|---|---|---|---|
| (6,2) | 4.00 | 4.00 | 8.00 |
| (8,2) | 5.00 | 9.00 | 18.00 |
| (11,2) | 6.50 | 20.25 | 40.50 |
| (2,3) | 2.50 | 0.25 | 0.50 |
| (3,3) | 3.00 | 0.00 | 0.00 |
| (6,3) | 4.50 | 2.50 | 5.00 |
| (8,3) | 5.50 | 6.25 | 12.50 |
| (11,3) | 7.00 | 16.00 | 32.00 |
| (2,6) | 4.00 | 4.00 | 8.00 |
| (3,6) | 4.50 | 2.25 | 4.50 |
| (6,6) | 6.00 | 0.00 | 0.00 |
| (8,6) | 7.00 | 1.00 | 2.00 |
| (11,6) | 8.50 | 6.25 | 12.50 |
| (2,8) | 5.00 | 9.00 | 18.00 |
| (3,8) | 5.50 | 6.25 | 12.50 |
| (6,8) | 7.00 | 1.00 | 2.00 |
| (8,8) | 8.00 | 0.00 | 0.00 |
| (11,8) | 9.50 | 2.25 | 4.50 |
| (2,11) | 6.50 | 20.25 | 40.50 |
| (3,11) | 7.00 | 16.00 | 32.00 |
| (6,11) | 8.50 | 6.25 | 12.50 |
| (8,11) | 9.50 | 2.25 | 4.50 |
| (11,11) | 11.00 | 0.00 | 0.00 |

La media de las medias muéstrales es:

$$\mu_{\overline{X}} = \frac{\text{suma de todas las medias muestrales (columna2)}}{25} = \frac{150}{25} = 6{,}0$$

Que coincide con la media de la población, es decir:

$$\mu_{\overline{X}} = \mu$$

La varianza de la distribución muestral de las medias, será la varianza de los elementos de la columna 2 (medias muéstrales), que es:

$$\sigma^2_{\bar{X}} = \frac{(2-6)^2 + \cdots + (11-6)^2}{25} = \frac{135}{25} = 5{,}40$$

Por tanto, la relación entre la varianza de la distribución de las medias muéstrales y la varianza de la población es:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{N} \quad ; \quad \text{N es el tamaño de las muestras, en este caso, N} = 2$$

El valor más esperado, media o esperanza, de las varianzas de las muestras, valores de la columna 3, es:

$$E(s_n^2) = \frac{0.00 + 0.25 + \cdots + 2.25 + 0.00}{25} = 5.41 \quad \text{Valor sesgado respecto a } \sigma^2 = 10.8$$

## Distribución muestral de la proporción.

Si tenemos $n$ observaciones ($X_1, X_2, ..., X_n$) dicotómicas y definimos:

$X_i$ : "Número de aciertos con probabilidad $\pi$"

Entonces: $X \rightarrow B(x, n, \pi)$ $\quad E(X) = n \cdot \pi$

$$\sigma(X) = \sqrt{n\pi \cdot (1-\pi)}$$

**EJEMPLO:**

Distribución del *número de aciertos* en un test de 5 ítems con $\pi = 0{,}50$

| $X_i$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $f(x_i)$ | 0,031 | 0,156 | 0,312 | 0,312 | 0,156 | 0,031 |

Donde: $E(X) = (5)\,(0{,}50) = 2{,}5$

$$\sigma(X) = \sqrt{2{,}5 \cdot (0{,}50)} = 1{,}12$$

Si ahora definimos la variable

$$\boxed{P_i = \frac{X_i}{n}}$$ … "Proporción de aciertos con probabilidad $\pi$"

| |
|---|
| El estadístico proporción ($P$) se distribuye mediante el modelo Binomial: B ($x$, $n$, $\pi$). |

Donde: $E(P) = \pi$

$\sigma(P) = \sqrt{\pi \cdot (1 - \pi)/n}$ … Error típico de la proporción

Las probabilidades asociadas al estadístico $P$ pueden obtenerse mediante la tabla de la Binomial con parámetros $n$ y $\pi$.

En el ejemplo:

Distribución de la *propoción de aciertos* en un test de 5 ítems con $\pi = 0{,}50$

| $X_i$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $P_i$ | 0 | 0,20 | 0,40 | 0,60 | 0,80 | 1,00 |
| $f(x_i)$ | 0,031 | 0,156 | 0,312 | 0,312 | 0,156 | 0,031 |

Donde: $E(P) = 0{,}50$

$\sigma(P) = \sqrt{0{,}50 \cdot (0{,}50)/5} = 0{,}2236$

Por tanto:

1) Probabilidad de que se acierten el 40% de los ítems:

$P(P_i = 0{,}40) = P(X_i = 2) = 0{,}312$

2) Probabilidad de que se acierten como máximo el 60% de los ítems:

$P(P_i \leq 0{,}60) = P(X_i \leq 3) = 0{,}811$

## El error estándar de las medias muéstrales

El error estándar de las medias muéstrales es la desviación estándar de la distribución de muestreo de las medias muéstrales.
Se calcula mediante

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

**σx** es el símbolo del error estándar de las medias muéstrales.
**σ** es la desviación estándar de la población.
**n** es el tamaño de la muestra.

Si σ no se conoce y n σ 30, la desviación estándar de la muestra, denotada por s, se usa para aproximar la desviación estándar poblacional. La fórmula para el error estándar se convierte en:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

## La varianza y el error estándar de las medias muestrales.

Si se extrae una muestra al azar de tamaño n, de una población infinita con media μ y una varianza $s^2$, entonces las observaciones de la muestra son variables aleatorias independientes e idénticamente distribuidas. La media de la muestra, calculada como

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \bullet\bullet\bullet + x_n)$$

Que es una combinación lineal de variables aleatorias dividida por una constante, que

También es una variable aleatoria normal, y el valor esperado y la varianza de la distribución por muestreo de $\bar{x}$ puede derivarse sencillamente. Primero, observamos que

$$E(\bar{x}) = E\left[\frac{1}{n}(x_1 + x_2 + \bullet\bullet\bullet\ x_n)\right]$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \bullet\bullet\bullet + E(x_n)\right]$$

$$= \frac{1}{n}(n\mu) = \mu$$

Es decir, esperanza de la media de la muestra es la media de la población.

Luego, puesto que se considera que las observaciones de la muestra son variables aleatorias independientes, la propiedad de aditividad se verifica para la varianza. Es decir, la varianza de la suma es la suma de las varianzas. Además, puesto que $V(x_i) = \sigma^2$ tenemos

$$V(\bar{x}) - V\left(\frac{1}{n}\sum x_i\right)$$

$$= \frac{1}{n^2}\left[V(x_1) + V(x_2) + \cdots V(x_n)\right]$$

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

$$= \frac{\sigma^2}{n}$$

En esta derivación hemos empleado el teorema de que la varianza de una constante multiplicado por una variable es igual al cuadrado de la constante multiplicado por la varianza de la variable.

El error estándar de la media, mide la variabilidad entre medias muéstrales.

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

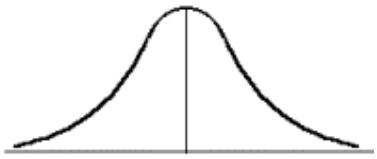
lo que revela que $\sigma_{\bar{x}}$ es menor que $\sigma$. Además, indica que cuando $n \to \infty$, $\sigma_{\bar{x}} \to 0$. Así, cuanto mayor es la muestra, tanto menor es la fluctuación entre medias muéstrales extraídas de la misma población.

Si se toman muestras de una población finita, sin reposición, como en los casos anteriores, debe de introducirse un factor de corrección para poblaciones finitas para calcular el error estándar de la media. A saber:

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

Cuando la población progenitora es normal, la distribución de $\bar{X}$ por muestreo es también normal, por pequeña que sea la muestra. ¿Qué ocurre cuando no puede especificarse la distribución de probabilidad de la población a partir de la cual se obtiene la muestra? Para obtener una idea con respecto a la distribución de muestreo de $\bar{X}$ cuando el modelo de probabilidad de la población de interés no se especifica.

Por lo tanto podemos decir que la



$\mu_{\bar{x}}$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

De esta manera la ecuación para la transformación de cualquier media muestral en una variable normal estándar será:

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

**Propiedades de la Media aritmética**

Entre varias propiedades matemáticas importantes de la media aritmética para una distribución normal están:

**Insesgamiento**

Implica el hecho de que el promedio de todas las medias muéstrales posibles (de un tamaño de muestra dado *n*) será igual a la media de población $\mu_x$ .

**Eficiente**
Se refiere a la precisión de la muestra de estadística como un estimador del parámetro de población.
Para distribuciones como la normal, la media aritmética se considera más estable de muestra a muestra que otras mediciones de tendencia central. Para una muestra de tamaño n, la media de muestra se acercará más, en promedio, a la media de población que cualquier otro estimador.

**Consistencia**

Se refiere al efecto del tamaño de muestra sobre la utilidad de un estimador. Al incrementarse el tamaño de muestra, la variación de la media de muestra de la media de población se hace más pequeña, de manera que la media aritmética de muestra se vuelve una mejor estimación de la media de población.

# Teorema del límite central.

El **teorema del límite central** o **teorema central del límite** indica que, en condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución normal (también llamada *distribución gaussiana* o *curva de Gauss* o *campana de Gauss*) cuando la cantidad de variables es muy grande.[1]

**Teorema**: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con media **μ** y varianza $\boldsymbol{\sigma^2}$. Entonces, si **n** es suficientemente grande, la variable aleatoria

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

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$.

También se cumple que si

$$T_0 = \sum_{i=1}^{n} X_i$$

Tiene aproximadamente una distribución normal con $\mu_{T_0} = n\mu$ y $\sigma^2_{T_0} = n\sigma^2$, cuanto más grande sea el valor de **n**, mejor será la aproximación.

El teorema del límite central garantiza una distribución normal cuando n es suficientemente grande.

Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.

La aproximación entre las dos distribuciones es, en general, mayor en el centro de las mismas que en sus extremos o colas, motivo por el cual se prefiere el nombre "teorema del límite central" ("central" califica al límite, más que al teorema).

Este teorema, perteneciente a la teoría de la probabilidad, encuentra aplicación en muchos campos relacionados, tales como la inferencia estadística o la teoría de renovación.

# Teorema del límite central.

***Teorema:***

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una distribución con media **μ** y varianza $\sigma^2$. Entonces, si **n** es suficientemente grande, **X** tiene aproximadamente una distribución normal con $\mu_x = \mu$ y $\sigma_x^2 = \sigma^2/n$, y $T_0$ tiene también aproximadamente una distribución normal con $\mu_{T0} = n.\mu$, $\sigma^2_{T0} = n.\sigma^2$. Cuanto mas grande sea el valor de **n**, mejor será la aproximación.

El Teorema del Límite Central garantiza una distribución normal cuando **n** es suficientemente grande

Si n > 30, se puede usar el TLC.

Si la distribución madre es normal, la distribución de la media muestral también es normal, independientemente del tamaño.

**x ≈** $N(\mu_x; \sigma_x)$ Þ $x \approx N(\mu_x; \sigma_x)$

Ejemplo 1:

Si se sabe que la dureza Rockwell de pernos de cierto tipo tiene un valor medio de 50 y desviación estándar de 1,5.

**a)** Si la distribución es normal, ¿cuál es la probabilidad de que la dureza muestral media para una muestra aleatoria de 9 pernos sea por lo menos 52?

**b)** ¿Cuál es la probabilidad (aproximada) de que la dureza muestral media para una muestra aleatoria de 40 pernos sea al menos 52?

**x =** 50

**σ =** 1,5

**x ≈** N(50; 1,5)

a)

n = 9

x = 52

x ≈ N(50; 1,5.√9)

z = (x - μ)/(σ/√n)

La probabilidad de que la media muestral sea superior a 52 es:

P(x ≥ 52) = $P\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{52-50}{\frac{1.5}{\sqrt{9}}}\right)$ Þ **P(z ≥ 4) = 0**

Con el valor de **z** obtenido de **<u>tablas</u>**:

P($x_1$ ≤ x ≤ $x_2$) = $P\left(z_1 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_2\right)$ Þ P($z_1$ ≤ z ≤ $z_2$) = ϕ(z)

Tener en cuenta que los valores para:

ϕ(z) = P(z ≤ $z_1$)

b)

n = 40

Con el valor de **z** obtenido de **<u>tablas</u>**:

P(x ≥ 52) = $P\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{52-50}{\frac{1.5}{\sqrt{40}}}\right)$ Þ **P(z ≥ 8,4327) = 0**

**EJERCICIOS:**

1. **.** En una localidad de 6000 habitantes, la proporción de menores de 16 años es p=1/4.

a) ¿Cuál es la distribución de la proporción de menores de 16 años en muestras de 50 habitantes de dicha población?.

b) Halla la probabilidad de que, en una muestra de 50 habitantes, haya entre 14 y 20 habitantes menores de 16 años?

2. La renta media de los habitantes de un país se distribuye uniformemente entre 4,0 millones ptas. y 10,0 millones ptas. Calcular la probabilidad de que al seleccionar al azar a 100 personas la suma de sus rentas supere los 725 millones ptas. Cada renta personal es una variable independiente que se distribuye según una función uniforme. Por ello, a la suma de las rentas de 100 personas se le puede aplicar el **Teorema Central del Límite.**

3. Un psicólogo clínico afirma que con su terapia para tratar "el miedo a volar en avión" se recupera el 80% de los pacientes. Si seleccionamos al azar 16 pacientes que han acudido a su consulta durante los últimos 3 meses por este tema, ¿cuál es la probabilidad de que al menos el 75% se hayan recuperado y puedan tomar aviones?

4. El peso de los libros de texto de un instituto se distribuyen de forma normal, con un peso media de $\mu = 400g$ y una desviación típica de $\sigma = 50g$. Si tomamos una muestra aleatoria de tamaño n = 16, hallar la probabilidad de que el peso medio esté entre 375 y 425 g

5. Explica si en los casos siguientes manejamos una población o una muestra:
   a) Un campesino tiene 87 gallinas. Para determinar si un nuevo pienso es eficaz, las pesa a todas antes y después de los 55 días durante los cuales las alimenta con él.
   b) Un granjero prueba con 100 de sus gallinas si un nuevo tipo de alimentación mejora el rendimiento.

6. En una asignatura del colegio la probabilidad de que te saquen a la pizarra en cada clase es del 10%. A lo largo del año tienes 100 clases de esa asignatura. ¿Cuál es la probabilidad de tener que salir a la pizarra más de 15 veces? Se vuelve a aplicar el **Teorema Central del Límite**

# CAPÍTULO 7: ESTIMACIÓN CON INTERVALOS DE CONFIANZA

## INTERVALO DE CONFIANZA

Intervalo de valores que tiene designada una probabilidad que incluya el valor real del parámetro de población.[1]

Se llama intervalo de confianza en estadística a un par de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un parámetro poblacional. La probabilidad de éxito en la estimación se representa por 1 - α y se denomina nivel de confianza. En estas circunstancias, α es el llamado error aleatorio o nivel de significación, esto es, una medida de las posibilidades de fallar en la estimación mediante tal intervalo.[2]

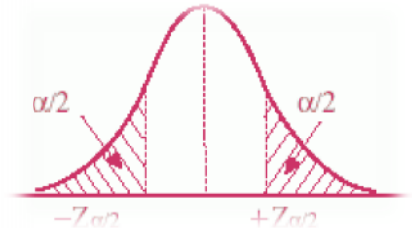
## Cálculo de intervalos de confianza.

En una estimación paramétrica, el intervalo de confianza [a, b] debe contener en su interior a la media de la población m con una probabilidad igual a 1 - a, expresión que se conoce como **nivel de confianza**. Es decir:

$$P\ (a \leq \mu \leq b) = 1 - \alpha$$

En una **distribución muestral de las medias** con media poblacional m, desviación típica poblacional s, tamaño de la muestra n, media muestral $\mu_{\bar{x}}$ e intervalo de confianza predeterminado 1 - a (expresado en porcentaje; por ejemplo, 95%), es posible calcular el intervalo de confianza a partir de la expresión:

$$P\ (\sigma_{\bar{x}} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \sigma_{\bar{x}} + z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 1 - \alpha$$




[1] Richard I. Levin & David S. Rubin
[2] Rius Díaz, Francisca

En una **distribución muestral de las proporciones** de tipo N (p, $\sqrt{pq/n}$), puede determinarse el intervalo de confianza, para el cual existe una proporción p de elementos que poseen una cierta característica, a partir de una muestra representativa, donde la proporción es p', por medio de la siguiente expresión:

$$P\,(p' - z_{\alpha/2} \cdot \sqrt{pq}/n \leq p \leq p' + z_{\alpha/2} \cdot \sqrt{pq}/n) = 1 - \alpha$$

# Intervalo de confianza para la media de una población.

De una población de media μ y desviación típica σ se pueden tomar muestras de *n* elementos. Cada una de estas muestras tiene a su vez una media ($\tilde{x}$). Se puede demostrar que la media de todas las medias muestrales coincide con la media poblacional: $\mu_{\bar{x}} = \mu$

Pero además, si el tamaño de las muestras es lo suficientemente grande, la distribución de medias muestrales es, prácticamente, una distribución normal (o gaussiana) con media μ y una desviación típica dada por la siguiente expresión: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Esto se representa como sigue: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ .

Si estandarizamos, se sigue que: $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1)$

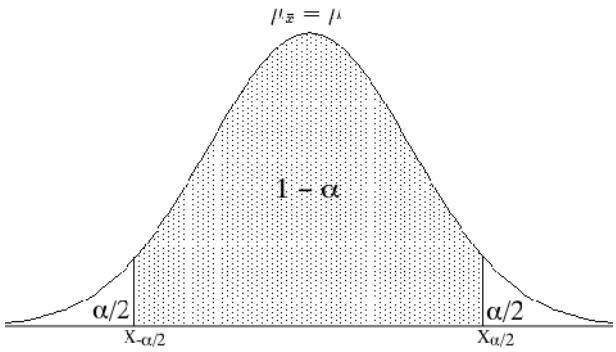
En una distribución $Z \sim N(0, 1)$ puede calcularse fácilmente un intervalo dentro del cual "caigan" un determinado porcentaje de las observaciones, esto es, es sencillo hallar $z_1$ y $z_2$ tales que $P[z_1 \leq z \leq z_2] = 1 - \alpha$, donde $(1 - \alpha)\cdot 100$ es el porcentaje deseado (*véase el uso de las tablas en una distribución normal*).

Se desea obtener una expresión tal que $P\left[\mu_1 \leq \mu \leq \mu_2\right] = 1 - \alpha$

En esta distribución normal de medias se puede calcular el intervalo de confianza

donde se encontrará la media poblacional si sólo se conoce una media muestral (
), con una confianza determinada. Habitualmente se manejan valores de confianza del 95% y 99%. A este valor se le llamará 1 − α (debido a que α es el error que se cometerá, un término opuesto).

Para ello se necesita calcular el punto $X_{\alpha/2}$ —o mejor dicho su versión estandarizada $Z_{\alpha/2}$— junto con su "opuesto en la distribución" $X_{-\alpha/2}$. Estos puntos delimitan la probabilidad para el intervalo, como se muestra en la siguiente imagen:




Dicho punto es el número tal que:

$$\mathbb{P}[\bar{x} \geq X_{\alpha/2}] = \mathbb{P}[z \geq z_{\alpha/2}] = \alpha/2$$

Y en la versión estandarizada se cumple que: $z_{-\alpha/2} = -z_{\alpha/2}$

Así: $$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right] = 1-\alpha$$

Haciendo operaciones es posible despejar μ para obtener el intervalo:

$$\mathbb{P}\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Resultando el intervalo de confianza:

$$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

Si σ no es conocida y n es grande (habitualmente se toma $n \geq 30$):

$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$, donde s es la desviación típica de una muestra.

Aproximaciones para el valor $z_{\alpha/2}$ para los niveles de confianza estándar son 1,96 para 1 − α = 95% y 2,576 para 1 − α = 99%.

# Estimación por intervalos de confianza para la media.

Sea desconocida la media poblacional de una cierta variable que deseamos estudiar, sacamos una muestra y se trata de obtener un intervalo (L1,L2) de forma que tengamos una probabilidad alta (1-alfa)% de que la media poblacional esté en ese intervalo. El **nivel de confianza del intervalo** (1-alfa)% lo fijamos nosotros., se suele trabajar con 95% y a veces con 99% o el 90%; es decir, con probabilidad 0.05, 0.01 o 0.1.

Si se cumple una de las siguientes hipótesis:

- El tamaño de la muestra es mayor de 30 y la variable sigue un modelo normal.
- El tamaño de la muestra es mayor de 100.

El intervalo de confianza para la media poblacional es:

$$\tilde{x} \pm z\frac{s}{\sqrt{n}}$$

Donde z es el valor que en la distribución N(0,1) deja a su derecha un área de alfa/2, $\tilde{x}$ es la media en la muestra, s la cuasidesviación típica (raíz cuadrada de la cuasivarianza) o la desviación típica y n el tamaño de la muestra.

## Intervalo de confianza para una proporción.

El intervalo de confianza para estimar una proporción *p*, conocida una proporción muestral $p_n$ de una muestra de tamaño *n*, a un nivel de confianza del (1-α)·100% es:

$$Pn \pm z_{\alpha/2}\sqrt{\frac{Pn(1-Pn)}{n}}$$

En la demostración de estas fórmulas están involucrados el Teorema Central del Límite y la aproximación de una binomial por una normal.

## Estimación por intervalos de confianza para la proporción.

Sea p desconocida la proporción de elementos en la población pertenecientes a una categoría C, sacamos una muestra y se trata de obtener un intervalo (L1,L2) de forma que tengamos una probabilidad alta (1-alfa)% de que la proporción esté en ese intervalo.

Si se cumple una de las siguientes hipótesis, y que habrá de comprobarlas en todos los problemas son:

- $n.\hat{p} < 5$
- $n.(1-\hat{p}) < 5$

En estas condiciones se obtienen los siguientes intervalos según el tamaño de la muestra:

- El tamaño de la muestra es mayor de 30 y menor o igual de 100.

$$\hat{p} \pm z\sqrt{\frac{1}{4n}}$$

- El tamaño de la muestra es mayor de 100.

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Donde $\hat{p} = \frac{n\ elementos\ de\ la\ muestras\ que\ pertenecen\ a\ C}{tamaño\ de\ la\ muestra}$, z es el valor que en la distribución N(0,1) deja a su derecha un área de alfa/2, n el tamaño de la muestra.

# Intervalo de confianza para la media poblacional - Muestras grandes

Uno de los usos más comunes de los intervalos de confianza es estimar la media poblacional. Un fabricante puede querer estimar la producción mensual promedio de su planta: un representante de mercadeo puede interesarse en a reducción en las ventas semanales promedio: el jefe financiero de una firma, que aparece entre las 500 mejores firmas en la revista *Fortune.* Puede querer estimar los rendimientos trimestrales promedio que se tuvieron en operaciones corporativas. El número de circunstancias que se encuentran comúnmente en el mundo de los ne$^{g}$ocios y que requiere de una estimación de la media poblacional es casi ilimitado.

Se debe recordar que el intervalo se forma utilizando la media muestra¡ como una estimación puntual para el cual se adiciona y se resta un cierto valor para obtener los límites superior e inferior del finten alo de confianza, respectivamente. Por tanto el intervalo es

| |
|---|
| *Intervalo de confianza para estimar y cuando u es conocido* |
| *I-C. para estimar* = $\mu = \tilde{x} \pm Z\sigma_x$ |

Cuánto debe sumarse y restarse, depende en parte del nivel de confianza deseado, estipulado por el valor de Z en la fórmula (7.1). $^{Un}$ nivel de confianza del 95% requiere un valor de Z de 1.96 (0.9512 = 0.4750). El área de 0.4750 corresponde a un valor de Z de 1.96.

***Ejemplo:***

Consideremos el caso de un promotor inmobiliario quien intenta construir un gran centro comercial. Puede estimaren el área el ingreso promedio por familia como indicador de las ventas esperadas. Una muestra de *n* = 100 familias da una media de -Y= US$35.500. Se asume que la desviación estándar poblacional es a= US$7.200. Dado que $\sigma_x = \frac{\sigma}{\sqrt{n}}$ se estima un intervalo del 95% como

$$I.C\ para\ estimar\ \mu = 35.500 \pm (1{,}96)\frac{7.200}{\sqrt{100}}$$

$$= 34.088{,}80 \leq \mu \leq 36.911{,}20$$

## A. Interpretación de un intervalo de confianza.

El promotor puede interpretar los resultados de su intervalo de confianza de dos formas. La primera, y la más común, establece que el promotor tiene un -95% de confianza en que la media poblacional real desconocida esté entre US$34,088,80 y US$36,911.20". Aunque el valor real para la media poblacional sigue siendo desconocido, el promotor tiene un 95% de confianza en que esté entre estos dos valores.

La segunda interpretación reconoce que se pueden desarrollar muchos intervalos de confianza diferentes. Otra muestra probablemente produciría una inedia muestra¡ diferente debido al error de muestreo. Con una X diferente, el intervalo tendría límite superior e inferior distintos. Por tanto, la segunda interpretación establece que si se construyen todos los C ¸intervalos de confianza, el 95% de ellos contendrá la inedia poblacional desconocida.

Si una segunda muestra da una inedia de US$35,600 en lugar de US$35,500, el intervalo es

$$I.C\ para\ estimar\ \mu = 35.600 \pm (1{,}96)\frac{7.200}{\sqrt{100}}$$

$$= 34.188{,}80 \leq \mu \leq 37.011{,}20$$

El promotor puede estar un 95% seguro de que la media poblacional está comprendida entre US$34,188.80 y US$37,011.20. Si todos los intervalos posibles se construyeran con base en todas las medias muestrales diferentes, el 95% de ellas contendría la media poblacional desconocida.

Esto por supuesto significa que el 5% de todos los intervalos estaría errado - no contendrían la media poblacional-. Este 5%, hallado como (1 - coeficiente de confianza). es denominado el valor alfa y representa la probabilidad de error. El valor alfa es la probabilidad de que cualquier intervalo dado no contenga la media poblacional.

Valor alfa: Es la probabilidad de error o la probabilidad de que un intervalo dado no contenga la media poblacional desconocida.

## B. Intervalo de confianza cuando a es desconocida.

La fórmula (7.1) requiere la suposición improbable que la desviación estándar poblacional aes conocida. En el evento probable que a sea desconocida, la desviación estándar de la muestra debe substituirse:

| |
|---|
| *Intervalo de confianza para estimar cuando a es desconocida* |
| *I-C. para estimar* = $\mu = \tilde{x} \pm Zs_x$ |

en donde $s_x = s/\sqrt{n}$

***Ejemplos:***

Gerry Gerber, CPA, acaba de registrar las declaraciones de impuestos de sus clientes. Desea estimar la cantidad promedio que deben al Servicio de Renta Interna. De los 50 clientes que seleccionó en su muestra, la cantidad promedio que se adeudaba era de US$652.68. Ya que la desviación estándar de todos sus clientes ctes desconocida, Gerber debe estimar <Ycon la desviación estándar de la muestra de s = US$217.43. Si se desea un nivel del 99% de confianza, el valor de Z apropiado es 2.58 (0.99/2 = 0.4950). De la tabla Z, un área de 0.4950 revela que Z= 2.58. Utilizando la fórmula (7.2)

$$\text{I.C para estimar } \mu = \bar{X} \pm Zs_x$$
$$= 652{,}68 \pm 2{,}58\frac{217{,}43}{\sqrt{50}}$$
$$= 573{,}35 \leq \mu \leq 732{,}01$$

El señor Gerber puede tener un 99% de confianza en que la cantidad promedio que deben todos sus clientes al Servicio de Renta Interna (SRI) está entre US$573.35 y US$732.01.

¿Qué pasaría a este intervalo si el señor Gerber estuviera dispuesto a aceptar un nivel de confianza del 95%7 Con un valor de Z de 1.96, el intervalo sería

$$= 652{,}68 \pm 1{,}96\frac{217{,}43}{\sqrt{50}}$$
$$= 592{,}41 \leq \mu \leq 712{,}96$$

Los resultados son tanto buenos como malos. Las buenas noticias son que el intervalo del 95% es más estrecho y ofrece mayor precisión. Un intervalo amplio no es especialmente útil. Revelaría muy poco si el profesor le pidiera que la media del siguiente examen estuviera entre el 0 y el 100%. Entre más estrecho sea el intervalo, más significativo es.

Las malas noticias son que el señor Gerber ahora está el 95% seguro de que el intervalo contiene en realidad ,u. Aunque el intervalo es más preciso (niás estrecho), la probabilidad de que contenga M se ha reducido del 99 al 95%. El señor Gerber tuvo que abandonar algo de confianza y ganar más precisión.

## Intervalo de confianza para la media en el caso de muestras pequeñas - la distribución t.

En todos' los ejemplos anteriores, el tamaño de la muestra era mayor *(ti?* 30). Sin embargo, no siempre puede ser posible obtener por lo menos 30 observaciones. Para una compañía de seguros que prueba la resistencia al impacto de lo., autos, destruir a propósito 30 vehículos de lujo puede volverse un poco costoso. Un investigador médico que prueba una nueva medicina puede no encontrar 30 personas dispuestas a actuar como conejillo de indias. En muchos casos una muestra grande no es posible.

Cuando debe tomarse una muestra pequeña, la distribución normal puede no

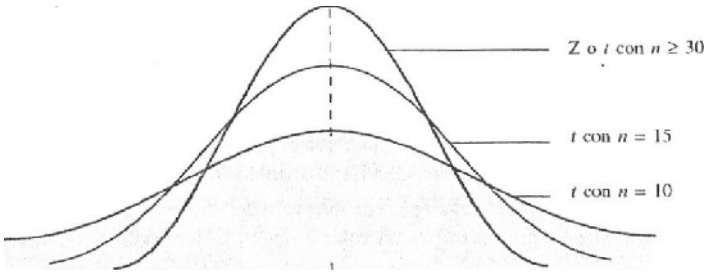
aplicarse. El teorema del límite central asegura normalidad en el proceso de muestreo sólo si la muestra es grande. Cuando se utiliza una muestra pequeña, puede ser necesaria una distribución alternativa, la distribución *t* Student (o simplemente la distribución *t).* Específicamente, la distribución *t* se utiliza cuando se cumplen las tres condiciones: (1) la muestra es pequeña, (2) aes desconocida. y (3) la población es normal o casi normal. Si aes conocida, la distribución Z se usa inclusive si la muestra es pequeña. Además, si no puede asumirse tina población normal, se aumenta el tamaño de la muestra para utilizar la distribución Z y de no ser posible se debe confiar en las pruebas *no paraniétricas.*

La distribución *t* Student fue desarrollada en 1908 por William S. Gosset (1876 - 1937), quien trabajó como experto cervecero para Guinnes Breweries en Dublín, Irlanda. Guiness no permitía que sus empleados publicaran su investigación, de manera que Gosset (a quien le gustaba 'Jugar con los números para relajarse") informó por primera vez sobre su distribución *i,* aunque publicó bajo el seudónimo de "Student" para proteger su trabajo.

Al igual que la distribución *Z,* la distribución *t* tiene una media de cero, es simétrica con especto a la media y oscila entre - ∞y + ∞. Sin embargo, mientras que la distribución *Z* tiene una varianza de $o^2 = 1$, la varianza de la distribución *t* es mayor que 1. Por tanto, es más plana y más dispersa que la distribución Z. La varianza para la distribución t es

| |
|---|
| *Varianza de la distribución t* $\sigma^2 = \frac{n-1}{n-3}$ |

En realidad la distribución *t* es una familia de distribuciones cada una con su propia varianza. La varianza depende de los *grados de libertad* (g.l.). Definidos como el número de observaciones que se pueden escoger libremente. Es el número de observaciones menos el número de restricciones impuestas sobre tales observaciones, en donde una restricción es algún valor que tales observaciones deben poseer. Se asume que se tienen *ti* = 4 observaciones que deben producir una media de 10. La media de 10 sirve como una restricción y hay *ir* - 1 = 3 grados de libertad. Por tanto, se pueden escoger tres observaciones cualesquiera: por ejemplo se puede escoger 8, 9 y 11. Después de que se seleccionan estos tres valores, ya no hay libertad para escoger la última observación. El cuarto valor *debe* ser 12 si se quiere tener un promedio de 10. Vale la pena destacaren la figura 7.3 que a medida que *ti* aumenta, la distribución *t* se aproxima a la distribución Z. Es por esto que se puede utilizar la distribución Z cuando *n* > 30.




***Grados de libertad:*** El número de observaciones menos el número de restricciones impuestas sobre tales observaciones.

Como se verá en breve, para todo conjunto de condiciones dadas la distribución *t*

producirá un intervalo más amplio que la distribución Z, si ésta se utilizara. Este ancho adicional es necesario debido a que se pierde algo de precisión porque a es desconocida y debe estimarse.

El estadístico *t* se calcula en gran parte como el estadístico Z.

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

Reescribiendo algebraicamente para expresarlo como un intervalo de confianza para estimar *p,* se tiene que

*Intervalo de confianza para estimar la media poblacional - muestras pequeñas*

*I.C. para estimar* $\mu = \bar{X} \pm (t)(s_{\bar{x}}) = \bar{X} \pm t\frac{s}{\sqrt{n}}$

El valor apropiado de *t* puede hallarse de la tabla F en el apéndice III. Para ilustrar, se asume que se desea un intervalo de confianza del 95% y se tiene una muestra de 20 observaciones. Debido a que *n* = 20, los grados de libertad son g.l. = *n* -1= 19. Bajando por la primera columna en la tabla F bajo g.l. hasta 19. Se mueve a través de dicha fila hacia la columna encabezada por un nivel de confianza de 0.95 para las pruebas de dos colas. (Se ignoran las dos filas referentes a las pruebas de una cola. Estas se tratarán en el capítulo 8). La entrada resultante de 2.093 es el valor *t* apropiado para un intervalo de confianza del 95% con un tamaño maestral de 20 (g.l. = 19).

***Ejemplo:***

Consideremos el si$^{g}$uiente problema tomado de *The lWill Sitect* Une empresa de cons$_{t}$rucción fue culpada *de* inflar los comprobantes que registra para los contratos de construcción con el sobiemo federal. El contrato estableció que un cieno tipo de trabajo debería promediar USS 1,150. Por motivos de tiempo, los directivos de sólo 12 agencias del gobienio fueron llamados a dar testimonio ante la corte respecto a los comprobantes de la empresa. Si se descubrió a partir del testimonio una media de USS1,275 y tina desviación estándar de USS235, ¿un iiite~,alo de confianza del 95$^{q}$ apoyaría el caso legal de la empresa? Se asume que los montos de los comprobantes son normales.

Un nivel de confianza del 95% con g.l. = 12 -1 = 11 resulta de la tabla F un valor *t* de 2.201. Entonces

$$\text{I.C para calcular } \mu = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

$$= 1.275 \pm (2{,}201)\frac{235}{\sqrt{12}}$$

$$= 1.275 \pm 149{,}31$$

$$1.125{,}69 \leq \mu \leq 1.424{,}31$$

La corte puede tener un 95% de confianza en que el promedio de todos los comprobantes está entre US$ 1.1255 y US$1.424. Este intervalo contiene los US$1,150 acordados, fortaleciendo la defensa de la empresa.

Vale la pena destacar que el valor *t* para un intervalo del 95% es 2.201 (dado g.l. = 11),

mientras que un intervalo del 95% de una muestra grande utiliza un valor Z de 1.96. El intervalo con base en un valor *t* es, por tanto, más amplio.

## Intervalo de confianza para la proporción poblacional.

Las decisiones dependen con frecuencia de parámetros que son binarios, parámetros con sólo dos posibles categorías dentro de las cuales pueden clasificarse las respuestas. En este evento, el parámetro de interés es la *proporción* poblacional. Una empresa puede desear saber qué proporción de sus clientes paga a crédito en oposición a quienes utilizan efectivo. Las corporaciones con frecuencia están interesadas en qué porcentaje de sus productos son defectuosos en oposición al porcentaje que no es defectuoso, o qué proporción de sus empleados renuncian después de un año en contraste con la proporción que no renuncia después de un año. En cada uno de estos casos, existen sólo dos posibles resultados. Por tanto. la preocupación se concentra en la proporción de respuestas que queda dentro de uno de estos dos resultados.

En el capítulo anterior se encontró que si *n n* y *n (* 1 - 7r ) son ambos mayores que 5, la distribución de las proporciones muestrales será normal y la distribución muestra) de la proporción muestra] tendrá una media igual i la proporción poblacional *z* y un error estándar de

*El error estándar de la distribución muestral de las proporciones muestrales*

$$\sigma_{\boldsymbol{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Sin embargo, la fórmula contiene $\pi$, el parámetro que se desea estimar. Por tanto, la proporción maestral *p* se utiliza como estimador de $\pi$.

La fórmula puede replantearse como

*Estimación del error estándar de la distribución de las proporciones muestrales*

$$s_{\boldsymbol{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Intervalo de confianza para estimar la proporción poblacional

$$I.C\ para\ estimar\ \pi = p \pm Zs_p$$

**Ejemplo:**

El gerente de tina estación de televisión debe determinar en la ciudad qué porcentaje de casas tiene más de un televisor. Una muestra aleatoria de 500 casas revela que 275 tienen dos o más televisores. ¿Cuál es el intervalo de confianza del 90% para estimar la proporción de todas las casas que tienen (toso más televisores'? Dados estos datos. *p = 275/1500 =* 0.55, y

$$s_{\boldsymbol{p}} = \sqrt{\frac{(0{,}55)(0{,}45)}{500}}$$

La tabla E da un valor de Z de 1.65 para un intervalo de confianza del 90%

$$\text{I.C. para estimar } \pi = 0{,}55 \pm (1{,}65)(0{,}022)$$
$$= 0{,}55 \pm 0{,}036$$
$$0{,}514 \leq \pi \leq 0{,}586$$

El gerente puede tener un 90% de confianza que entre el 51.4% y el 58.6% de las casas de la ciudad tienen más de un televisor.

## Control del ancho de un intervalo.

Como se expresó anteriormente, es preferible un intervalo más estrecho debido a la precisión adicional que proporciona. Hay dos métodos principales para lograr un intervalo más preciso: (1) reducir el nivel de confianza y (2) incrementar el tamaño de la muestra.

## A. Reducción del nivel de confianza.

Ya se ha visto, en el intento del señor Gerber por estimarla declaración de impuestos promedio de sus clientes, que un incremento en la precisión puede obtenerse aceptando un nivel inferior de confianza. Su intervalo de confianza del 99$^{9}$1 oscilaba entre US$573 y US$732, mientras que el intervalo del 95% era más estrecho de US$594 a US$712. Esto resultó del hecho que el intervalo de confianza del 99% requirió un valor de Z de 2.58 en lugar de 1.96 que utilizo el intervalo del 95%.

Sin embargo, había un costo involucrado en lograr esta precisión mayor: el nivel de confianza bajó a 95%, produciendo un 5% de probabilidad de error en lugar del 1% relacionado con el intervalo de confianza del 99%. ¿Existe al$_{g}$una manea en la que se pueda reducir el intervalo sin sufrir una pérdida de confianza? Sí, incrementando el tamaño maestral.

## B. Incremento del tamaño muestral.

Incrementando el tamaño muestral se puede reducir el error estándar $\sigma/\sqrt{n}$. Si *el* tamaño muestral del señor Gerber se incrementa a 80. El intervalo del 99% presenta un grado de precisión similar al intervalo más estrecho del 95%. Sin ninguna pérdida de confianza. Con n = 80. El intervalo del 99% es

$$I.C. para\ estimar\ \mu = 652{,}68 \pm 2{,}58\frac{217{,}43}{\sqrt{80}}$$
$$589{,}96 \leq \mu \leq 715{,}39$$

Esto está muy cercano al intercalo más preciso del 95% de USS592.41 a USS712,96,

pero mantiene un nivel de confianza del 99%.

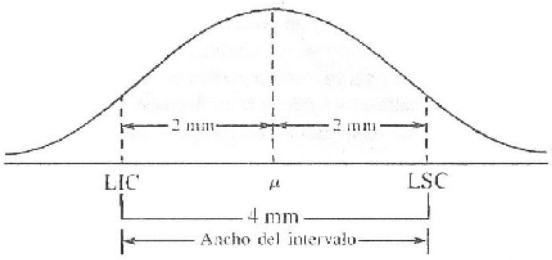
Infortunadamente, esta ventaja no se gana sin un precio. El tamaño avis $_{g}$rande de la ¡nuestra significa más tiempo y más; dinero que deben gastarse al recolectar y manejar los datos. De nuevo, debe tomare una decisión. Se vuele una decisión gerencia) respecto a qué método tomar.

# Determinación del tamaño apropiado de la muestra.

El tamaño de la muestra juega un papel importante al determinarla probabilidad de error así como en la precisión de la estimación. Una vez que se ha seleccionado el nivel de confianza, dos factores importantes influyen en el tamaño muestral: ( 1 ) la varianza de la población $\sigma^2$y (2) el tamaño del *error tolerable* que el investigador está dispuesto a aceptar. Mientras que el primer factor está más allá del control del investigador (no hay nada que se pueda hacer sobre la varianza de la población). Sí es posible limitar el tamaño del error.

El tamaño del error que un investigador puede tolerar depende de qué tan crítico es el trabajo. Al$^g$unas tareas extremadamente delicadas requieren de resultados exactos: los procedimientos médicos vitales de los cuales dependen vidas humanas, o la producción de piezas de tina máquina que deba cumplir medidas precisas, pueden tolerar sólo un pequeño error. En otros casos, los errores más grandes pueden tener consecuencias menos graves.

Se supone que en la fabricación (le una pieza para los reproductores de discos compactos (CD), un error de 2 milímetros (mm) en el diámetro no causaría problema alguno: sin embar$^g$o. todo error superior a 2 mm resultará en un reproductor defectuoso. Si una pieza puede variar por encima y por debajo de algún diámetro deseado en 2 mm, se permite un intervalo de 4 mm. Todo intervalo dado es dos veces el error tolerable. Ver la figura para observar una ilustración




# A. Tamaño de la muestra para estimar $\mu$.

Vale la pena recordar que la desviación normal Z puede expresarse como

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

Esto puede reescribirse algebraicamente como

*Tamaño muestral para intervalos de la media poblacional*

$$n = \frac{Z^2\sigma^2}{(\bar{X} - \mu)^2}$$

en donde la diferencia entre la media maestral y la media poblacional (A^ *p)* es el error. En el ejemplo anterior de los reproductores de CD's, con $^{un}$ error tolerable de 2 aun, la fórmula (7.9) se escribiría como

$$n = \frac{Z^2\sigma^2}{(2)^2}$$

El valor de Z depende del nivel de confianza requerido. Esto deja por determinar sólo 6' para calcular el tamaño muestra) apropiado. En el evento probable que $6^2$ sea desconocido, puede estimarse mediante la desviación estándar maestral *s,* utilizando una muestra piloto de cualquier tamaño razonable ( $n \geq 30$). La varianza calculada de esta muestra preliminar puede entonces utilizarse en la fórmula (7.9).

**Ejemplo:**

Por ejemplo, se asume que el fabricante de los reproductores de discos compactos desea construir un intervalo del 95% para el tamaño promedio de la pieza. Una muestra piloto ha revelado una desviación estándar de 6 mm. ¿Qué tan grande debería ser la muestra? Un intervalo del 95% da un valor de Z de 1.96. Por tanto

$$n = \frac{(1{,}96)^2(6)^2}{(2)^2} = 34{,}5\ o\ 35$$

El fabricante debería seleccionar una muestra de 35 piezas. De esta muestra, un intervalo de 95% podría construirse para el tamaño promedio. El intervalo tendría un error no superior a 2 mm.

## B. Tamaño de la muestra para estimar $\pi$

$$Z = \frac{p - \pi}{\sigma_p}$$

en donde

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Se puede reescribir esto para producir una expresión para el tamaño maestral

*Tamaño muestral para intervalos para la proporción poblacional*

$$n = \frac{Z^2\pi(1 - \pi)}{(p - \pi)^2}$$

en donde *(p - $\pi$)* es la diferencia entre la proporción muestral y la proporción poblacional, y por tanto es el error.

La fórmula requiere el valor de $\pi$. Sin embargo, $\pi$: es el parámetro que se desea estimar y es desconocido. Este problema puede tratarse en una de las dos maneras- Se podría tornar una muestra piloto para obtener un valor preliminar para $\pi$, tal y como se hizo al determinar el tamaño maestral apropiado para la media. O se puede determinar que $\pi$ = 0.5, para efectos de determinar el tamaño maestral. Frecuentemente se prefiere este método porque es muy "seguro" o conservador – garantizará el tamaño muestral más grande posible, dado cualquier nivel de confianza y error deseados– Esta muestra más grande resulta del hecho que el numerador de la

fórmula , el cual contiene $\pi(1 - \pi)$ es máximo (por tanto, *n* se maximizará) cuando $\pi = 1 - \pi = 0.5$. No existe valor distinto a 0.5 que pueda asignarse a raque haga más grande $\pi(1 - \pi)$. Si $\pi = 0.5$, entonces $\pi(1 - \pi) = 0.25$. Todo valor distinto a 0.5 resultaría en n $(1 - \pi) < 0.25$. Por tanto, *n* sería más pequeña.

**Ejemplo:**

Wally Simpleton está postulado para gobernador. Él desea estimar dentro de 1 punto porcentual la proporción de personas que votarán por él. También desea tener el 95% de confianza en sus hallazgos. ¿Qué tan grande debería ser el tamaño maestral?

$$n = \frac{(1{,}96)^2(0{,}5)(0{,}5)}{(0{,}01)^2}$$

$$= 9.604 \; votantes$$

Una muestra de 9,604 votantes permitirá a Wally estimar $\pi$ con un error de un 1% y un nivel de confianza del 95%.

# CAPÍTULO 8: PRUEBA DE HIPÓTESIS

## Hipótesis.

Es el enunciado acerca de una población elaborada con el propósito de ponerse a prueba y para verificar si la afirmación es razonable de los datos utilizados.

**Ejemplo:**

- El 45.5 % de los estudiantes de comunicación y medios, fuma.
- Un dentista reclama que el 5% de sus pacientes sufren enfermedades en las encías.
- Prueba de hipótesis: procedimiento basado en la evidencia muestral y en la teoría de probabilidad que se emplea para determinar si la hipótesis es un enunciado razonable y no debe rechazarse o si no es razonable y debe ser rechazado.

## Prueba de hipótesis:

Es un procedimiento basado en la evidencia muestral (una o más poblaciones) y la teoría de probabilidad; se emplea para determinar si la hipótesis es una afirmación razonable. Nunca se sabe con absoluta certeza la verdad o falsedad de una hipótesis estadística, a no ser que se examine la población entera.

Se debe tomar una muestra aleatoria de la población de interés y se utilizan los datos que contiene tal muestra para proporcionar evidencia que confirme o no la hipótesis. La evidencia de la muestra que es un constante con la hipótesis planteada conduce a un rechazo de la misma mientras que la evidencia que apoya la hipótesis conduce a su aceptación.

La prueba de hipótesis cuantifica el proceso de toma de decisiones.

## Objetivo de la prueba de hipótesis.

El propósito de la prueba de hipótesis no es cuestionar el valor calculado del estadístico (muestral), sino hacer un juicio con respecto a la **diferencia** entre estadístico de muestra y un valor planteado del parámetro.

## Pasos de la prueba de hipótesis.

# CONCEPTOS BÁSICOS PARA EL PROCEDIMIENTO DE PRUEBAS DE HIPOTESIS

## Tipos de hipótesis.

### ♣ Hipótesis Estadística:

Al intentar alcanzar una decisión, es útil hacer hipótesis (o conjeturas) sobre la población aplicada.

Tales hipótesis, que pueden ser o no ciertas, se llaman hipótesis estadísticas.

Son, en general, enunciados acerca de las distribuciones de probabilidad de las poblaciones. Por cada tipo de prueba de hipótesis se puede calcular una prueba estadística apropiada.

### ♣ La hipótesis nula (Ho).

Se refiere siempre a un valor especificado del parámetro de población, no a una estadística de muestra. La letra H significa hipótesis y el subíndice cero no hay diferencia. Por lo general hay un "no" en la hipótesis nula que indica que "no hay cambio" Podemos rechazar o aceptar Ho.

Esta es una afirmación que no se rechaza a menos que los datos maestrales proporcionen evidencia convincente de que es falsa. El planteamiento de la hipótesis nula siempre contiene un signo de igualdad con respecto al valor especificado del parámetro. Suele llevar los signos igual, mayor o igual y menor o igual.

**Ejemplo 1:** si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea p = 0,5, donde p es la probabilidad de cara). Analógicamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que no hay diferencia entre ellos (o sea. Que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la misma población).

**Ejemplo 2:** supongamos que un investigador cree que si un grupo de jóvenes se somete a un entrenamiento intensivo de natación, éstos serán mejores nadadores que aquellos que no recibieron entrenamiento. Para demostrar su hipótesis toma al azar una muestra de jóvenes, y también al azar los distribuye en dos grupos: uno que llamaremos experimental, el cual recibirá entrenamiento, y otro que no recibirá entrenamiento alguno, al que llamaremos control. La hipótesis nula señalará que no hay diferencia en el desempeño de la natación entre el grupo de jóvenes que recibió el entrenamiento y el que no lo recibió.

## ♣ La hipótesis alternativa (H1).

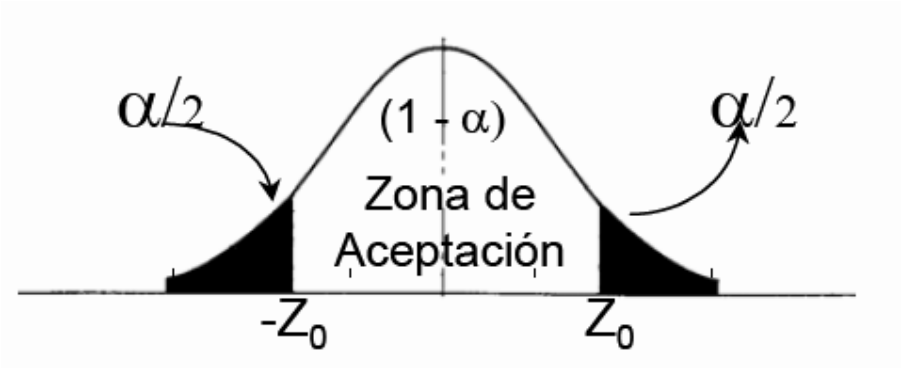
Es cualquier hipótesis que difiera de la hipótesis nula. Es una afirmación que se acepta si los datos maestrales proporcionan evidencia suficiente de que la hipótesis nula es falsa. Se le conoce también como la hipótesis de investigación. El planteamiento de la hipótesis alternativa nunca contiene un signo de igualdad con respecto al valor especificado del parámetro. Suele llevar los signos distintos, mayor y menor.Toda hipótesis que difiere de una dada se llamará una hipótesis alternativa.

**Ejemplo 1:** Si una hipótesis es p = 0,5, hipótesis alternativa podrían ser p = 0,7, p " 0,5 ó p > 0,5.

## Nivel de significancia.

Probabilidad de rechazar la hipótesis nula cuando es verdadera. Se le denota mediante la letra griega α, también es denominada como nivel de riesgo, este término es más adecuado ya que se corre el riesgo de rechazar la hipótesis nula, cuando en realidad es verdadera. Este nivel esta bajo el control de la persona que realiza la prueba.

Si suponemos que la hipótesis planteada es verdadera, entonces, el nivel de significación indicará la probabilidad de no aceptarla, es decir, estén **fuera** de área de aceptación. El **nivel de confianza** (1-α), indica la probabilidad de aceptar la hipótesis planteada, cuando es verdadera en la población.




La distribución de muestreo de la estadística de prueba se divide en dos regiones, una región de rechazo (conocida como región crítica) y una región de no rechazo (aceptación). Si la estadística de prueba cae dentro de la región de aceptación, no se puede rechazar la hipótesis nula.

La región de rechazo puede considerarse como el conjunto de valores de la estadística de prueba que no tienen posibilidad de presentarse si la hipótesis nula es verdadera. Por otro lado, estos valores no son tan improbables de presentarse si la hipótesis nula es falsa. El valor crítico separa la región de no rechazo de la de rechazo.

**Ejemplo** se escoge el nivel de significación 0,05 (ó 5%) al diseñar una regla de decisión, entonces hay unas cinco (05) oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado; Es decir, tenemos un 95% de confianza de que hemos adoptado la decisión correcta. En tal caso decimos que la hipótesis ha sido rechazada al nivel de significación 0,05, lo cual quiere decir que tal hipótesis tiene una probabilidad 0,05 de ser falsa.

## Tipos de errores.

***Error tipo I*** se presenta si la hipótesis nula Ho es rechazada cuando es verdadera y debía ser aceptada. La probabilidad de cometer un error tipo I se denomina con la letra alfa α

***Error tipo II*** se denota con la letra griega β se presenta si la hipótesis nula es aceptada cuando de hecho es falsa y debía ser rechazada.

En la siguiente tabla se muestran las decisiones que pueden tomar el investigador y las consecuencias posibles.

| | Investigador | |
|---|---|---|
| Hipotesis nula | Se acepta Ho | Se rechaza Ho |
| Ho es verdadrea | Decision correcta | Error tipo I |
| Ho es falsa | Error tipo II | Decision orrecta |

Un error de tipo II denotada con la letra griega beta β, depende de la diferencia entre los valores supuesto y real del parámetro de la población. Como es más fácil encontrar diferencias grandes, si la diferencia entre la estadística de muestra y el correspondiente parámetro de población es grande, la probabilidad de cometer un error de tipo II, probablemente sea pequeña.

El estudio y las conclusiones que obtengamos para una población cualquiera, se habrán apoyado exclusivamente en el análisis de una parte de ésta. De la probabilidad con la que estemos dispuestos a asumir estos errores, dependerá, por ejemplo, el tamaño de la muestra requerida. Las contrastaciones se apoyan en que los datos de partida siguen una distribución normal

Existe una relación inversa entre la magnitud de los errores α y β: conforme a aumenta, β disminuye. Esto obliga a establecer con cuidado el valor de a para las pruebas estadísticas. Lo ideal sería establecer α y β. En la práctica se establece el nivel α y para disminuir el Error β se incrementa el número de observaciones en la muestra, pues así se acortan los limites de confianza respecto a la hipótesis planteada.La meta de las pruebas estadísticas es rechazar la hipótesis planteada. En otras palabras, es deseable aumentar cuando ésta es verdadera, o sea, incrementar lo que se llama **poder de la prueba (**1- β**)** La aceptación de la hipótesis planteada debe interpretarse como que la información aleatoria de la muestra disponible no permite detectar la falsedad de esta hipótesis.

# Tipos de prueba.

Cuando estudiamos ambos valores estadísticos es decir, ambos lados de la media lo llamamos prueba de uno y dos extremos o contraste de una y dos colas.

Con frecuencia no obstante, estaremos interesados tan sólo en valores extremos a un lado de la media (o sea, en uno de los extremos de la distribución), tal como sucede

cuando se contrasta la hipótesis de que un proceso es mejor que otro (lo cual no es lo mismo que contrastar si un proceso es mejor o peor que el otro) tales contrastes se llaman unilaterales, o de un extremo. En tales situaciones, la región crítica es una región situada a un lado de la distribución, con área igual al nivel de significación.

## a) Prueba bilateral o de dos extremos.

La hipótesis planteada se formula con la igualdad

Ejemplo

H0 : μ = 200

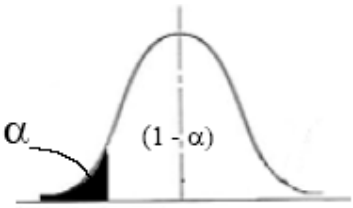
H1 : μ ≠ 200

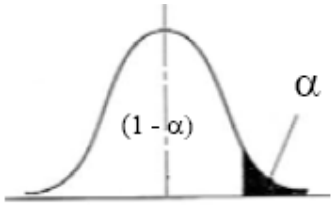
## b) Pruebas unilateral o de un extremo:

La hipótesis planteada se formula con ≥ o ≤

H0 : μ ≥ 200 H0 : μ ≤ 200

H1 : μ < 200 H1 : μ > 200




En las pruebas de hipótesis para la media (μ), cuando se conoce la desviación estándar (σ) poblacional, o cuando el valor de la muestra es grande (30 o más), el valor estadístico de prueba es z y se determina a partir de:

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

El valor estadístico z, para muestra grande y desviación estándar poblacional desconocida se determina por la ecuación:

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

En la prueba para una media poblacional con muestra pequeña y desviación estándar poblacional desconocida se utiliza el valor estadístico t.

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

## Formular la regla de decisión.

Se establece las condiciones específicas en la que se rechaza la hipótesis nula y las condiciones en que no se rechaza la hipótesis nula. La región de rechazo define la ubicación de todos los valores que son tan grandes o tan pequeños, que la probabilidad de que se presenten bajo la suposición de que la hipótesis nula es verdadera, es muy remota

Distribución muestral del valor estadístico z, con prueba de una cola a la derecha

**Valor critico:** Es el punto de división entre la región en la que se rechaza la hipótesis nula y la región en la que no se rechaza la hipótesis nula.

## Tomar una decisión.

En este último paso de la prueba de hipótesis, se calcula el estadístico de prueba, se compara con el valor crítico y se toma la decisión de rechazar o no la hipótesis nula. Tenga presente que en una prueba de hipótesis solo se puede tomar una de dos decisiones: aceptar o rechazar la hipótesis nula. Debe subrayarse que siempre existe la

posibilidad de rechazar la hipótesis nula cuando no debería haberse rechazado (error tipo I). También existe la posibilidad de que la hipótesis nula se acepte cuando debería haberse rechazado (error de tipo II).

**Ejemplo:**

El consumo de un automóvil en km/lts tiene distribución normal con un desvío 0.8 km/lts Con el objeto de estimar el consumo medio de nafta se realizaron 40 pruebas, obteniéndose un rendimiento medio de 12.8 km/litro

Si el fabricante afirmo que su rendimiento medio es de 12.4 km/litro, ¿puede rechazarse esta hipótesis a un nivel de significancia del 5%? Exprese la regla de decisión en términos de la media muestral.

**PASO 1:** Definir juego de Hipótesis

Ho: M = 12.4

Ha: M ≠ 12.4

**PASO 2:** Nivel de Significancia

5% = 0.05

(Debido a que es una prueba de dos colas se divide entre dos)

0.05/2 = 0.025

0.05-0.025 = 0.475 Este es el numero que buscamos en la tabla de Z.

0.475 = 1.96

**PASO 3:** Calcular estadístico de prueba:

Debido a que si conocemos la desviación estándar utilizaremos la siguiente fórmula:

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

$$Z = \frac{12.8 - 12.4}{\frac{.8}{\sqrt{40}}}$$

Z= -1.74

**PASÓ 4:** Formular Regla de decisión:

Rechazaremos Ho si Zc es ≠ al valor crítico.

**PASÓ 5:** Tomar la decisión.

¡SI SE RECHAZA ¡ Debido a que -1.74 es diferente a 1.96

**Ejemplo:**

La fábrica de llantas tiene dos turnos de operarios, turno de día y turno mixto. Se selecciona una muestra aleatoria de 100 llantas producidas por cada turno para ayudar al gerente a sacar conclusiones de cada una de las siguientes preguntas:

1.- Es la duración promedio de las llantas producidas en el turno de día igual a 25000 millas?

2.- ¿Es la duración promedio de las llantas producidas en el turno mixto menor de 25 000 millas?

3.- Se revienta más de un 8% de las llantas producidas por el turno de día antes de las 10 000 millas

# Prueba de hipótesis para la media.

En la fábrica de llantas la hipótesis nula y alternativa para el problema se plantearon como sigue:

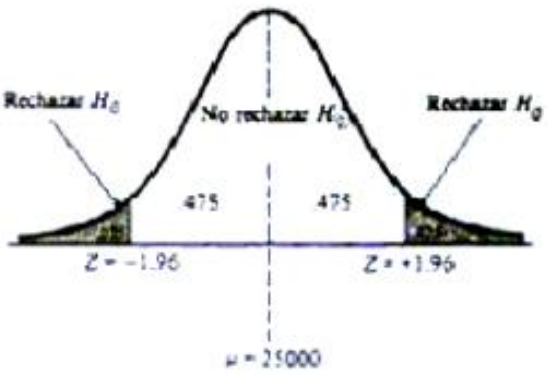
*Ho:* $\mu = 25\,000$
*H1:* $\mu \neq 25\,000$

Si se considera la desviación estándar σ las llantas producidas en el turno de día, entonces, con base en el teorema de limite central, la distribución en el muestreo de la media seguiría la distribución normal, y la prueba estadística que está basada en la diferencia entre la media de la muestra y la media μ hipotética se encontrara como sigue:

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

Si el tamaño de la región α de rechazo se estableciera en 5% entonces se podrían determinar los valores críticos de la distribución. Dado que la región de rechazo está dividida en las dos colas de la distribución, el 5% se divide en dos partes iguales de 2.5%.

Dado que ya se tiene la distribución normal, los valores críticos se pueden expresar en unidades de desviación. Una región de rechazo de 0.25 en cada cola de la distribución normal, da por resultado un área de .475 entre la media hipotética y el valor crítico. Si se busca está área en la distribución normal, se encuentra que los valores críticos que dividen las regiones de rechazo y no rechazo son + 1.96 y - 1.96




Por tanto, la regla para decisión sería:

Rechazar *Ho* si Z > + 1.96
O si Z < - 1.96
De lo contrario, no rechazar *Ho*

## Problemas resueltos.

**Un criador de pollos sabe por experiencia que el peso de los pollos de cinco meses es 4,35 libras. Los pesos siguen una distribución normal. Para tratar de aumentar el peso de dichas aves se le agrega un aditivo al alimento. En una muestra de pollos de cinco meses se obtuvieron los siguientes pesos (en libras).**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4,41 | 4,37 | 4,33 | 4,35 | 4,30 | 4,39 | 4,36 | 4,38 | 4,40 | 4,39 |

En el nivel 0,01, el aditivoa ha aumentado el peso medio de los pollos? Estime el valor de p.

n = 10
u = 4,35
Xmed = 43,68/10 = 4,368

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{0.01036}{10-1}}$$

$$s = \sqrt{0.00115} = 0.0339$$

Planteamiento de hipótesis

$$H_o : u \leq 4.35$$
$$H_1 : u > 4.35$$

| Peso librasX | X-Xmed | (X-Xmed)^2 | X^2 |
|---|---|---|---|
| 4,41 | 0,042 | 0,001764 | 19,4481 |
| 4,37 | 0,002 | 4E-06 | 19,0969 |
| 4,33 | -0,038 | 0,001444 | 18,7489 |
| 4,35 | -0,018 | 0,000324 | 18,9225 |
| 4,3 | -0,068 | 0,004624 | 18,49 |
| 4,39 | 0,022 | 0,000484 | 19,2721 |
| 4,36 | -0,008 | 6,4E-05 | 19,0096 |
| 4,38 | 0,012 | 0,000144 | 19,1844 |
| 4,4 | 0,032 | 0,001024 | 19,36 |
| 4,39 | 0,022 | 0,000484 | 19,2721 |
| 43,68 | | 0,01036 | 190,8046 |
| X med = | 4,368 | | |

a) Prueba de una cola
b) nivel de significancia 0,01
c) Estadístico de prueba

$$t = \frac{\bar{X}-u}{s/\sqrt{n}} = \frac{4.368-4.35}{0.0339/\sqrt{10}} = \frac{0.018}{0.01072} = 1.68$$

Area = 0,4535

d) Plantear la regla de desición:
alfa = 0,01 y gl = n - 1 =10 -1 = 9

Si t > 2,821 Se rechaza Ho y se acepta H1

Tomar la decisión:

Como t ( 1,68) < 2,821se acepta la hipótesis nula y se rechaza H1 y se concluye que el aditivo no aumenta el peso medio de los pollos en un 4,35.

Valor p = 1,68 es 0,4535
p = 0,50 - 0,4535 = 0,046

# EJERCICIOS

**Una empresa que se dedica a hacer en cuestas se queja de que un agente realiza en promedio 53 encuestas por semana. Se ha introducido una forma más moderna de realizar las encuetas y la empresa quiere evaluar su efectividad. Los números de encuestas realizadas en una semana por una muestra aleatoria de agentes son:**

**53 57 50 55 58 54 60 52 59 62 60 60 51 59 56**

**En el nivel de significancia 0,05, puede concluirse que la cantidad media de entrevistas realizadas por los agentes es superior a 53 por semana? Evalúe el valor p.**

u = 53

n = 15

Xmed = 56,4

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{1956}{15-1}} = \sqrt{13971} = 3.73$$

Planteamiento de hipótesis

$$H_o : u \leq 53$$
$$H_1 : u > 53$$

a) Prueba de una cola

b) nivel de significancia 0,05

c) Estadístico de prueba

$$t = \frac{\bar{X}-u}{s/\sqrt{n}} = \frac{56.4-53}{3.73/\sqrt{15}} = \frac{3.4}{0.963} = 3.53$$

| # encuestas | X-Xmed | (X-Xmed)^2 | X^2 |
|---|---|---|---|
| 53 | -3,4 | 11,56 | 2809 |
| 57 | 0,6 | 0,36 | 3249 |
| 50 | -6,4 | 40,96 | 2500 |
| 55 | -1,4 | 1,96 | 3025 |
| 58 | 1,6 | 2,56 | 3364 |
| 54 | -2,4 | 5,76 | 2916 |
| 60 | 3,6 | 12,96 | 3600 |
| 52 | -4,4 | 19,36 | 2704 |
| 59 | 2,6 | 6,76 | 3481 |
| 62 | 5,6 | 31,36 | 3844 |
| 60 | 3,6 | 12,96 | 3600 |
| 60 | 3,6 | 12,96 | 3600 |
| 51 | -5,4 | 29,16 | 2601 |
| 59 | 2,6 | 6,76 | 3481 |
| 56 | -0,4 | 0,16 | 3136 |
| 846 | | 195,6 | 47910 |

Xmed = 846/15 = 56,4

Area = 0,4989

d) Plantear la regla de desición:

alfa = 0,05 y gl = n - 1 =15 -1 = 14

Si t < 1,761 Se rechaza Ho y se acepta H1

Tomar la decisión:

Como t ( 3,53) > 1,761 se rechaza la hipótesis nula y se acepta H1 y se concluye que la cantidad media de entrevistas realizadas por los agentes es mayor a 53 por semana.

Valor p = 1,761 es 0,4989

p = 0,50 - 0,4989 = 0,0011

**Lisa Monnin es directora de presupuesto en la empresa New Process Company, desea comparar los gastos diarios de transporte del equipo de ventas y del personal de cobranza. Recopiló la siguiente información muestral (importe en dólares).**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Ventas ($)** | **131** | **135** | **146** | **165** | **136** | **142** | |
| **Cobranza ($)** | **130** | **102** | **129** | **143** | **149** | **120** | **139** |

**Al nivel de significancia de 0,10, puede concluirse que los gastos medios diarios del equipo de ventas son mayores? cuál es el valor p?**

| ventas | cobranzas | | Prueba de hipótesis |
|---|---|---|---|
| 131 | 130 | | $H_o : u1 \leq u2$ |
| 135 | 102 | | $H_1 : u1 > u2$ |
| 146 | 129 | | |
| 165 | 143 | | |
| 136 | 149 | | a) Es esta una prueba de una o de dos colas |
| 142 | 120 | | Esta es una prueba a una sola cola |
| | 139 | | |
| 142,5 | 130,285714 | promedio | |
| 12,2433656 | 15,7872764 | desviación st | |
| n = 6 | n = 7 | | |
| alfa = 0,10 | | | |

b) Establezca la regla de desición

Si Z > que el valor crítico, se rechaza la hipótesis nula y se acepta H1

c) Calcule el valor del estadístico de prueba

Datos:

| | |
|---|---|
| n1 = 6 | n2 = 7 |
| Prom 1 = 142,5 | Prom 2 = 130,3 |
| s1 = 12,2 | s2 = 15,8 |
| Alfa = 0.10 | |

Grados de libertad = 6+7 - 2 =11 =1,363 de la tabla con 0,10

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$s_p^2 = \frac{(6-1)148.84 + (7-1)249.64}{6+7-2} = \frac{744,2 + 1497.84}{11} = 203.82$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad t = \frac{142,5 - 130,3}{\sqrt{203.82\left(\frac{1}{6} + \frac{1}{7}\right)}} = \frac{12,2}{7,819} = 1,56$$

d) Cuál es la desición respecto a la hipótesis nula

t calculado 1,56 es mayor que 1,363 rechazo la hipóteis nula

Por lo tanto rechazo la hipótesis nula y concluyo que los gastos medios diarios de las ventas realizadas son mayores.

e) Cual es su valor p

El valor se encuentra entre 0,10 y 0,5

**De una población se toma una muestra de 40 observaciones. La media muestral es de 102 y la desviación estándar 5. De otra población se toma una muestra de 50 observaciones. La media muestral es ahora 99 y la desviación estándar es 6. Realice la siguiente prueba de hipótesis usando como nivel de significancia 0,04.**

Ho: u1 = u2
Ho: u1 ≠ u2

**a) Es esta una prueba de una o de dos colas?**

Esta es una prueba de hipótesis de dos colas

**b ) Establezca la regla de decisión**

Si Z > que le valor crítico, se rechaza la hipótesis nula y se acepta la hipótesis alternativa

**c) Calcule el valor del estadístico de prueba**

Si Z > que el valor crítico, se rechaza la hipótesis nula y se acepta H1

| | | | | | | |
|---|---|---|---|---|---|---|
| Datos: | | | | | | |
| n1 = 40 | | n2 = 50 | | | | |
| Prom 1 = 102 | | Prom 2 = 99 | | | | |
| s1 = 5 | | s2 = 6 | | | | |

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad z = \frac{102 - 99}{\sqrt{\frac{(5)^2}{40} + \frac{(6)^2}{50}}} = \frac{3}{\sqrt{0,625 + 0,72}} = \frac{3}{\sqrt{1,345}} = 2,59$$

**d) Cuál es su decisión respecto a la hipótesis nula?**

Como su valor calculado Z (2,59) > 2,05; se rechaza la hipótesis nula y se acepta la hipótesis alternativa

Si Z tabulada es 0,5 - 0,02 = 0,48 este valor en la tabla es 2,05

**e) Cuál es el valor p?**

Z = 2,59 Area 0,4952

0,5 - 0,4952 = 0,0048 * 2 = 0,0096

Problemas resueltos.
1- las puntuaciones en un test que mide la variable creatividad siguen, en la población general de adolescentes, una distribución Normal de media 11,5. En un centro escolar que ha implantado un programa de estimulación de la creatividad una muestra de 30 alumnos ha
Proporcionado las siguientes puntuaciones:
11, 9, 12, 17, 8, 11, 9, 4, 5, 9, 14, 9, 17, 24, 19, 10, 17, 17, 8,23, 8, 6, 14, 16, 6, 7, 15, 20,14,15.
A un nivel de confianza del 95% ¿Puede afirmarse que el programa es efectivo?

Solución:
Ho m = 11,5
H1 m > 11,5
El estadístico de contraste en este caso es:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

La media muestral es 12,47 y la desviación típica de la muestra es 5,22, sustituyendo En el estadístico estos valores se obtiene:

$$t = \frac{12.47 - 11.5}{5.22/\sqrt{29}}$$

$$t = 1.00$$

Como el contraste es unilateral, buscamos en las tablas de la t de Student, con 29 grados de libertad, el valor que deja por debajo de sí una probabilidad de 0,95, que resulta ser 1,699

El valor del estadístico es menor que el valor crítico, por consiguiente se acepta la hipótesis nula.

La interpretación sería que no hay evidencia de que el programa sea efectivo.

**En una muestra de 1000 nacimientos el número de varones ha sido 542 ¿Puede considerarse, con un nivel de significación del 10%, que en general nacen más niños que niñas?**

Solución:

La hipótesis nula sería que nacen igual número de niños que de niñas, o lo que es lo mismo que la proporción de niños nacidos es igual 1/2.

Por consiguiente: Ho P = 0,5
H1 P > 0,5
El estadístico de contraste es:

$$\frac{p - Po}{\sqrt{\frac{Po * Qo}{n}}}$$

Como la proporción muestral es 542/1000 = 0,542, sustituyendo se obtiene el valor del estadístico:

$$\frac{0.542 - 0.5}{\sqrt{\frac{0.5 * 0.5}{1000}}}$$

Como el contraste es unilateral, buscamos en las tablas de la Normal el valor de la variable que deja por debajo de sí una probabilidad de 0,9, este valor es 1,282.

El valor del estadístico 2,66 es mayor que el valor crítico 1,282 por consiguiente, se rechaza la hipótesis nula.

Efectivamente, nacen en mayor proporción niños que niñas.

**En una muestra de 66 alumnos se ha calculado el coeficiente de correlación de Pearson entre sus puntuaciones en el primer parcial de Análisis de Datos y el tiempo que se emplea en desplazarse desde su domicilio hasta la Facultad, obteniéndose que r vale 0,24. Podemos mantener, con un nivel de confianza del 95%, la idea de que estas variables son incorreladas, o por el contrario debemos rechazarla.**

Solución:
Ho r = 0
H1 r ¹ 0
El estadístico de contraste es:

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Sustituyendo tenemos:

$$t=\frac{0.24\sqrt{64}}{\sqrt{1-0.0576}}$$

t= 1.98

El contraste es bilateral, por ello buscamos en las tablas de la t de Student, con 60 grados de libertad (el valor más próximo a 64 que figura en nuestras tablas), el valor que deja por debajo una probabilidad de 0,975 que es 2. Por tanto la región de aceptación será el intervalo (-2 ,, 2).

El valor del estadístico pertenece a la región de aceptación, por consiguiente se acepta la hipótesis nula.

No existe correlación entre ambas variables, de donde se deduce que el tiempo empleado no influye en la calificación.

**Suponga una variable aleatoria X para designar el peso de un pasajero de avión, que se interesa en conocer el peso promedio de todos los pasajeros. Como hay limitaciones de tiempo y dinero para pesarlos a todos, se toma una muestra de 36**

**pasajeros de la cual se obtiene una media muestral x= 160 lbs. Suponga además que la distribución de los pasajeros tenga una distribución normal con desviación estándar de 30, con un nivel de significancia de 0,05. Se puede concluir que el peso promedio de todos los pasajeros es menor que 170 lbs?**

***Datos***

$n = 36$

$\bar{x} = 160$

$\sigma = 30$

$\alpha = 0{,}05$

$H_0 : \mu \geq 170$

$H_1 : \mu < 170$

$\mu_0 = 170$

$$Z = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}} = \frac{160 - 170}{30 / \sqrt{36}} = \frac{-10}{5} = -2$$

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

Rechazo Ho si $Z < -Z_{1-\alpha}$

No rechazo Ho si $Z \geq -Z_{1-\alpha}$

$$Z_{1-0.05} = Z_{0,95} = 1{,}65$$

$$-2 < -1{,}65$$

## Problemas propuestos.

1. En un cultivo se viene trabajando a nivel comercial con una variedad cuyo rendimiento promedio es de 4 ton/ha y el desvío estándar es de 0.5 ton/ha. Mediante un proceso de selección, o sea de mejoramiento genético, los fitotecnistas "mejoran" la variedad buscando aumentar el rendimiento de la misma. Para decidir si la variedad mejoró el rendimiento se plantaron 25 parcelas, a la cosecha el rendimiento promedio que se obtuvo fue de 4.3 ton/ha.

2. Una compañía productora de cigarrillos sostiene que una marca que ellos producen tiene un contenido promedio de nicotina menor a 0.7 miligramos por cigarrillo. Para realizar la prueba se determinó el contenido de nicotina de 30 cigarrillos. Con esta muestra se obtuvo una media de 0.6920 y un desvío estándar de 0.0653.

3. En el primer ejemplo planteado, se trabajó con una variedad a la que se había hecho un proceso de mejoramiento. La varianza poblacional de la variedad antes de ser mejorada es de 0.25 kg2 y se está interesado en que se mantenga en ese valor durante el proceso de mejoramiento. En la muestra de 25 parcelas la varianza que se obtuvo fue de 0.29.

4. Un genetista que estudia el mecanismo hereditario del color de la flor en una especie sostiene la teoría que al cruzar dos tipos de dicha especie se obtendrán tres plantas con flores rojas por cada dos plantas con flores blancas. Si la evidencia experimental no apoya su afirmación, su teoría sobre el mecanismo hereditario no se sostendrá**.** Para poner a prueba su hipótesis el genetista observa el color de la flor en 10 plantas de cada una de 50 macetas que tiene sembradas con semillas provenientes de dicho cruzamiento y observa que en total 356 plantas tienen flores rojas.

5. Se ha tomado una muestra aleatoria de 100 individuos a los que se ha medido el nivel de glucosa en sangre, obteniédose una media muestral de 110 mg/cc. Se sabe que la desviación típica de la población es de 20 mg/cc.
   **a)** Obtén un intervalo de confianza, al 90%, para el nivel de glucosa en sangre en la población
   **b)** ¿Qué error máximo se comete con la estimación anterior?

6. La media de edad de los alumnos que se presentan a pruebas de acceso a la Universidad es de 18,1 años, y la desviación típica 0,6 años.
   **a)** De los alumnos anteriores se elige, al azar, una muestra de 100. ¿ Cuál es la probabilidad de que la media de la edad de la muestra esté comprendida entre 17,9 y 18,2 años?.
   **b)** ¿Qué tamaño debe tener una muestra de dicha población para que su media esté comprendida entre 17,9 y 18,3 años, con una confianza del 99,5%?

7. Las medidas de los diámetros de una muestra tomada al azar, de 200 cojinetes de bolas, hechos por una determinada máquina, dieron una media de 2 cm y una desviación típica de 0,1 cm. Hallar los intervalos de confianza del :

   - 68,26%
   - 95,44%
   - 99,73%

Para el diámetro de todos los cojinetes.

8. Se sabe que el contenido de fructosa de cierto alimento sigue una distribución normal, cuya varianza es conocida, teniendo un valor de 0,25. Se desea estimar el valor de la media poblacional mediante el valor de la media de una muestra, admitiendose un error máximo de 0,2 con una confianza del 95%. ¿Cuál ha de ser el tamaño de la muestra?

9. En una determinada población juvenil, el peso, en Kgs sigue una distribución normal N(50,10).
   Si se extrae una muestra aleatoria de 25 jóvenes y para un nivel de significación del 5%,¿ en qué condiciones se rechazaría la hipótesis de que la media de la población es de 50 kgs ?

10. De 120 alumnos, la proporción de que tengan dos o más hermanos es de 48/120. Indica los parámetros de la distribución a la que se ajustarían las muestras de tamaño 30.

11. En un isntituto de Enseñanza Secundaria hay matriculados 800 alumnos. A una muestra seleccionada aleatoriamente de un 15% de ellos, se les preguntó si utilizaban la cafetería del instituto. Contestaron negativamente un total de 24 alumnos.

    a. Estima el porcentaje de alumnos que utilizan la cafetería del instituto
    b. Determina, con una confianza del 99%, el error máximo cometido con dicha estimación

    Para estimar la proporción de familias de una determinada ciudad que poseen microondas, se quiere realizar una muestra aleatoria de medida n.

    12. Calcula el valor mínimo de n para garantizar que, a un nivel de confianza del 95%, el error en la estimación sea menor que 0,05. (Como se desconoce la proporción, se ha de tomar el caso más desfavorable, que será 0,5)

12. El peso medio de una muestra aleatoria de 81 personas de una determinada población es de 63,6 kg. Se sabe que la desviación típica poblacional es de 6 kg. Con un nivel de significación del 0,05, ¿hay suficientes evidencias para rechazar la afirmación de que el peso medio poblacional es de 65 kg?.

13. Una encuesta realizada a 64 empleados de una fábrica, concluyó que el tiempo medio de duración de un empleo en la misma es de 6,5 años, con una desviación típica de 4. ¿Sirve esta información para aceptar, con un nivel de significación del 5%, que el tiempo medio de empleo en esa fábrica es menor o igual que 6?. Justifica adecuadamente la respuesta.

14. En un determinado barrio se selaccionó al azar una muestra de 100 personas cuya media de ingresos mensuales resultaba igual a 106.000 pta. con una desviación típica de 20.000 PTAS.

I. Si se toma un nivel de confianza del 95%, ¿cuál es el intervalo de confianza para la media de los ingresos mensuales de toda la población?
II. Si se toma un nivel de significación igual a 0,01, ¿cuál es el tamaño muestral necesario para estimar la media de ingresos mensuales con un error menor de 3.000 PTAS.?

15. Se desea estimar la proporción p de individuos daltónicos de una población a través del porcentaje observado en una muestra aleatoria de individuos de tamaño n.

    a. Si el porcentaje de individuos daltónicos en la muestra es igual al 30%, calcula el valor de n para que, con un nivel de confianza dde 0,95, el error cometido en la estimación sea inferior al 3,1%.
    b. Si el tamaño de la muestra es de 64 individuos y el porcentaje de individuos daltónicos en la muestra es del 35%, determina, usando un nivel de significación del 1%, el correspondiente intervalo de confianza para la proporción de daltónicos de la población.

16. Supongamos que, a partir de una muestra aleatoria de tamaño n=25, se ha calculado el intervalo de confianza para la media de una población normal, obteniéndose una amplitud de ±4.
    Si el tamaño de la muestra hubiera sido n=100, permaneciendo invariable todos los demás valores que intervienen en el cálculo, ¿cuál habría sido la amplitud del intervalo?

17. Se desea estudiar el gasto semanal de fotocopias, en pesetas, de los estudiantes de bachillerato de Madrid. Para ello, se ha elegido una mustra aleatoria de 9 de estos estudiantes, resultando los valores siguientes para estos gastos:

    100 150 90 70 75 105 200 120 80

    Se supone que la variable aleatoria objeto de estudio sigue una distribución normal de media desconocida y de desviación típica igual a 12.

18. Determina un intervalo de confianza del 95% para la media del gasto semanal en fotocopias por estudiante.

19. Se sabe que la renta anual de los individuos de una localidad sigue una distribución normal de media desconocida y de desviación típica 0,24 millones. Se ha observado la renta anual de 16 individuos de esa localidad escogidos al azar, y se ha obtenido un valor medio de 1,6 millones de pesetas.
    Contrasta, a un nivel de significación del 5%, si la media de la distribución es de 1,45 millones de pesetas.

    a. ¿Cuáles son las hipótesis nula y alternativa del contraste?
    b. Determina la forma de la región crítica.
    c. ¿Se acepta la hipótesis nula con el nivel de significación indicado?

20. La media de las medidas de los diámetros de una muestra aleatoria de 200 bolas de rodamiento fabricadas por cierta máquina fue de 0,824 cm y la desviación típica fue de 0,042 cm. Halla los límites de confianza al 95% para el diámetro medio de las bolas fabricadas por esa máquina.

21. La Concejalía de la Juventud de un Ayuntamiento maneja el dato de que la edad a la que los hijos se independizan de sus padres es una variable normal con media 29 años y desviación típica 3 años. Aunque la desviación típica no plantea dudas, si se sospecha que la media ha descendido, sobre todo por la politica de ayuda al empleo que ha llevado a cabo el Ayuntamiento. Así de un estudio reciente sobre 100 jóvenes que se acaban de independizar, se ha obtenido una media de 28,1 años de edad.

    - Con un nivel de significación del 1%, ¿ puede defenderse que la edad media no ha disminuido, frente a que si lo ha hecho como parecen indicar los datos? Plantea el contraste o test de hipótesis y resuélvelo.
    - Explica en el contexto del problema, en qué consisten cada uno de los errores de tipo I y II.

    (Algunos valores de la función de distribución de la Normal de media 0 y desviación típica 1 : F(100)=1 ; F(3)=0,999 ; F(2,33)=0,99 ; F(0,01)=0,504).

22. Tras realizar un test de cultura general entre los habitantes de cierta población, se observa que las puntuaciones siguen una distribución normal, de media 68 y desviación típica 18. Se desea clasificar a los habitantes en tres grupos ( de baja cultura general, de cultura general aceptable, de cultura general excelente), de manera que el primer grupo abarque un 20% de la población, el segundo un 65% y el tercero el 15% restante. ¿Cuáles son las puntuaciones que marcan el paso de un grupo a otro?

23. En los folletos de propaganda, una empresa asegura que las bombillas que fabrica tienen una duración media de 1600 horas. A fin de contrastar este dato, se tomó una muestra aleatoria de 100 bombillas, obteniendose una duración media de 1.570 horas, con una desviación típica de 120 horas. ¿Puede aceptarse la información de los folletos con un nivel de confianza del 95%?

24. En una gran ciudad española, la altura media de sus habitantes tiene una desviación típica de 8 cm. Se pide :

    - Si la altura media de dichos habitantes fuera de 175 cm. ¿cuál sería la probabilidad de que la altura media de una muestra de 100 individuos tomada al azar fuera superior a 176 cm? Explica los pasos seguidos para obtener la respuesta.
    - Si se considera una muestra aleatoria de 100 individuos de esta ciudad, se obtiene una altura media de 178 cm. Determina un intervalo de confianza del 95% para la altura media de los habitantes de esta ciudad. Explica los pasos seguidos para obtener la respuesta.

# CAPÍTULO 9: PRUEBAS CUANDO SE TIENEN DOS COLAS

## Diferencia entre muestras independientes y dependientes.

"Dos muestras son independientes o dependientes entre sí, en función de si las observaciones de las muestras se han obtenido de los mismos individuos u objetos o no."[3]

Si ambas muestras se obtienen de distintos individuos, máquinas, empresas, objetos, etc.…no hay nada en común en dichas muestras lo que hace que ambas sean **"independientes".**

Sin embargo, si las observaciones o valores de ambas muestras se obtienen de los mismos individuos, empresas, agentes, etc., diremos que hay algo en común en dichas muestras por lo que serán muestras **"dependientes" o "no independientes".**

## Intervalo de confianza para la diferencia de medias de dos distribuciones normales, varianzas desconocidas.

En esta sección se verá el caso en donde se tienen dos poblaciones con medias y varianzas desconocidas, y se desea encontrar un intervalo de confianza para la diferencia de dos medias $\mu_1 - \mu_2$ Si los tamaños de muestras $n_1$ y $n_2$ son mayores que 30, entonces, puede emplearse el intervalo de confianza de la distribución normal. Sin embargo, cuando se toman muestras pequeñas se supone que las poblaciones de interés están distribuidas de manera normal, y los intervalos de confianza se basan en la distribución t.

## Intervalo de confianza para la diferencia de medias de dos distribuciones normales, varianzas desconocidas pero iguales.

Si $s_1^2$ y $s_2^2$ son las medias y las varianzas de dos muestras aleatorias de tamaño $n_1$ y $n_2$, respectivamente, tomadas de dos poblaciones normales e independientes con varianzas desconocidas pero iguales, entonces un intervalo de confianza del $100(1-\alpha)$ por ciento para la diferencia entre medias es:

$$\mu_1 - \mu_2 = (\overline{x_1} - \overline{x_2}) \pm t s_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

[3] Alexis Acevedo Rubilar

En donde: es el estimador combinado de la desviación estándar común de la población con $n_1 + n_2 - 2$ grados de libertad.

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

**Ejemplos:**

1. Un artículo publicado dio a conocer los resultados de un análisis del peso de calcio en cemento estándar y en cemento contaminado con plomo. Los niveles bajos de calcio indican que el mecanismo de hidratación del cemento queda bloqueado y esto permite que el agua ataque varias partes de una estructura de cemento. Al tomar diez muestras de cemento estándar, se encontró que el peso promedio de calcio es de 90 con una desviación estándar de 5; los resultados obtenidos con 15 muestras de cemento contaminado con plomo fueron de 87 en promedio con una desviación estándar de 4. Supóngase que el porcentaje de peso de calcio está distribuido de manera normal. Encuéntrese un intervalo de confianza del 95% para la diferencia entre medias de los dos tipos de cementos. Por otra parte, supóngase **que las dos poblaciones normales tienen la misma desviación estándar.**

**Solución:**

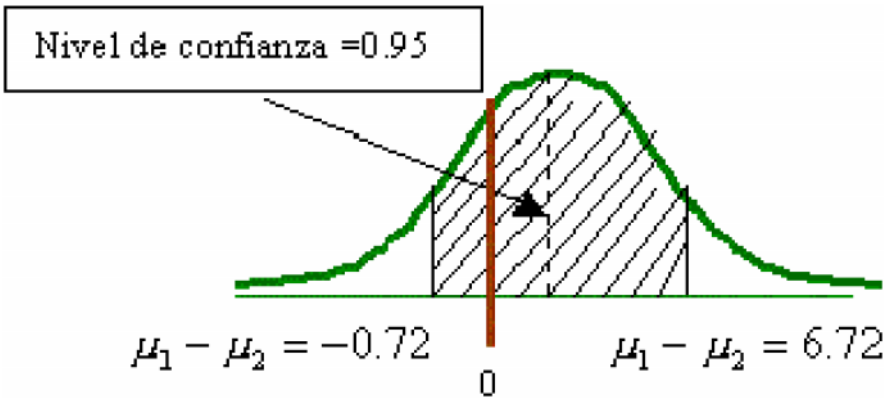
El estimador combinado de la desviación estándar es:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} = \frac{5^2(10 - 1) + 4^2(15 - 1)}{10 + 15 - 2} = 19{,}52$$

Al calcularle raíz cuadrada a este valor nos queda que $s_p = 4{,}41$ expresión que se reduce a $-0{,}72 \leq \mu_2 - \mu_1 \leq 6{,}72$

$$\mu_1 - \mu_2 = (\overline{x_1} - \overline{x_2}) \pm t s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (90 - 87) \pm (2{,}069)(4{,}41) \sqrt{\frac{1}{10} + \frac{1}{15}}$$

Nótese que el intervalo de confianza del 95% incluye al cero; por consiguiente, para este nivel confianza, no puede concluirse la existencia de una diferencia entre las medias.




2. Se realizó un experimento para comparar el tiempo promedio requerido por el cuerpo humano para absorber dos medicamentos, A y B. Suponga que el tiempo necesario para que cada medicamento alcance un nivel específico en el torrente sanguíneo se distribuye normalmente. Se eligieron al azar a doce personas para ensayar cada fármaco registrándose el tiempo en minutos que tardó en alcanzar un nivel específico en la sangre. Calcule un intervalo de confianza del 95% para la diferencia del tiempo promedio. Suponga varianzas iguales.

| Medicamento A | Medicamento B |
|---|---|
| $n_A = 12$ | $n_B = 12$ |
| $\overline{x} = 26{,}8$ | $\overline{x} = 32{,}6$ |
| $s_A^2 = 15.57$ | $s_B^2 = 17.54$ |

**Solución:**

$$\boldsymbol{S_P = \sqrt{\frac{S_A^2(n_A-1)+S_B^2(n_B-1)}{n_A+n_B-2}} = \sqrt{\frac{15{,}57(12-1)+17{,}54(12-1)}{12+12-2}} = 4{,}07}$$

$$\mu_B - \mu_A = (\overline{x_B} - \overline{x_A}) \pm t s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = (32{,}6 - 26{,}8) \pm (2{,}074)(4{,}07)\sqrt{\frac{1}{12} + \frac{1}{12}}$$

$$2{,}35 \leq \mu_B - \mu_A \leq 9{,}25$$

Con un nivel confianza del 95% se sabe que el tiempo promedio para alcanzar un nivel específico es mayor para el medicamento B.

# Prueba sobre dos medias, poblaciones normales, varianzas desconocidas pero iguales.

Las situaciones que más prevalecen e implican pruebas sobre dos medias son las que tienen varianzas desconocidas. Si el científico prueba mediante una prueba F, que las varianzas de las dos poblaciones son iguales, se utiliza la siguiente fórmula:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{{}^{1}/_{n_1} + {}^{1}/_{n_2}}}$$

Donde:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Los grados de libertad están dados por:

$$\alpha = n_1 + n_2 - 2$$

**Ejemplos:**

1. Para encontrar si un nuevo suero detiene la leucemia, se seleccionan nueve ratones, todos con una etapa avanzada de la enfermedad. Cinco ratones reciben el tratamiento y cuatro no. Los tiempos de sobrevivencia en años, a partir del momento en que comienza el experimento son los siguientes:

| | | | | | |
|---|---|---|---|---|---|
| **Con Tratamiento** | 2.1 | 5.3 | 1.4 | 4.6 | 0.9 |
| **Sin Tratamiento** | 1.9 | 0.5 | 2.8 | 3.1 | |

¿Se puede decir en el nivel de significancia del 0.05 que el suero es efectivo? Suponga que las dos poblaciones se distribuyen normalmente con varianzas iguales.

Solución:

Primero se probará el supuesto de varianzas iguales con un ensayo de hipótesis bilateral utilizando la distribución Fisher.

Datos:

Con tratamiento

- $\overline{X} = 2{,}86$

- $S = 1{,}97$
- $n = 5$

Sin tratamiento

1- $\overline{X} = 2{,}75$
2- $S = 1{,}1672$
3- $n = 4$

## Ensayo de hipótesis:

$$H_0; \sigma^2_{CT} \Big/ \sigma^2_{ST} = 1$$

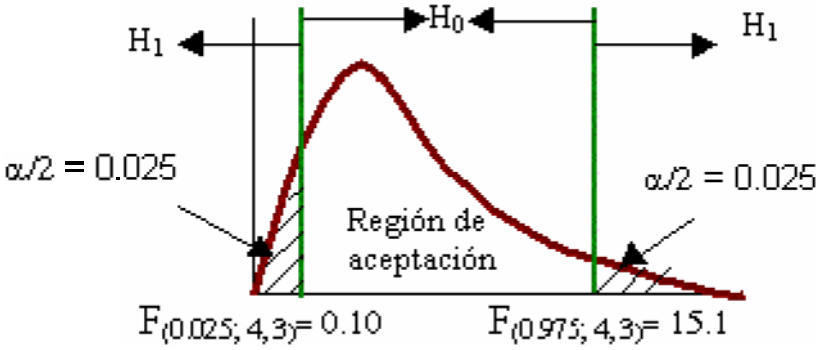
$$H_0; \sigma^2_{CT} \Big/ \sigma^2_{ST} \neq 1$$

## Estadístico de prueba:

$$F = \frac{S^2_{CT}}{S^2_{ST}}$$

La sugerencia que se hace es que el numerador sea el de valor mayor.

Entonces los grados de libertad uno será el tamaño de la muestra de la población uno menos uno. $\nu_1$= 5-1 = 4 y $\nu_2$ = 4-1=3.




## Regla de decisión:

- Si $0.10 \leq F_c \leq 15.1$ No se rechaza $H_o$,
- Si la $F_c < 0.10$ ó si $F_c > 15.1$ se rechaza $H_o$.

## Cálculo:

$$F = \frac{S_{CT}^2}{S_{ST}^2} = \frac{1{,}97^2}{1{,}167^2} = 2{,}85$$

## Decisión y Justificación:

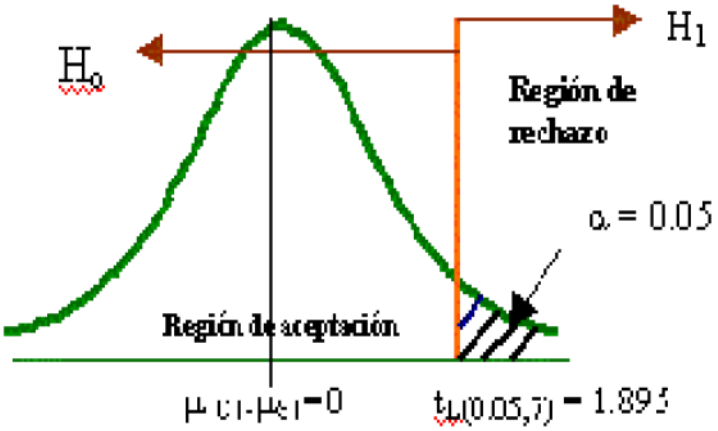
Como 2.85 está entre los dos valores de $H_o$ no se rechaza, y se concluye con un $\alpha$ = 0.05 que existe suficiente evidencia para decir que las varianza de las poblaciones son iguales.

Con la decisión anterior se procede a comparar las medias:

Ensayo de Hipótesis

$$H_0; \mu_{CT} - \mu_{ST} = 0$$

$$H_1; \mu_{CT} - \mu_{ST} > 0$$




Los grados de libertad son (5+4-2) = 7

## Regla de decisión:

- Si $t_R \leq$ 1.895 No se Rechaza $H_o$
- Si $t_R$ > 1.895 se rechaza $H_o$

## Cálculos:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} = \frac{1{,}97(5-1) + 1{,}672^2(4-1)}{5+4-2} = 3{,}415$$

Por lo tanto $s_p$ = 1.848

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} = \frac{(2{,}86 - 2{,}075) - 0}{1{,}848\sqrt{1/5 + 1/4}} = 0{,}6332$$

## Justificación y decisión:

Como 0.6332 es menor que 1.895, no se rechaza $H_o$, y se concluye con un nivel de significancia del 0.05 que no existe suficiente evidencia para decir que el suero detiene la leucemia.

2. Se realizó un experimento para comparar el tiempo promedio requerido por el cuerpo humano para absorber dos medicamentos, A y B. Suponga que el tiempo necesario para que cada medicamento alcance un nivel específico en el torrente sanguíneo se distribuye normalmente. Se eligieron al azar a doce personas para ensayar cada fármaco registrándose el tiempo en minutos que tardó en alcanzar un nivel específico en la sangre. Calcule con $\alpha$

= 0.05 si existe diferencia entre los tiempos promedio y obtenga el valor de P. Suponga varianzas iguales.

| Medicamento A | Medicamento B |
|---|---|
| $\boldsymbol{n_A = 12}$ | $\boldsymbol{n_B = 12}$ |
| $\bar{x}_A = 26.8$ | $\bar{x}_B = 32.6$ |
| $\boldsymbol{S_A^2 = 15.57}$ | $\boldsymbol{S_B^2 = 17.54}$ |

## Solución:

Primero se pondrá a prueba el supuesto de varianzas iguales mediante una prueba de hipótesis con $\alpha$ = 0.10.

Ensayo de hipótesis:

$H_0; \sigma_B^2 - \sigma_A^2 = 1$

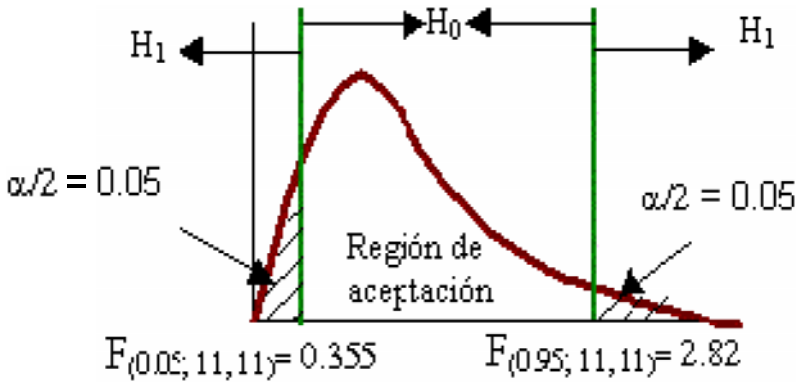
$H_0; \sigma_B^2 - \sigma_A^2 \neq 1$

Estadístico de prueba:

$$F = \frac{S_{CT}^2}{S_{ST}^2}$$

La sugerencia que se hace es que el numerador sea el de valor mayor.

Entonces los grados de libertad uno será el tamaño de la muestra de la población uno menos uno. $\nu_1$=12-1=11 y $\nu_2$=12-1=11.




Regla de decisión:

Si 0.355 $\leq$ $F_c$ $\leq$ 2.82 No se rechaza $H_o$,

Si la $F_c$ < 0.355 ó si $F_c$ > 2.82 se rechaza $H_o$.

Cálculo:

$$F = \frac{S_{CT}^2}{S_{ST}^2} = \frac{17,54}{15,57} = 1,13$$

## Decisión y Justificación:

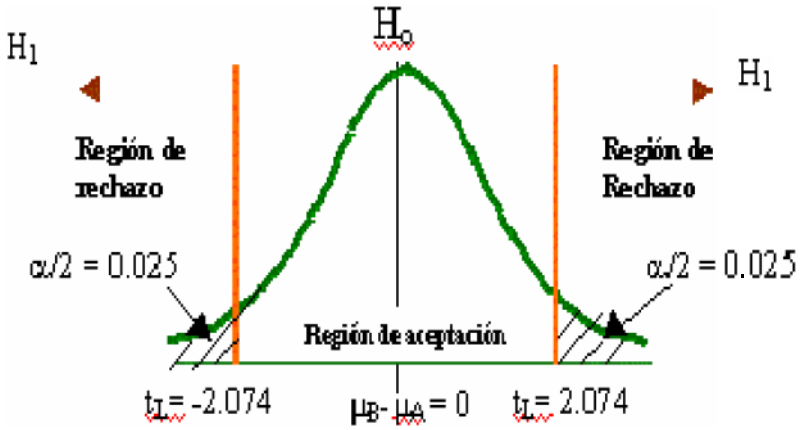
Como 1.13 está entre los dos valores de $H_o$ no se rechaza, y se concluye con un $\alpha$ = 0.10 que existe suficiente evidencia para decir que las varianza de las poblaciones son iguales.

Con la decisión anterior se procede a comparar las medias:

## Ensayo de Hipótesis

$H_0; \mu_{CT} - \mu_{ST} = 0$

$H_0; \mu_{CT} - \mu_{ST} \neq 0$




Los grados de libertad son (12+12-2) = 22

Regla de decisión:

Si $-2.074 \leq t_c \leq 2.074$ No se rechaza $H_o$,

Si la $t_c < -2.074$ ó si $t_c > 2.074$ se rechaza $H_o$.

## Cálculos:

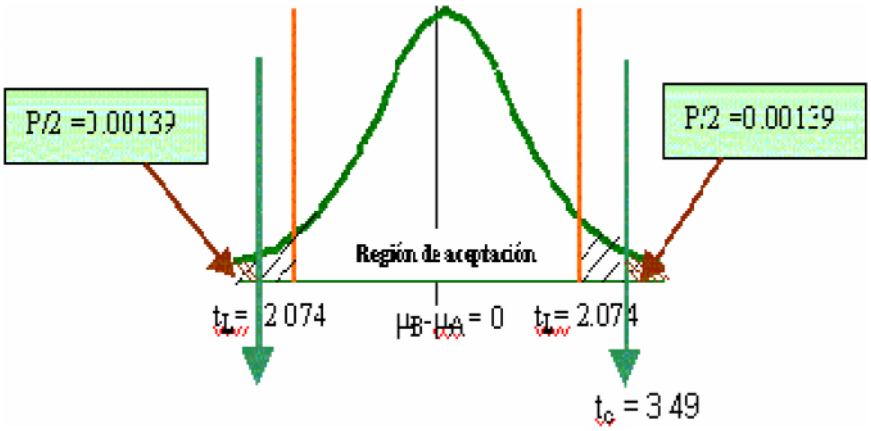
$$s_p^2 = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}} = \sqrt{\frac{15{,}57(12-1)+17{,}54(12-1)}{12+12-2}} = 4{,}07$$

$$t = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} = \frac{(32{,}6-26{,}8)-0}{1{,}848\sqrt{1/12+1/12}} = 3{,}49$$

## Justificación y decisión:

Como 3.49 es mayor que 2.074, no se rechaza $H_o$, y se concluye con un nivel de significancia del 0.05 que la media del tiempo para que el medicamento A llegue a un nivel específico en el torrente sanguíneo es distinta de la que toma al fármaco B alcanzar ese mismo nivel.

Para calcular el valor de P se ubicará la t calculada en la gráfica para proceder a buscar el área y multiplicarla por dos ya que es bilateral.




P = (2) (0.00139) = 0.00278

## Intervalo de confianza para la diferencia de medias de dos distribuciones normales, varianzas desconocidas pero diferentes.

Consideremos ahora el problema de encontrar una estimación por intervalos de $\mu_1 - \mu_2$ cuando no es probable que las varianzas poblacionales desconocidas sean iguales. La estadística que se usa con más frecuencia en este caso es:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

Que tiene aproximadamente una distribución t con $\alpha$ grados de libertad, donde:

$$\alpha = \frac{(S_1^1/n_1 * S_2^2/n_2)^2}{\left\{\left(S_1^2/n_1\right)^2 / (n_1 - 1)\right\}\left\{\left(S_2^2/n_2\right)^2 / (n_2 - 1)\right\}}$$

Como $\alpha$ rara vez es número entero, lo redondeamos al número entero más cercano menor. Esto es si el valor de nu es de 15.9 se redondeará a 15.

Al despejar la diferencia de medias poblacionales de la formula de t nos queda:

$$\mu_B - \mu_A = (\overline{x_1} - \overline{x_2}) \pm t\sqrt{\frac{S_1^2}{n_A} + \frac{S_2^2}{n_B}}$$

**Ejemplos:**

1. El departamento de zoología de la Universidad de Virginia llevó a cabo un estudio para estimar la diferencia en la cantidad de orto fósforo químico medido en dos estaciones diferentes del río James. El orto fósforo se mide en miligramos por litro. Se reunieron 15 muestras de la estación 1 y se obtuvo una media de 3.84 con una desviación estándar de 3.07 miligramos por litro, mientras que 12 muestras de la estación 2 tuvieron un contenido promedio de 1.49 con una desviación estándar 0.80 miligramos por litro. Encuentre un intervalo de confianza de 95% para la diferencia del contenido promedio real de orto fósforo en estas dos estaciones, suponga que las observaciones vienen de poblaciones normales con varianzas diferentes.

Solución:

| Estación 1 | Estación 2 |
|---|---|
| $n_1 = 16$ | $n_2 = 12$ |
| $\overline{x} = 3{,}84$ | $\overline{x} = 1{,}49$ |
| $S_1 = 3{,}07$ | $S_2 = 0{,}8$ |

Primero se procederá a calcular los grados de libertad:

$$\alpha = \frac{\left(S_1^1/n_1 * S_2^2/n_2\right)^2}{\left\{\left(S_1^2/n_1\right)^2 / (n_1-1)\right\}\left\{\left(S_2^2/n_2\right)^2 / (n_2-1)\right\}} =$$

$$\frac{\left(3{,}07^2/15 * 0{,}8^2/12\right)^2}{\left\{\left(13{,}07^2/15\right)^2 / (14)\right\}\left\{\left(0{,}8^2/12\right)^2 / (11)\right\}} = 16$$

Al usar $\alpha$=0.05, encontramos en la tabla con 16 grados de libertad que el valor de t es 2.120, por lo tanto:

$$\mu_B - \mu_A = (\overline{x_1} - \overline{x_2}) \pm t\sqrt{\frac{S_1^2}{n_A} + \frac{S_2^2}{n_B}} = (3{,}84 - 1{,}49) \pm 2{,}120\sqrt{\frac{3{,}07^2}{15} + \frac{0{,}8^2}{12}}$$

Que se simplifica a:

$0{,}6 \leq \mu_B - \mu_A \leq 4{,}1$

Por ello se tiene una confianza del 95% de que el intervalo de 0.60 a 4.10 miligramos por litro contiene la diferencia de los contenidos promedios reales de orto fósforo para estos dos lugares.

## Prueba sobre dos medias, poblaciones normales, varianzas desconocidas pero diferentes.

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{{}^1/_{n_1} + {}^1/_{n_2}}}$$

$$\alpha = \frac{\left({}^{S_1^1}/_{n_1} * {}^{S_2^2}/_{n_2}\right)^2}{\left\{\left({}^{S_1^2}/_{n_1}\right)^2 / (n_1 - 1)\right\}\left\{\left({}^{S_2^2}/_{n_2}\right)^2 / (n_2 - 1)\right\}}$$

**Ejemplo:**

1. Un fabricante de monitores prueba dos diseños de microcircuitos para determinar si producen un flujo de corriente equivalente. El departamento de ingeniería ha obtenido los datos siguientes:

| | | | |
|---|---|---|---|
| **Diseño 1** | $n_1 = 16$ | $\bar{x}_1 = 24.2$ | $s_1^2 = 10$ |
| **Diseño 2** | $n_2 = 10$ | $\bar{x}_2 = 23.9$ | $s_2^2 = 40$ |

Con $\alpha$ = 0.05, se desea determinar si existe alguna diferencia significativa en el flujo de corriente promedio entre los dos diseños, donde se supone que las dos poblaciones son normales, pero no es posible suponer que las varianzas desconocidas sean iguales.

Solución:

**Primero se probarán varianzas desiguales.**

Ensayo de hipótesis:

$$H_0; \sigma_2^2 - \sigma_1^2 = 1$$

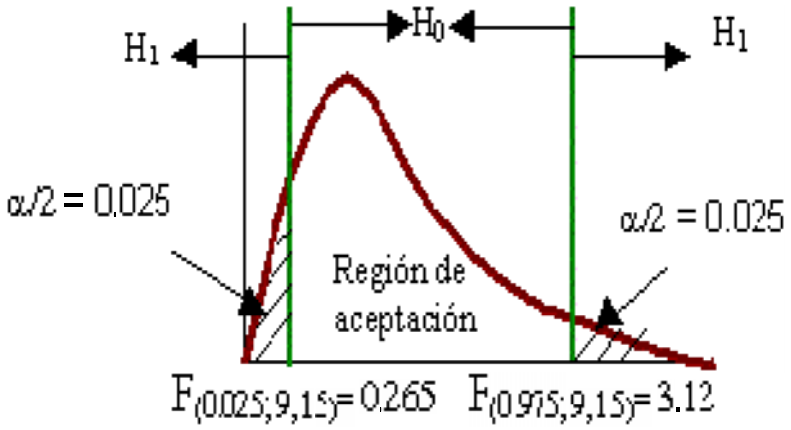
$$H_0; \sigma_2^2 - \sigma_1^2 \neq 1$$

Estadístico de prueba:

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

La sugerencia que se hace es que el numerador sea el de valor mayor.

Entonces los grados de libertad uno será el tamaño de la muestra de la población uno menos uno. $\nu_1$= 10-1 = 9 y $\nu_2$ = 16-1=15.




Regla de decisión:

Si 0.265 $\leq$ $F_c$ $\leq$ 3.12 No se rechaza $H_o$,

Si la $F_c$ < 0.265 ó si $F_c$ > 3.12 se rechaza $H_o$.

Cálculo:

$$F = \frac{S_2^2}{S_1^2} = \frac{40}{10} = 4$$

## Decisión y Justificación:

Como 4 es mayor que 3.12 se rechaza $H_o$, y se concluye con un $\alpha$ = 0.05 que existe suficiente evidencia para decir que las varianza de las poblaciones son diferentes.

Con la decisión anterior se procede a comparar las medias:

## Ensayo de Hipótesis

$$H_0; \mu_{CT} - \mu_{ST} = 0$$

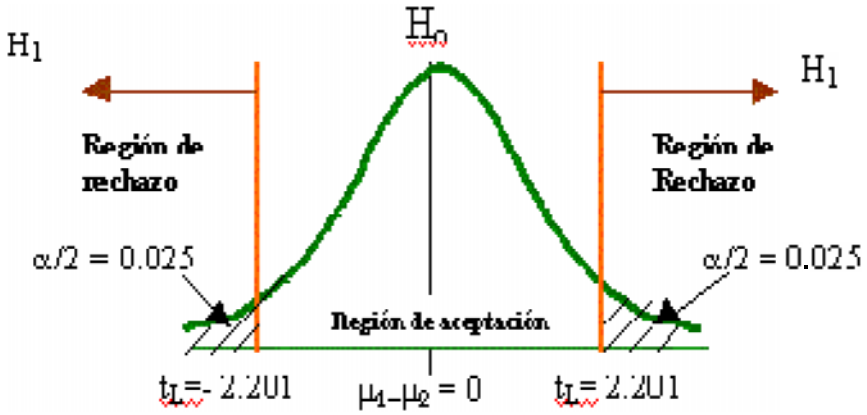
$$H_0; \mu_{CT} - \mu_{ST} \neq 0$$

Para poder buscar el valor de t en la tabla, se necesita saber el valor de los grados de libertad:

$$\alpha = \frac{\left({}^{S_1^1}/_{n_1} * {}^{S_2^2}/_{n_2}\right)^2}{\left\{\left({}^{S_1^2}/_{n_1}\right)^2 \Big/ (n_1 - 1)\right\}\left\{\left({}^{S_2^2}/_{n_2}\right)^2 \Big/ (n_2 - 1)\right\}}$$

$$= \frac{\left({}^{10}/_{16} * {}^{40}/_{10}\right)^2}{\left\{\left({}^{10}/_{16}\right)^2 \Big/ (15)\right\}\left\{\left({}^{40}/_{10}\right)^2 \Big/ (9)\right\}} = 11$$

Este valor se redondea al próximo menor que sería 11.




Regla de decisión:

Si $-2.201 \leq t_R \leq 2.201$ No se rechaza $H_o$

Si $t_R < -2.201$ ó si $t_R > 2.201$ se rechaza $H_o$

**Cálculos:**

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_{21}^2/n_2}} = \frac{(24{,}2 - 23{,}9) - 0}{\sqrt{10/16 + 40/10}} = 0{,}1395$$

**Justificación y decisión:**

Como 0.1395 está entre –2.201 y 2.201, no se rechaza $H_o$ y se concluye con un $\alpha$ = 0.05, que no existe diferencia significativa en el flujo de corriente promedio entre los dos diseños.

2. Dos proveedores fabrican un engrane de plástico utilizado en una impresora láser. Una característica importante de estos engranes es la resistencia al impacto la cual se mide en pies-libras. Una muestra aleatoria de 10 engranes suministrados por el primer proveedor arroja los siguientes resultados: $\overline{X_1}$ =290 y $s_1$ = 12. Del segundo proveedor se toma una muestra aleatoria de 16 engranes, donde los resultados son $\overline{X_2}$ =321 y $s_2$ = 45. ¿Existe evidencia que apoye la afirmación de que los engranes del proveedor 2 tienen una mayor resistencia promedio al impacto? Use un nivel de significancia de 0.05. Calcule el valor de P.

Solución:

Estación 1 Estación 2

$$n_1 = 10 \qquad n_2 = 16$$

$$\overline{x} = 290 \qquad \overline{x} = 321$$

$$S_1 = 12 \qquad S_2 = 45$$

**Primero se probarán varianzas desiguales.**

Ensayo de hipótesis:

$$H_0; \sigma_2^2 - \sigma_1^2 = 1$$

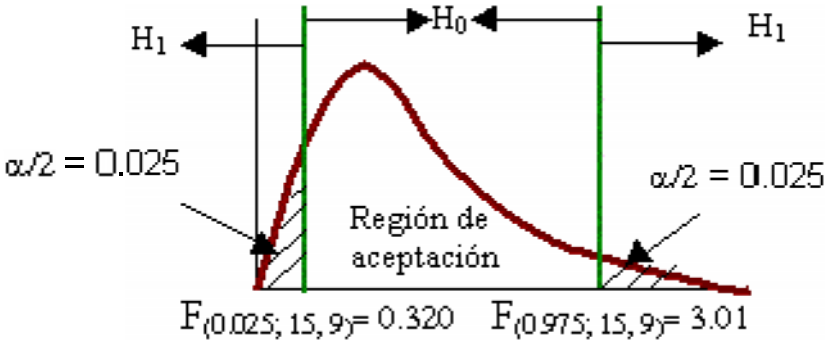
$$H_0; \sigma_2^2 - \sigma_1^2 \neq 1$$

Estadístico de prueba:

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

La sugerencia que se hace es que el numerador sea el de valor mayor.

Entonces los grados de libertad uno será el tamaño de la muestra de la población uno menos uno. $\nu_1$= 16-1 = 15 y $\nu_2$ = 10-1=9.




Regla de decisión:

Si 0.320 $\leq$ $F_c$ $\leq$ 3.01 No se rechaza $H_o$,

Si la $F_c$ < 0.320 ó si $F_c$ > 3.01 se rechaza $H_o$.

Cálculo:

$$F = \frac{S_2^2}{S_1^2} = \frac{45^2}{12^2} = 14{,}06$$

## Decisión y Justificación:

Como 14.06 es mayor que 3.01 se rechaza $H_o$, y se concluye con un $\alpha$ = 0.05 que existe suficiente evidencia para decir que las varianza de las poblaciones son diferentes.

Con la decisión anterior se procede a comparar las medias:

## Ensayo de Hipótesis

$H_0; \mu_{CT} - \mu_{ST} = 0$

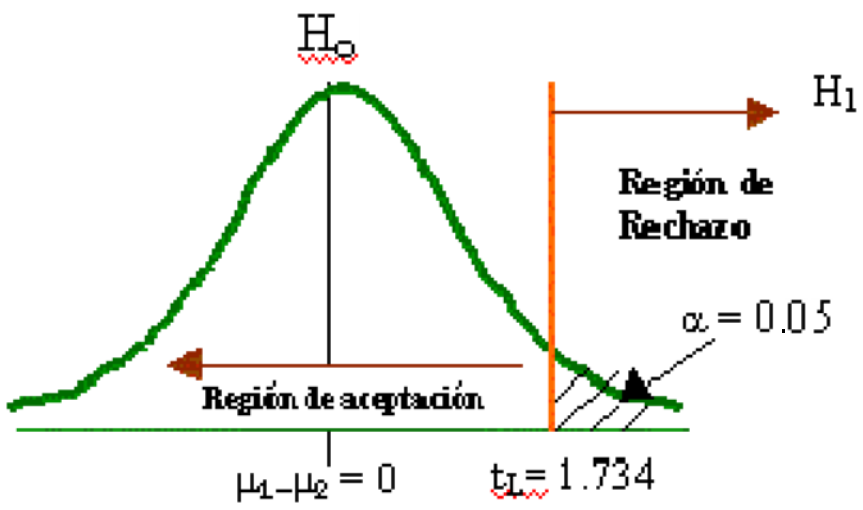
$H_0; \mu_{CT} - \mu_{ST} \neq 0$

Para poder buscar el valor de t en la tabla, se necesita saber el valor de los grados de libertad:

$$\alpha = \frac{\left({}^{S_1^1}/_{n_1} * {}^{S_2^2}/_{n_2}\right)^2}{\left\{\left({}^{S_1^2}/_{n_1}\right)^2 \Big/ (n_1 - 1)\right\}\left\{\left({}^{S_2^2}/_{n_2}\right)^2 \Big/ (n_2 - 1)\right\}}$$

$$= \frac{\left({}^{12^2}/_{16} * {}^{45^2}/_{10}\right)^2}{\left\{\left({}^{12^2}/_{16}\right)^2 \Big/ (9)\right\}\left\{\left({}^{45^2}/_{10}\right)^2 \Big/ (15)\right\}} = 18{,}21$$

Este valor se redondea al próximo menor que sería 18.




Regla de decisión:

Si $t_R \leq 1.734$ No se rechaza $H_o$

Si $t_R > 1.734$ se rechaza $H_o$

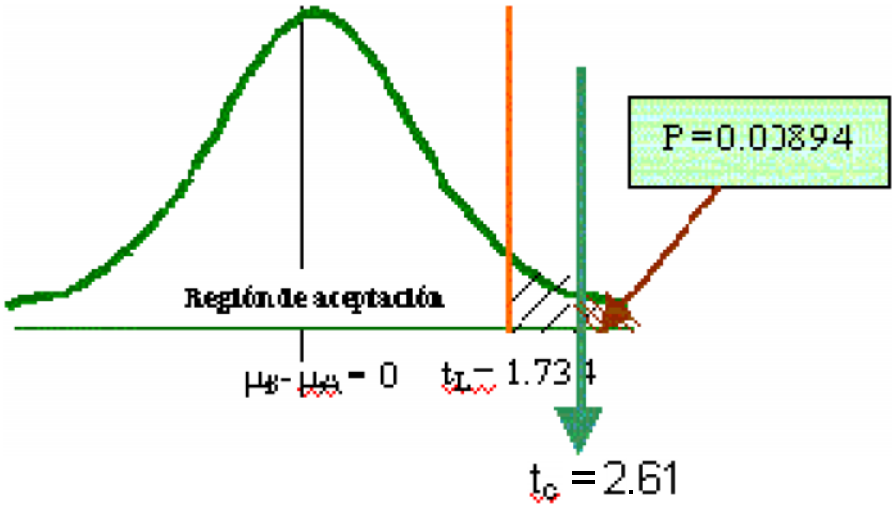
**Cálculos:**

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_{21}^2/n_2}} = \frac{(321 - 290) - 0}{\sqrt{12^2/16 + 45^2/10}} = 261$$

## Justificación y decisión:

Como 2.61 es mayor que 1.734, se rechaza $H_o$ y se concluye con un $\alpha$ =0.05, que existe evidencia suficiente para decir que el promedio de resistencia de los engranes del proveedor 2 es mayor a el promedio de resistencia de los engranes del proveedor 1.

Para calcular el valor de P se busca adentro de la tabla de t el valor de 2.61 con 18 grados de libertad y se observa que se encuentra entre dos áreas que son 0.01 y 0.0075, al interpolar nos da un valor de P = 0.00894.




## Ejercicios Del Capitulo.[4]

1- las pruebas con dos muestras

2- En un negocio de ropa se desea comparar la eficiencia (medida en el monto medio de sus ventas) de sus dos vendedores, Hernán y María Julia. Se sabe que el monto de las ventas para ambos vendedores tiene distribución normal. El desvío estándar de la distribución de Hernán es igual a $75 y el desvío estándar de la distribución de María Julia es $50. Una muestra de 36 ventas de Hernán proporcionó un monto promedio diario de $300, mientras que una muestra de 40 ventas de María Julia, proporcionó un monto promedio de $350. Con un nivel de significación del 5%, verificar si María Julia es más eficiente que Hernán.
RESPUESTA: Rc = z ≥ 1,645 e.p. = 3,38 Se rechaza $H_0$

3- Dos máquinas producen clavos cuyas longitudes siguen una ley normal. Con una muestra de 15 clavos de la maquina Amarilla se calculó una longitud media de 85 mm y un desvío están-dar de 7 mm. Con una muestra de 20 clavos de la máquina Verde, se calculó una media de 90 mm y un desvío de 4 mm.
Probar, con un nivel de significación del 5%, si existen diferencias significativas en la longitud media de los clavos producidos por ambas máquinas.
RESPUESTA: Varianzas desconocidas y diferentes Rc = t ≥ 2,08 y t ≤–2,08 e.p.= |2,479| Se rechaza $H_0$

4- Se desea comparar la eficiencia de una nueva droga para tratar una determinada enferme-dad. Se tomaron dos grupos de pacientes, que padecen dicha enfermedad. Al grupo A, constituido por 200 pacientes se les suministró una nueva medicación, y al grupo B formado por 100 pacientes se les mantiene la medicación habitual. Luego de un mes, 25 personas del grupo A y 9 personas del

[4] Mirta Misevicius

grupo B muestran una marcada mejoría. Con un nivel de significación del 5%, ¿puede afirmarse que la nueva medicación es mejor que la medicación habitual?
RESPUESTA: Rc = z ≥ 1,645 e.p. = 0,9015 No se rechaza $H_0$

5- Se toma una muestra de 50 deudores por ventas de la SUCURSAL NORTE, obteniéndose un monto medio de la deuda de $40; y otra muestra de 40 deudores por ventas de la SUCURSAL SUR obteniéndose monto medio de la deuda de $33. Se sabe que en la SUCURSAL NORTE la varianza es de $4, y para la SUCURSAL SUR la varianza es $9. Con una probabilidad de cometer Error de Tipo I, igual a 0,05, ¿puede afirmarse que el monto medio de las deudas de la SUCURSAL NORTE es superior en $5 al promedio de la deuda de la SUCURSAL SUR?
RESPUESTA: Rc = z ≤-1,96 y z ≥ 1,96 e.p. = 3,62 Se rechaza $H_0$

6- Se desea comparar los pesos medios de dos grupos vacas alimentadas con distintos tipos de suplementos nutricionales. Una muestra de 25 vacas alimentadas con el Suplemento **A** proporcionó un peso medio de 1000 Kg y un desvío estándar de 200 Kg; mientras que una muestra de 25 vacas alimentadas con Suplemento **B** proporcionó un peso medio de 920 Kg y un desvío estándar de 250 Kg. Sabiendo que los pesos de los animales se distribuyen normalmente, verificar, con un nivel de significación del 5%, si se puede afirmar que los pesos medios de ambos grupos son iguales.
RESPUESTA: Varianzas desconocidas e iguales Rc = t ≥ 2,011 y t ≤–2,011 e.p. = -1,25 No se rechaza $H_0$

7- Se desea comprobar si hay diferencia significativa entre las proporciones de personas que consumen carne de pescado por lo menos una vez por semana, correspondiente a dos ciudades. En una muestra de 180 personas de la ciudad **A**, se encontró que el 23% consumen carne de pescado, mientras que 56 personas de una muestra de 200 de la ciudad **B**, comen carne de pescado. Realizar la prueba correspondiente con una probabilidad de rechazar la hipótesis nula cuando esta es verdadera, igual a 0,05.
RESPUESTA: Rc = z ≤-1,96 y z ≥ 1,96 e.p. = -1,116 No se rechaza

8- Lisa Monín es directora de presupuesto en la empresa New Process Company, desea comparar los gastos diarios de transporte del equipo de ventas y del personal de cobranza. Recopiló la siguiente información muestral (importe en dólares).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Ventas ($) | 131 | 135 | 146 | 165 | 136 | 142 | |
| Cobranza ($) | 130 | 102 | 129 | 143 | 149 | 120 | 139 |

9- De una población se toma una muestra de 40 observaciones. La media muestral es de 102 y la desviación estándar 5. De otra población se toma una muestra de 50 observaciones. La media mustral es ahora 99 y la desviación estándar es 6. Realice la siguiente prueba de hipótesis usando como nivel de significancia 0,04.

a) ¿Es esta una prueba de una o de dos colas?
b) Establezca la regla de decisión
c) Calcule el valor del estadístico de prueba

d) ¿Cuál es su decisión respecto a la hipótesis nula?

e) ¿Cuál es el valor p?

# CAPÍTULO 11: REGRESIÓN SIMPLE Y CORRELACIÓN

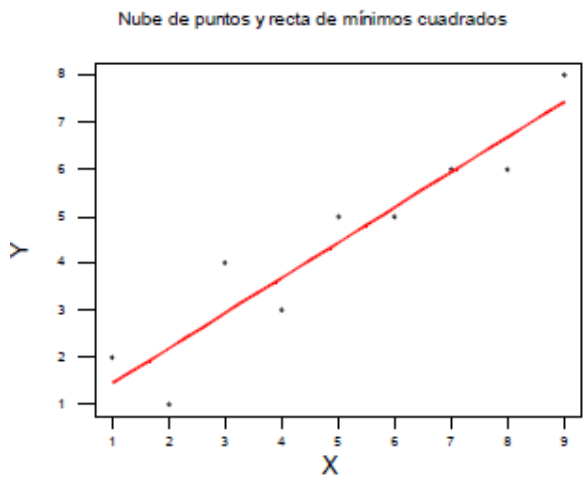
## Regresión simple y correlación.

- La Regresión y la correlación son dos técnicas estadísticas que se pueden utilizar para solucionar problemas comunes en los negocios.
- Muchos estudios se basan en la creencia de que es posible identificar y cuantificar alguna Relación Funcional entre dos o más variables, donde una variable depende de la otra variable.

## Regresión lineal.

- ✓ La regresión lineal técnica que usa variables aleatorias, continuas se diferencia del otro método analítica que es la correlación, porque esta última no distingue entre las variables respuesta y la variable explicativa por que las trata en forma simétrica.
- ✓ La regresión en forma grafica, trata de lograr que una dispersión de las frecuencias sea ajustada a una línea recta o curva.
- ✓ En el Modelo de Regresión es muy importante identificar cuál es la variable dependiente y cuál es la variable independiente.
- ✓ En el Modelo de Regresión Simple se establece que Y es una función de sólo una variable independiente, razón por la cual se le denomina también Regresión Divariada porque sólo hay dos variables, una dependiente y otra independiente y se representa así:
  Y = f (X)
  "Y está regresando por X"

En aquellos casos en que el coeficiente de regresión lineal sea “cercano” a +1 o a –1, tiene sentido considerar la ecuación de la recta que “mejor se ajuste” a la nube de puntos (recta de mínimos cuadrados). Uno de los principales usos de dicha recta será el de predecir o estimar los valores de Y que obtendríamos para distintos valores de X. Estos conceptos quedarán representados en lo que llamamos diagrama de dispersión:




Se puede decir que Y depende de X, en donde Y y X son dos variables cualquiera en un modelo de Regresión Simple.

"Y es una función de X"

Y = f(X)

Como Y depende de X,

Y es la variable dependiente, y

X es la variable independiente.

## Clases de regresión.

La regresión puede ser Lineal y Curvilínea o no lineal, ambos tipos de regresión pueden ser a su vez:

a. Esta regresión se utiliza con mayor frecuencia en las ciencias económicas, y sus disciplinas tecnológicas. Cualquier función no lineal, es linealizada para su estudio y efectos prácticos en las ciencias económicas, modelos no lineales y lineales multiecuacionales.

**Objetivo**:

Se utiliza la regresión lineal simple para:

1.- Determinar la relación de dependencia que tiene una variable respecto a otra.
2.- Ajustar la distribución de frecuencias de una línea, es decir, determinar la forma de la línea de regresión.
3.- Predecir un dato desconocido de una variable partiendo de los datos conocidos de otra variable.

## Coeficiente de regresión.

Indica el número de unidades en que se modifica la variable dependiente "Y" por efecto del cambio de la variable independiente "X" o viceversa en una unidad de medida.

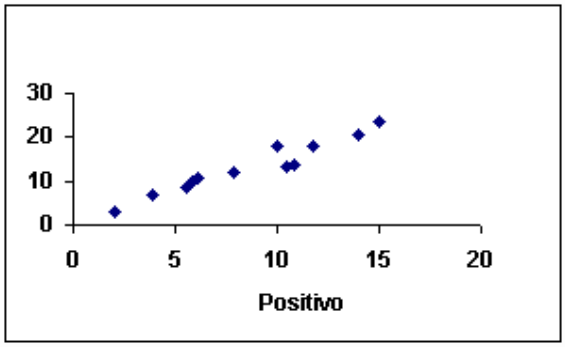
## Clases de coeficiente de regresión:

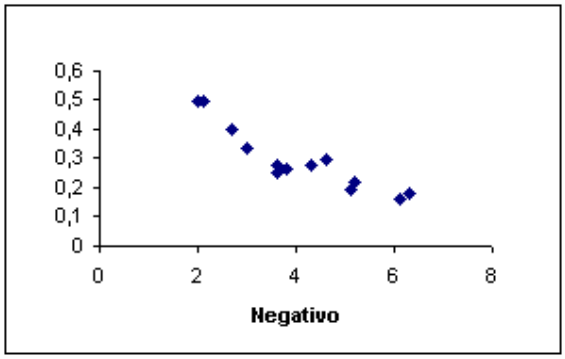
El coeficiente de regresión puede ser: Positivo, Negativo y Nulo.

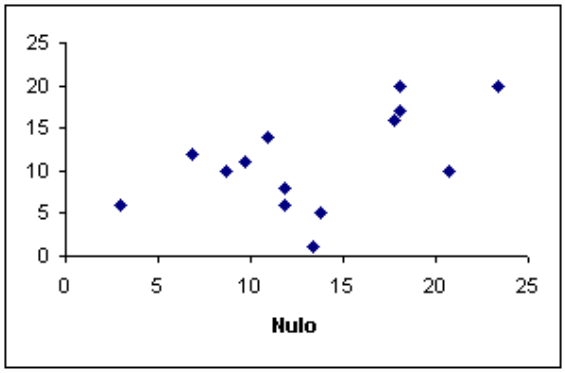
Es positivo cuando las variaciones de la variable independiente X son directamente proporcionales a las variaciones de la variable dependiente "Y"

Es negativo, cuando las variaciones de la variable independiente "X" son inversamente proporcionales a las variaciones de las variables dependientes "Y"

Es nulo o cero, cuando entre las variables dependientes "Y" e independientes "X" no existen relación alguna.

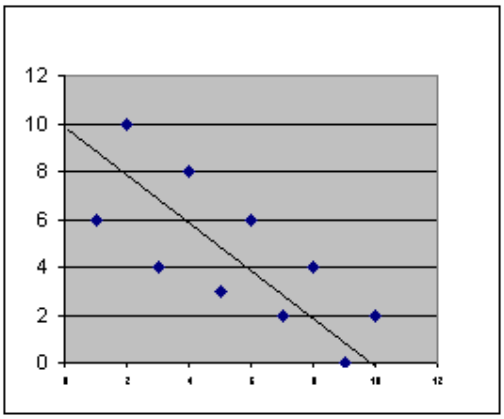


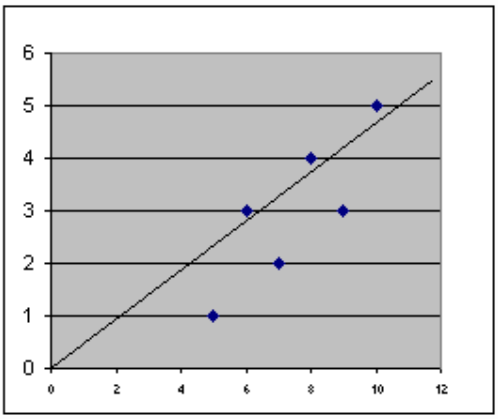



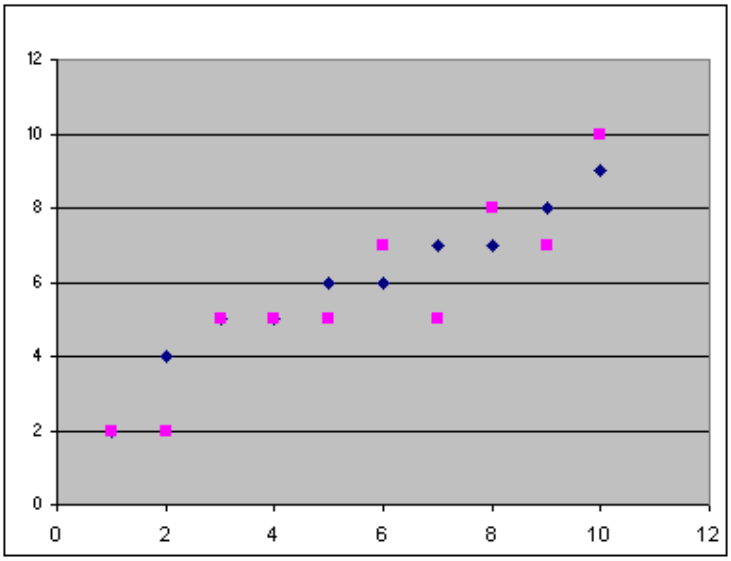



El coeficiente de correlación se puede clasificar de la siguiente manera:

## Graficas de recta de regresión.




Por último se pueden graficar las líneas de tendencia, herramienta muy útil para el mercadeo porque es utilizada para evaluar la resistencia que proyectan los precios. Cuando una línea de tendencia central se rompe ya sea con tendencia al alza o en la baja es porque ocurre un cambio en los precios, por lo tanto las líneas de tendencia

pueden ser alcista cuando se unen los puntos sucesivos y bajista cuando se unen los puntos máximos.

## Procedimiento para hallar el coeficiente de regresión.

Para determinar el valor del coeficiente de regresión de una manera fácil y exacta es utilizando el método de los Mínimos Cuadrados de dos maneras:

**1.- Forma Directa**

De la ecuación de la recta:

$$\boxed{Y = a_0 + a_1 x}$$

Si $a_0$ y $a_1$, se obtienen a partir de las ecuaciones normales:

$$\sum y = a_0 N + a_1 \sum x$$

$$\sum xy = a_0 \sum x + a_1 \sum x^2$$

Aplicando normales Y sobre X tenemos:

$$a_1 = \frac{n\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2}$$

El Coeficiente de Regresión es

$$= a_1$$

De la misma manera la recta de regresión de "X" sobre "Y" será dada de la siguiente manera:

$$\boxed{X = b_0 + b_1 y}$$

Donde: $b_0$ y $b_1$ se obtienen a partir de las ecuaciones normales:

$$\sum x = b_0 N + b_1 \sum y$$

$$\sum xy = b_0 \sum y + b_1 \sum y^2$$

Aplicando normales X sobre Y tenemos:

$$b_0 = \frac{\sum x \sum y^2 - \sum y \sum xy}{N\sum y^2 - \left(\sum y\right)^2}$$

$$b_1 = \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - \left(\sum y\right)^2}$$

El Coeficiente de Regresión es $= b_1$

2.- Forma Indirecta del Método de los Mínimos Cuadrados.

El fundamento de este método es de las desviaciones de X respecto a su media aritmética. X

$$y = \left[\frac{\sum xy}{\sum x^2}\right]x \qquad x = \left[\frac{\sum xy}{\sum y^2}\right]y$$

Ecuación de ***y*** sobre ***x*** Ecuación de ***y*** sobre ***x***

Ecuación de *y* sobre *x* Ecuación de *y* sobre *x*

Donde:

$$x = X - \overline{X}$$

$$y = Y - \overline{Y}$$

X, y = desviaciones

X = media aritmética

Y = media aritmética

b. Regresión Simple: Este tipo se presenta cuando una variable independiente ejerce influencia sobre otra variable dependiente. Ejemplo: Y = f(x)
c. Regresión Múltiple: Este tipo se presenta cuando dos o más variables independientes influyen sobre una variable dependiente. Ejemplo: Y = f(x, w, z).

**Ejemplo regresión:**

La matematización nos da ecuaciones para manipular los datos, como por ejemplo medir la circunferencia de los niños y niñas y que parece incrementarse entre las

edades de 2 meses y 18 años, aquí podemos inferir o predecir que las circunferencias del cráneo cambiara con la edad, en este ejercicio la circunferencia de la cabeza es la respuesta y la edad la variable explicativa.

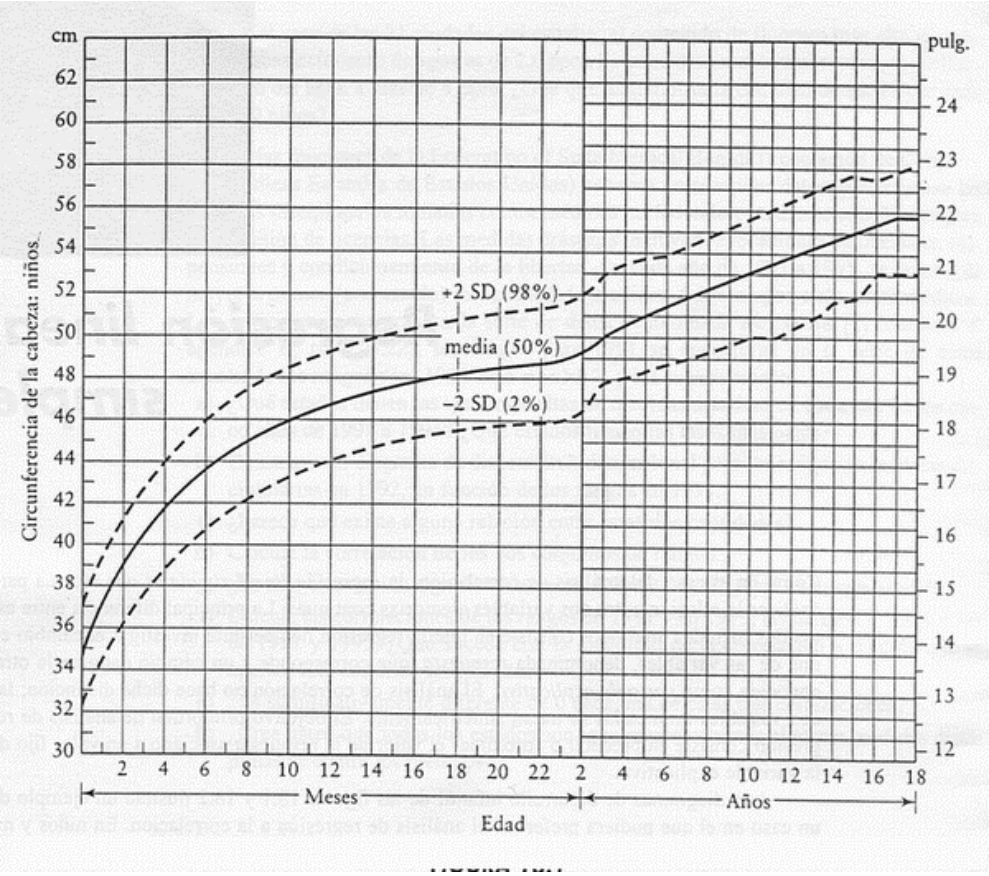
En la regresión tenemos ecuaciones que nos representan las diferentes clases de regresión:

Regresión Lineal: y = A + Bx

Regresión Logarítmica: y = A + BLn(x)

Regresión Exponencial: y = Ac (bx)

Regresión Cuadrática: y = A + Bx +Cx2



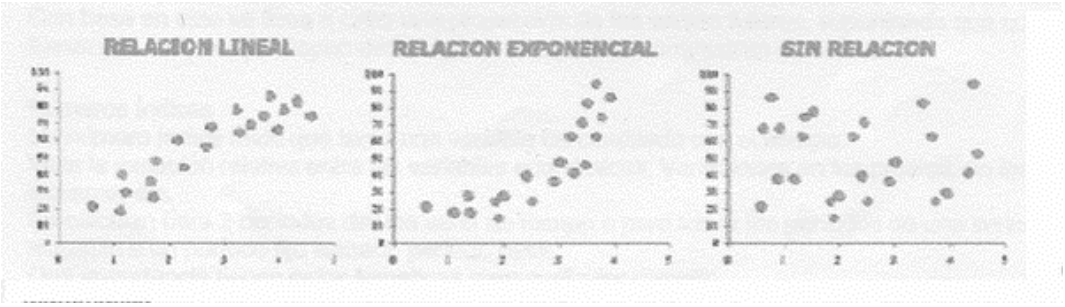

Para obtener un modelo de regresión es suficiente establecer la regresión para eso se hace uso del coeficiente de correlación: R.

R = Coeficiente de correlación, este método mide el grado de relación existente entre dos variables, el valor de R varía de -1 a 1, pero en la práctica se traba con un valor absoluto de R.

El valor del coeficiente de relación se interpreta de modo que a media que R se aproxima a 1, es más grande la relación entre los datos, por lo tanto R (coeficiente de correlación) mide la aproximación entre las variables.

También existen gráficos que representan la dispersión de datos dentro de las coordenadas cartesianas, ósea las nubes de puntos y que pueden darse según la relación que representa, que puede ser lineal, exponencial y sin relación, esta última cuando los puntos están dispersos en todo el cuadro sin agruparse lo cual sugiere que no hay relación.

Los gráficos siguientes nos muestran esta relación:




Matemáticamente las ecuaciones serían:

Ajuste Lineal: Y = Bx + A

Ajuste Logarítmico: Y =BLnX + A

Ajuste Exponencial: Y = AC BX

En el modelo de regresión lineal simple se utiliza la técnica de estimación de los mínimos cuadrados, este modelo tiene solo una variable de predicción y se supone una ecuación de regresión lineal.

## Análisis de regresión múltiple.

Dispone de una ecuación con dos variables independientes adicionales:

$$Y' = a' + b_1 x_1 + b_2 x_2$$

Se puede ampliar para cualquier número "m" de variables independientes:

$$Y' = a' + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_m x_m$$

Para poder resolver y obtener $a, b_1$ y $b_2$ en una ecuación de regresión múltiple el cálculo se presenta muy tediosa porque se tiene atender 3 ecuaciones que se generan por el método de mínimo de cuadrados:

$$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2$$

$$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

Para poder resolver se puede utilizar programas informáticos como AD+, SPSS y Minitab y Excel.

El error estándar de la regresión múltiple $(S_{xy})$

Es una medida de dispersión la estimación se hace más precisa conforme el grado de dispersión alrededor del plano de regresión se hace mas pequeño.

Para medirla se utiliza la formula:

$$\boxed{S_{xy} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - m - 1}}}$$

Y: Valores observados en la muestra

$\hat{Y}$ : Valores estimados a partir a partir de la ecuación de regresión

n : Número de datos

m : Número de variables independientes

El coeficiente de determinación múltiple $(r^2)$

Mide la tasa porcentual de los cambios de Y que pueden ser explicados por $x_1$, $x_2$ y $x_3$ simultáneamente.

$$r^2 = \frac{SCregresiòn}{SCTotal}$$

Lo que buscamos es construir un modelo para determinar la dependencia que exista de aprendizaje reflejada en las notas de la asignatura de PHP, conociendo las notas de las asignaturas Algoritmos, Base de Datos y Programación.

Se presentara la siguiente ecuación a resolver:

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3$$

Utilizando las formulas de las ecuaciones normales a los datos obtendremos los coeficientes de regresión o utilizando Regresión de Análisis de datos, en la Hoja de Cálculo de Excel podemos calcular también los coeficientes de regresión:

| | *Coeficientes* | *Error típico* | *Estadístico t* | *p-valor* | *Inferior 95%* | *Superior 95%* |
|---|---|---|---|---|---|---|
| Intercepción | 2.551 | 2.369 | 1.077 | 0.305 | -2.663 | 7.766 |
| Algoritmos | 0.583 | 0.267 | 2.186 | 0.051 | -0.004 | 1.169 |
| Base de Datos | 0.373 | 0.144 | 2.589 | 0.025 | 0.056 | 0.691 |
| Programación | -0.242 | 0.270 | -0.893 | 0.391 | -0.837 | 0.354 |

Por lo tanto podemos construir la ecuación de regresión que buscamos:

$$Y = 2.551 + 0.583x_1 + 0.373x_2 - 0.242x_3$$

$$S_{yx} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-m-1}}$$

$$r^2 = \frac{18.7737874}{26.9333333} = 0.69704656$$

## El error estándar de estimación.

Un error estándar de estimación se da cuando en el diagrama de dispersión no todos los puntos se encuentran en la recta de regresión

El error estándar de la estimación mide la dispersión de los valores observados alrededor de la recta de regresión.

Fórmulas usadas para calcular el error estándar:

$$S_{Y \cdot X} = \sqrt{\frac{\Sigma (Y - Y')^2}{n - 2}}$$

$$= \sqrt{\frac{\Sigma Y^2 - a(\Sigma Y) - b(\Sigma XY)}{n - 2}}$$

## Suposiciones de la regresión lineal.

- Para dar cada valor x, existen valores, y correspondientes, los cuales siguen una distribución normal.
- En la recta de regresión se encuentran ubicadas las medias de las distribuciones normales.
- Todas las desviaciones estándar de estas distribuciones normales son iguales.
- Lao valores de Y son estadísticamente independientes.

En el caso en que nuestras observaciones sean una muestra aleatoria proveniente de una población, estaremos interesados en realizar inferencias sobre la misma. A fin de que estas inferencias sean "estadísticamente razonables", se han de cumplir las siguientes condiciones:

1. En la población, la relación entre las variables X e Y debe ser aproximadamente lineal, i.e.: y = β + β x +ε 1 2, siendo ε la v.a. que representa los residuos (diferencias entre el valor estimado por el modelo y el verdadero valor de Y).

2. Los residuos se distribuyen según una Normal de media 0, i.e., ε ≈ N (0, σ 2).

3. Los residuos son independientes unos de otros.

4. Los residuos tienen varianza σ 2 constante.

Afortunadamente, el modelo de regresión lineal es bastante "robusto", lo que significa que no es necesario que las condiciones anteriores se cumplan con exactitud (en particular las tres últimas).

## Correlación y regresión lineal simple.

La correlación es utilizada para determinar la confiabilidad y validez de pruebas o instrumentos de medición. Tema que se considera el material preparado por este servidor sobre estadística multivalente.
La correlación, la regresión y la predicción una vez que se conoce la correlación entre dos variables es posible predecir el valor que le correspondería a un caso en una de las variables, si conoce su calificación o valor en la otra variable. (Recuérdese que se supone que las variables tienen una correlación lineal).
Su pude decir que cuando dos variables están relacionadas entre sí, dichas variables se encuentran una en función de la otra. El término función, se utilizó por primera vez en

1692 en un artículo sobre matemática publicado en el Acta Eruditorem y se atribuyó a Gottfried von Leibniz, en 1749 Leonard Euler lo definió como *la cantidad de una variable que es dependiente de otra cantidad* y más tarde Lejeune Dirichlet (1837) propuso que desde el punto de vista matemático la función es la correspondencia que asigna un valor único a la variable dependiente para cada valor permitido en una variable independiente.

**Ejemplo:**

Cuando compramos harina la pagamos por su peso (masa), podemos decir que el precio de la harina está en función de su peso (masa). Por ejemplo, si su precio es de .50 centavos de dólar por kilo, dos kilos costarán un dólar, medio kilo veinticinco centavos, diez kilos cinco dólares, etc. Esto se representa así:

$$f(x) = .50\,x$$

En este caso f(x) está definido sólo para que x ³ 0.
De manera que una función es una asociación entre dos o más variables, en la cual a cada valor de cada una de las variables independientes o argumentos, corresponde exactamente un valor de la variable de dependiente en un conjunto denominado específicamente, dominio de la función.
La función de una variable puede ser escrita f (x) lo cual se lee " f de x" o de manera más completa "el valor de la función f en x". De manera que si *y* es una variable que está en función de *x,* se acostumbra escribir la variable dependiente en el lado izquierdo del signo de igualdad en la ecuación, así:

$$y = x + 2 \text{ que es lo mismo que decir } f(x) = x + 2$$

Según la expresión anterior, tanto y como f(x) es la variable dependiente, siendo x la variable independiente.
También podría ser la expresión que aparece a continuación:

$$x = y + 2 \text{ que es lo mismo que decir } f(y) = y + 2$$

Nótese que en este caso, tanto x como y pueden intercambiar su papel como variable dependiente e independiente. Esto sucede así porque cuando se trata de la relación entre dos variables, los términos dependiente o independiente no trascienden el simple significado de determinar cuál variable es la que se pretende predecir en contraste con la variable que se utiliza para hacer la predicción. En otras palabras, la existencia de la relación, no permite por sí misma determinar que una variable se dependiente (efecto) de la independiente (causa). No hay lugar para la determinación

de causalidad. Esa es la razón por la cual autores como Kachigan (1991) prefieren utilizar los términos, variable criterio y variable (s) predictiva (s)

Antes de cerrar este tema de las correlaciones, vamos a señalar tres asuntos muy importantes y útiles.

1. Un coeficiente de correlación no muestra el porcentaje de relación entre dos variables como algunas veces se piensa. Sin embargo, el valor que se obtiene al elevar al cuadrado el coeficiente de correlación entre dos variables, muestra el porcentaje de la variabilidad (varianza) de una de las variables, que puede ser atribuido a la variabilidad (varianza) de la otra. De manera que ese cuadrado ofrece una aproximación del porcentaje de relación que existe entre las dos variables.

(Para el ejemplo que hemos tenido en esta sección, correspondería decir que la correlación de .82 puede estar mostrando aproximadamente un 67% de relación entre las dos variables observadas).

2. Los coeficientes de correlación no son directamente proporcionales. Como por ejemplo, en el caso de nuestra correlación .82 muestra más que el doble de relación que una correlación de .41. Se puede obtener una aproximación del valor relativo de estas dos correlaciones si elevamos al cuadrado ambos coeficientes y luego dividimos el más grande entre el más pequeño. Por ejemplo (.82)2 = .67, (.41)2 = .1681. De manera que al dividir .67 entre .1681 tenemos 3.98 lo que significa que la correlación .82 manifiesta cuatro veces más relación que la correlación.

3. Cuando dos variables muestran una correlación positiva no necesariamente significa que tienen una relación tan alta como lo muestra el coeficiente de correlación. Siendo que ellas pueden estar relacionadas con otra (s) variable (s) o factor (es) comunes. Puede que se usted observe una relación significativa entre la satisfacción en el trabajo y el grado de motivación del maestro, pero también hay que recordar que estas dos variables están relacionadas también con ciertas características del individuo como puede ser su nivel de inteligencia emocional. De manera que si podemos aislar el aporte que hace la inteligencia emocional en la variabilidad de la satisfacción y de la motivación, notaremos que la correlación entre estas últimas será menor. Esta técnica es conocida como correlación parcial y permite estimar la correlación residual entre dos variables habiendo retirado el efecto de una o más variables que intervienen.

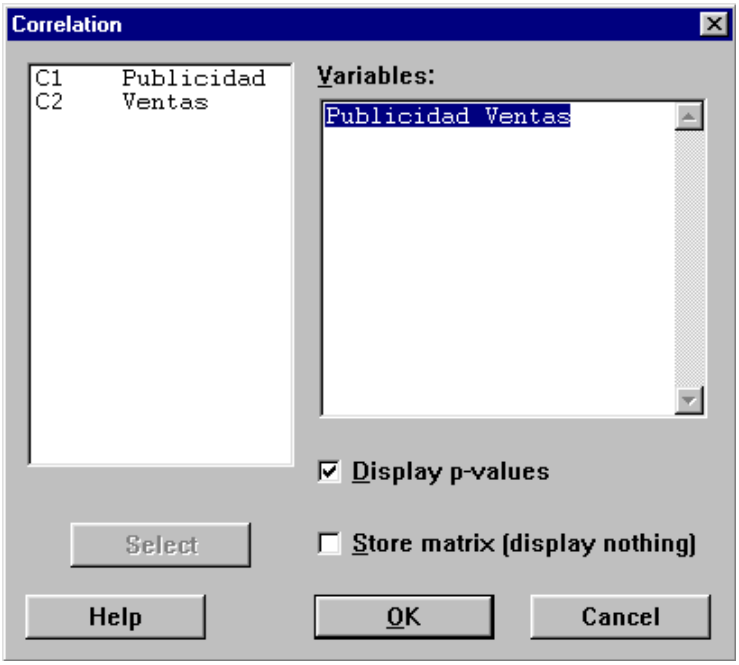
## Casos prácticos con software.

En la siguiente tabla, se muestran los datos registrados de las ventas en millones de euros y de los gastos incurridos en publicidad, también en millones de euros, por una empresa industrial que fabrica sillas abatibles para oficina:

| Gtos de publicidad (millones euros) (X) | Volumen de ventas (millones euros) (Y) |
|---|---|
| 14,2226 | 95,065 |
| 13,9336 | 97,281 |
| 15,5040 | 103,159 |
| 16,3105 | 107,607 |
| 17,4936 | 113,860 |
| 19,8906 | 121,153 |
| 21,4803 | 129,102 |
| 20,4046 | 132,340 |
| 21,4776 | 138,663 |
| 22,6821 | 142,856 |
| 20,9722 | 143,120 |
| 23,3538 | 147,928 |
| 26,1040 | 155,955 |
| 29,1101 | 164,946 |
| 27,2418 | 163,921 |
| 23,0096 | 163,426 |
| 27,6116 | 172,485 |
| 32,1111 | 180,519 |
| 36,1788 | 190,509 |
| 37,5671 | 196,497 |
| 33,5069 | 196,024 |
| 36,6088 | 200,832 |
| 31,1554 | 196,769 |
| 32,7752 | 205,341 |
| 41,1886 | 220,230 |
| 39,9715 | 228,703 |
| 39,6866 | 236,500 |
| 40,2991 | 244,560 |
| 40,9538 | 254,771 |
| 41,9323 | 263,683 |
| 39,8393 | 268,304 |

a) Calcular el coeficiente de correlación lineal entre las variables ventas y gastos de publicidad.

- Seleccionamos *Stat > Basic Statistics > Correlation* :



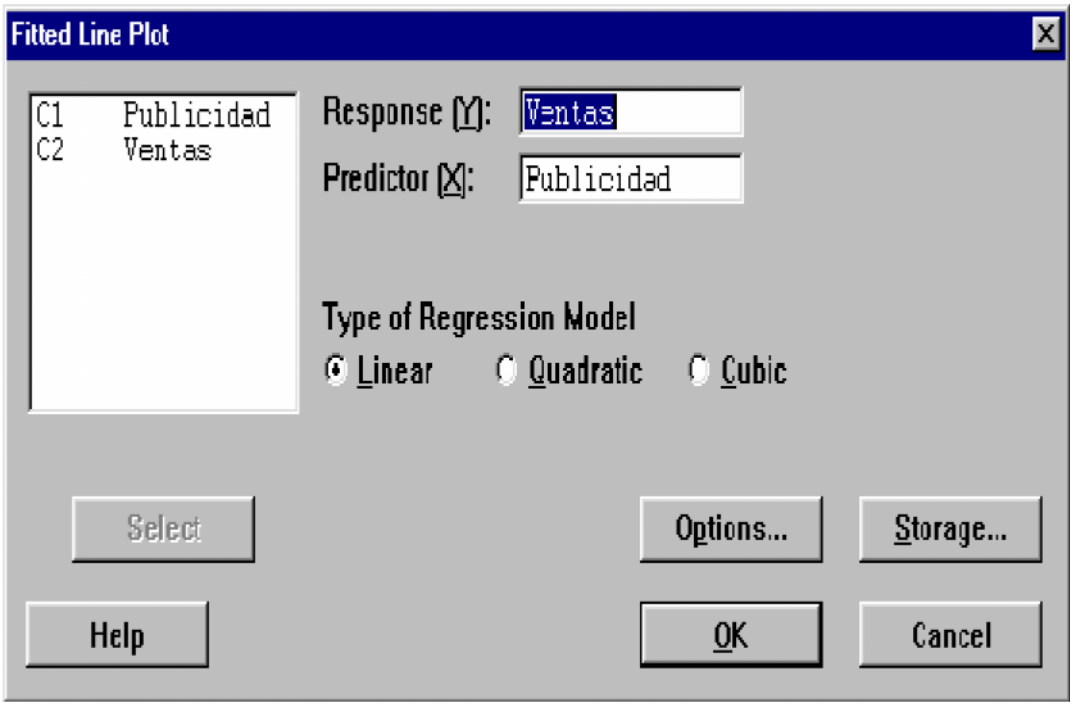
Correlations (Pearson)

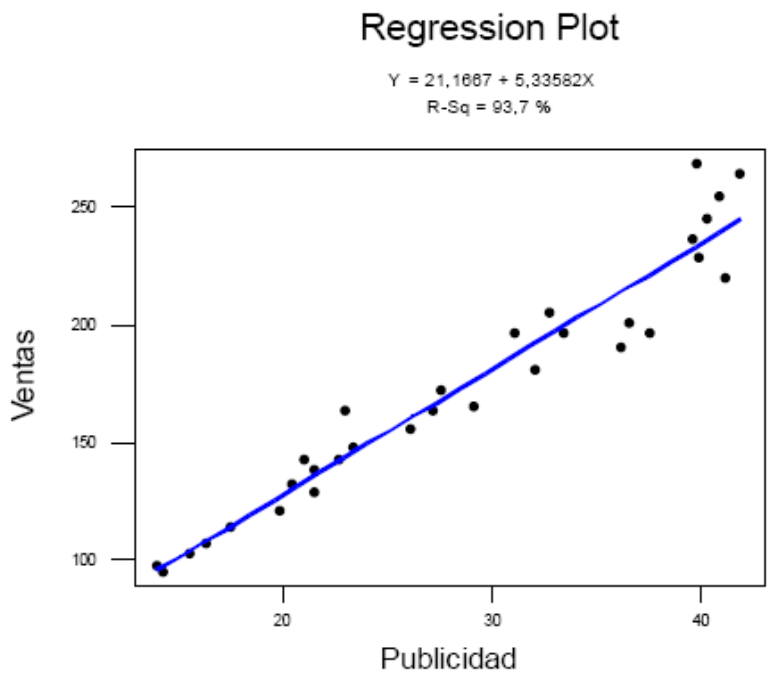
Correlation of Publicity y vents = 0.973, P-Value = 0.000

El valor obtenido para el coeficiente de correlación es de 0,973, lo cual hace suponer que, en principio, la correlación entre ambas variables es muy alta por lo que se puede prever que en la regresión obtendremos un buen ajuste.

b) Representar la nube de puntos (gráfico x-y) ventas vs. Publicidad, junto con la recta de regresión asociada. ¿Piensas que el modelo obtenido sirve para explicar las ventas obtenidas por esta empresa en los últimos treinta años en función de lo que se ha gastado en publicidad?

Seleccionamos Stat > Regression > Fitted Line Plot






```
Regression

The regression equation is
y = 21,2 + 5,34 x

Predictor        Coef       StDev          T        P
Constant       21,167       7,687       2,75    0,010
x              5,3358      0,2568      20,78    0,000

S = 12,94       R-Sq = 93,7%     R-Sq(adj) = 93,5%
```

Como se aprecia en el gráfico, el modelo lineal simple ajusta con mínimos errores la evolución de las ventas en función de los gastos en publicidad. De hecho, si nos fijamos en el valor del coeficiente de determinación R-sq, veremos que este modelo explica casi el 94% del comportamiento de las ventas a través de la evolución, por lo que es un buen ajuste y por tanto, los residuos son mínimos.

c) ¿Presenta la muestra suficiente evidencia, a un nivel de significación de 0,05, como para rechazar la hipótesis nula sobre la pendiente (H0: pendiente de la recta es cero)?

En el output anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es casi cero. Como este valor es menor que $\alpha = 0{,}05$, debemos rechazar la hipótesis nula, i.e., concluiremos que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo tiene sentido).

**EJERCICIOS PROPUESTOS**

1. La información estadística obtenida de una muestra de tamaño 12 sobre la relación existente entre la inversión hecha y el rendimiento obtenido en miles de euros para explotaciones agropecuarias se muestra la tabla siguiente:

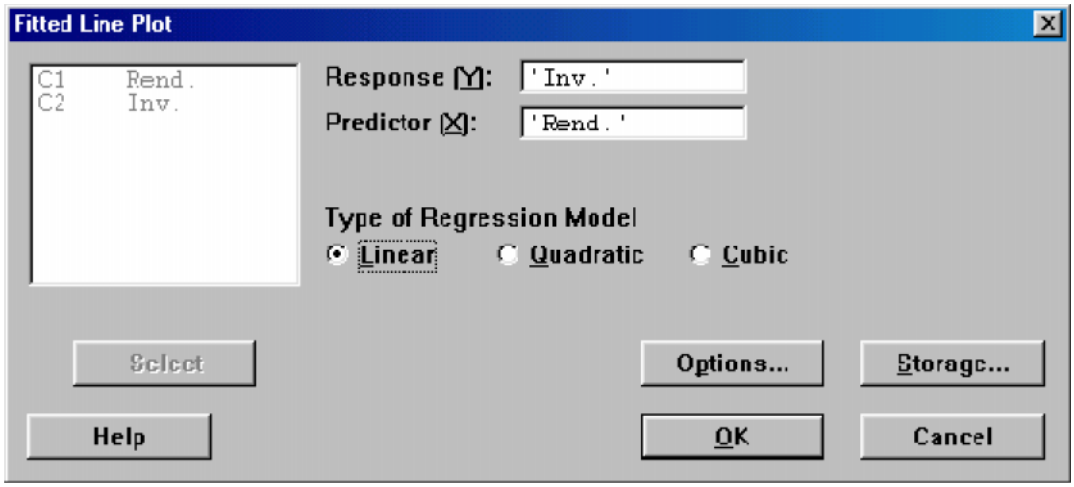
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Inv** | 11 | 14 | 16 | 15 | 16 | 18 | 20 | 31 | 14 | 20 | 19 | 11 |
| **Rend.** | 2 | 3 | 5 | 6 | 5 | 3 | 7 | 10 | 6 | 10 | 5 | 6 |

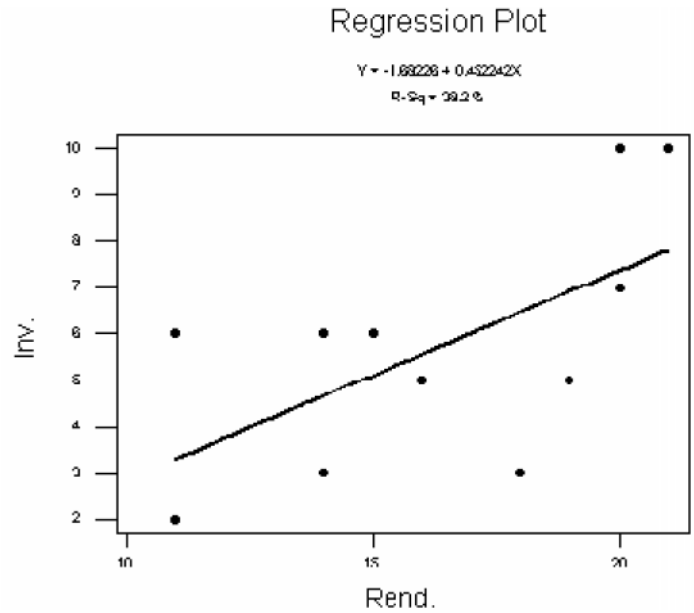
a) Calcula el coeficiente de correlación lineal, así como la recta de regresión. Calcula además, la previsión de inversión que se obtendrá con un rendimiento de 8000 €
Seleccionamos Stat > Basic Statistics > Correlation y obtenemos:

```
Correlations (Pearson)

Correlation of Rend. and Inv. = 0.618, P-Value = 0.032
```

Como el coeficiente de correlación lineal es 0.618 no podemos deducir que exista una relación fuerte ni débil, tendríamos que realizar un contraste de hipótesis para saberlo con claridad.
Calculemos ahora la recta de regresión. Para ello, seleccionaremos *Stat > Regression > Fitted Line Plot*:






A partir de este gráfico, observamos que no existe ninguna correlación entre las dos variables.

Para conocer más detalles, seleccionamos *Stat > Regression > Regression*

:

**Regression Analysis**

The regression equation is
Inv. = - 1.68 + 0.452 Rend.

| Predictor | Coef | StDev | T | P |
|---|---|---|---|---|
| Constant | -1.682 | 3.015 | -0.56 | 0.589 |
| Rend. | 0.4522 | 0.1819 | 2.49 | 0.032 |

S = 2.060 R-Sq = 38.2% R-Sq(adj) = 32.0%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 26.230 | 26.230 | 6.18 | 0.032 |
| Residual Error | 10 | 42.437 | 4.244 | | |
| Total | 11 | 68.667 | | | |

Así pues, la recta de regresión será:

Inv.=-1.68 + 0.452*Rend

Por tanto, para obtener un rendimiento de 8000 €, tendríamos que hacer una inversión de... Inv. = -1.68 + 0.452*8000 = 3614.32 €

b) ¿Presenta la muestra suficiente evidencia, a un nivel de significación de 0,05, como para rechazar la hipótesis nula sobre la pendiente (H0: pendiente de la recta es cero)?

En el output anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es 0,032. Como este valor es menor que α = 0,05, debemos rechazar la

hipótesis nula, i.e., concluiremos que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo tiene sentido).

**RELACIÓN DE EJERCICIOS Y PROBLEMAS SOBRE REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE**

1. Sobre un conjunto de conductores se ha realizado una encuesta para analizar su edad (Y) y el número de accidentes que han sufrido (X). A partir de la misma, se obtuvieron los siguientes resultados:

| X/Y | (20,30] | (30,40] | (40,50] | (50,60] | (60,70] |
|---|---|---|---|---|---|
| **0** | 74 | 82 | 78 | 72 | 7 |
| **1** | 7 | 6 | 5 | 6 | 5 |
| **2** | 3 | 2 | 2 | 1 | 1 |

A partir de estos datos, se le pide que determine para esta distribución las curvas de regresión de Y sobre X y de X sobre Y.

2. Para la economía española, disponemos de los datos anuales redondeados sobre consumo final de los hogares a precios corrientes (Y) y renta nacional disponible neta (X), tomados de la Contabilidad Nacional de España base 1995 del INE , para el período 1995-2002, ambos expresados en miles de millones de euros:

| Año | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|---|---|
| $Y_t$ | 258'6 | 273'6 | 289'7 | 308'9 | 331'0 | 355'0 | 377'1 | 400'4 |
| $X_t$ | 381'7 | 402'2 | 426'5 | 454'3 | 486'5 | 520'2 | 553'3 | 590'0 |

Considerando que el consumo se puede expresar como función lineal de la renta ($Yt=a+b\cdot Xt$), determine:
a) Los parámetros a y b de la recta de regresión.
b) La varianza de la variable consumo y su descomposición en varianza explicada y no explicada por el modelo.
c) El coeficiente de determinación de dicha regresión.
d) La predicción del valor que tomará el consumo para una renta de 650.000 millones de euros.

3. Se supone que se puede establecer cierta relación lineal entre las exportaciones de un país y la producción interna de dicho país. En el caso de España, tenemos los datos anuales (expresados en miles de millones de pesetas) para tales variables correspondientes al quinquenio 1992-96 en la siguiente tabla:

| Años | Producción | Exportaciones |
|---|---|---|
| 1992 | 52.654 | 10.420 |
| 1993 | 53.972 | 11.841 |
| 1994 | 57.383 | 14.443 |
| 1995 | 61.829 | 16.732 |
| 1996 | 65.381 | 18.760 |

A partir de tal información, y considerando como válida dicha relación lineal, se pide:
a) Si la producción para el año 1997 fue de 2.210.6100 millones de pesetas, ¿cuál sería la predicción de las exportaciones para este año? ¿Qué grado de precisión tendría dicha predicción? Regresión Lineal Simple y Múltiple
b) Si sabemos que las exportaciones para 1997 fueron de 69.045.704 millones de pesetas, ¿cuál sería la producción interna aproximada para ese año? ¿Qué grado de confianza daría usted a esta predicción?
c) ¿Qué tanto por ciento de la varianza de las exportaciones no vienen explicadas por la producción interna, y se debe a otro tipo de variables?

4. De una distribución de dos variables se conocen los siguientes datos:

$$r_{xy} = 0'9;\ s_x = 1'2;\ s_y = 2'1;\ \bar{x} = 5;\ \bar{y} = 10$$

A partir de los mismos, obténganse las rectas de regresión mínimo cuadráticas de X sobre Y y de
Y sobre X.

5. Para un mismo grupo de observaciones de las variables X e Y, hemos obtenido las dos rectas de regresión siguientes:

$$3x + 2y = 26$$
$$6x + 2y = 32$$

En función de las mismas, responda a las siguientes cuestiones:
a) ¿Qué valores tomarían las medias de X e Y?
b) Represente gráficamente ambas rectas de regresión.
c) Determine el valor del coeficiente de correlación lineal rxy.
d) ¿Porqué la regresión de Y sobre X y la de X sobre Y no coinciden?

6. A partir de un conjunto de valores de las variables X e Y, se ha determinado la regresión de Y sobre X, obteniéndose la siguiente recta:

$$Y^* = 10 + 0'45X; \qquad R^2 = 0'9; \qquad \bar{x} = 20$$

Se pide que, a partir de la definición de la anterior recta, determine los parámetros de la recta de regresión de X sobre Y.

7. Se han observado, en varios modelos de automóviles, los datos de potencia del motor (X), en caballos, y la aceleración (Y), medida en el número de segundos necesarios para acelerar de 0 a 100 Km./h. La tabla adjunta refleja los valores obtenidos.

| | | | | | | |
|---|---|---|---|---|---|---|
| **X** | 50 | 75 | 90 | 100 | 120 | 150 |
| **Y** | 15 | 12 | 10'5 | 10 | 9 | 8 |

A partir de tales datos, se ha decidido expresar la aceleración en función de la potencia mediante el ajuste de una función potencial (mediante el correspondiente cambio logarítmico).
Bajo esta hipótesis:
a) Determine la función de ajuste y el ECM conseguido.
b) Si aumenta la potencia de un motor en un 10%, ¿en qué porcentaje repercutirá dicho aumento en la aceleración prevista? ¿Depende ello de la potencia que tenga el motor en cuestión?

8. En un nuevo proceso artesanal de fabricación de cierto artículo que está implantado, se ha considerado que era interesante ir anotando periódicamente el tiempo medio (medido en minutos) que se utiliza para realizar una pieza (variable Y) y el número de días desde que empezó dicho proceso de fabricación (variable X). Con ello, se pretende analizar cómo los operarios van adaptándose al nuevo proceso, mejorando paulatinamente su ritmo de producción conforme van adquiriendo más experiencia en él. A partir de las cifras recogidas, que aparecen en la tabla adjunta, se decide ajustar una función exponencial que explique el tiempo de fabricación en función del número de días que se lleva trabajando con ese método.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **X** | 10 | 20 | 30 | 40 | 50 | 60 | 70 |
| **Y** | 35 | 28 | 23 | 20 | 18 | 15 | 13 |

Desde el correspondiente ajuste propuesto, se pide que determine:
a) ¿Qué tiempo se predeciría para la fabricación del artículo cuando se lleven 100 días?
b) ¿Qué tiempo transcurriría hasta que el tiempo de fabricación que se prediga sea de 10 minutos?
c) ¿Qué porcentaje de tiempo se reduce por cada día que pasa?

9. La empresa COLOBONA S. A .L. ha trabajado hasta ahora con la hipótesis de que las ventas de un período dependen linealmente de los gastos de publicidad efectuados en el período anterior. En este momento, le solicitan a usted la realización de un análisis que ponga de manifiesto si la hipótesis, hasta ahora mantenida, se puede seguir sosteniendo en función de los datos que le suministran.

| AÑOS | GASTOS | VENTAS |
|---|---|---|
| 1987 | 21 | ... |
| 1988 | 22 | 19 |
| 1989 | 25 | 20 |
| 1990 | 26 | 22 |
| 1991 | 27 | 23 |
| 1992 | 29 | 24 |
| 1993 | 30 | 26 |

En el informe final de su análisis, deberá responder a las siguientes preguntas:

a) ¿Se incrementarán las ventas del período siguiente al aumentar los gastos en publicidad del período actual?

b) ¿Es adecuado suponer que el ajuste entre estas variables es efectivamente lineal teniendo en cuenta los valores de las variables? Ajuste el modelo lineal e intérprete los coeficientes del mismo. ¿Qué porcentaje de la varianza de las ventas no son explicadas por las variaciones de los gastos en publicidad?

c) ¿Cuál será la predicción de las ventas para 1994? ¿Qué precisión tendrá ese pronóstico?

d) Si para el año 1994 se piensa incrementar los gastos de publicidad en un 10%, ¿qué incremento relativo cabría esperar para las ventas de 1995 con respecto a las de 1994, según el modelo ajustado?

# Bibliografía.


Probabilidad y estadísticas para ingeniería y ciencias". Jay L. Devore. 1998.
Grinstead, Charles M.; Snell, J. Laurie (1997). «9. Central Limit Theorem», *Introduction to Probability*
http://es.wikipedia.org/wiki/Teorema_del_límite_central
http://es.wikipedia.org/wiki/Intervalo_de_confianza
http://maralboran.org/wikipedia/index.php/Estimaci%C3%B3n_por_intervalos_de_confianza_de_medias_y_proporciones
http://descartes.cnice.mec.es/materiales_didacticos/Muestreo_Inferencia_Estadistica/estimacion_intervalos.html
http://www.ite.educacion.es/w3/eos/MaterialesEducativos/mem2001/estadistica/estimacion.htm
http://www.virtual.unal.edu.co/cursos/odontologia/2002890/lecciones/estimacion/estimacion.htm
http://www.gestiopolis.com/finanzas-contaduria/estadistica-intervalos-de-confianza.htm
http://www.iesxunqueira1.com/Download/pdf/teointervalos.pdf
http://thales.cica.es/rd/Recursos/rd97/UnidadesDidacticas/28-1-u-i.html
http://www.hiru.com/es/matematika/matematika_06600.html
www.estadistica-aplic.com.ar
http://www.google.com.ec/search?hl=es&source=hp&q=pruebas+cuando+se+tienen+dos+poblaciones&aq=f&aqi=g1&aql=&oq=&gs_rfai=
http://www.google.com.ec/search?hl=es&q=ejercicios+propuestos+de+pruebas+cuando+se+tienen+dos+poblaciones&aq=o&aqi=&aql=&oq=&gs_rfai=
http://www.itch.edu.mx/academic/industrial/estadistica1/cap03d.html
http://fltbw2.rug.ac.be/iloapp/Applets/Ap6b.html