**UNIVERSIDAD BOLIVARIANA DE VENEZUELA**
**DIRECCIÓN GENERAL ACADEMICA**
**DIRECCIÓN DE PLANIFICACION Y DESARROLLO CURRICULAR**

# ANÁLISIS DEL DATO ESTADÍSTICO II

**GUIA DIDACTICA**

**Por**
**Profesores**
**HERRERA, ROBERTO. PEÑA, CAROLINA,**
**GODOY, MANUEL Y BRICEÑO, GIOCONDA**
**DEL PFG EN GESTIÓN AMBIENTAL-UBV, SEDE CARACAS,**

**CARACAS, SEPTIEMBRE DE 2005**

# INDICE

**ANÁLISIS DEL DATO ESTADÍSTICO**
**GUIA DIDACTICA**

## PRESENTACIÓN

La Guía Didáctica de **"Análisis del Dato Estadístico II "** está diseñada para que los profesores - facilitadores de la UBV contribuyan al logro de la competencia en el uso de las herramientas estadísticas y generar bases que apoyen la toma de decisiones en proyecto..

La estadística Inferencial es una herramienta fundamental para la formulación, ejecución, supervisión, control de calidad y toma de decisiones de cualquier proyecto socio comunitario y de investigación. El egresado de la UBV tiene que manejar con soltura un leguaje estadístico revestido con una gran sencillez y comprensible, que permita una fluida comunicación dentro de un grupo trabajo interdisciplinario y que al mismo tiempo pueda apoyar la resolución de una gran cantidad de situaciones que requieran el estudio de un conjunto de datos para su mejor compresión y aporte de soluciones.

Esta guía didáctica expone de manera sencilla, elementos de probabilidad y de estadística Inferencial.

La sencillez de esta guía no disminuye su validez didáctica, apta para todo aquel que se inicia en el estudio de la estadística Inferencial. Esta herramienta es indispensable para los proyectos socio comunitarios y de investigación que aspiren a tener base cuantitativa, pues un proyecto sin datos estadísticos y su interpretación e inferencia presenta una gran debilidad al momento de tomar decisiones.

Esta unidad curricular ha sido diseñada de manera de suministrar una herramienta de utilidad, la cual apoyada en la antropogogía como estrategia didáctica impulsará el trabajo autónomo, responsable y participativo de los alumnos, en la ejecución eficaz del diagnóstico integral sociocomunitario que se realiza durante el Proyecto.

La estructura de la guía consta de 5 Capítulos

CAPITULO I Probabilidad básica y sus distribuciones

CAPITULO II: Prueba de hipótesis e intervalos de confianza

CAPITULO III Análisis de varianza

CAPITULO IV Regresión y correlación

CAPITULO V Pruebas no paramètricas

# GUÌA DIDÁCTICA DE LA UNIDAD CURRICULAR ANÀLISIS DEL DATO ESTADÍSTICO II

Por
HERRERA, Roberto. PEÑA, Carolina, GODOY, Manuel y BRICEÑO, Gioconda del PFG en Gestión Ambiental-UBV, Sede Caracas,
.

## CAPITULO 1

## TEMA 1: PROBABILIDAD BASICA Y SUS DISTRIBUCIONES

**OBJETIVO:** El estudiante comprende la importancia del uso de las probabilidades y distribuciones en la estadística y sus aplicaciones.

**COMPETENCIAS A LOGRAR**:

1. Comprende las definiciones básicas como son: Probabilidad, Diferentes tipos de probabilidad, diferentes distribuciones de probabilidades: Normal, Poisson, entre otras.
2. Comprende la importancia de obtener buenos resultados en las probabilidades y sus distribuciones.
3. Comprende cada uno de los procesos que involucra una investigación estadística inferencial.
4. Define la fuente y afina las herramientas para el manejo de la probabilidad y sus distribuciones en las necesidades y complejidad de los distintos escenarios.
5. Diseña el instrumento para la recolección de datos.
6. Comprende la Importancia de la probabilidad y sus distribuciones en la vida diaria y en la formulación, ejecución y validación de proyectos.

## CONTENIDOS

### 1. ¿Qué es la probabilidad?

Es el número al que tiende la frecuencia relativa de un suceso que esta asociada al número de veces que se realiza el experimento.

Por definición, entonces, la probabilidad se mide por un número en que se localiza entre cero y uno: Si un suceso *no ocurre nunca*, su probabilidad asociada es *cero*, mientras, que *si ocurriese* siempre su probabilidad sería igual a *uno*.

Así, las probabilidades suelen venir expresadas como decimales, fracciones o porcentajes.

Su fórmula se expresa de la siguiente manera:

$$P = \frac{f}{n}$$ ; Donde **f**: número de casos favorables

**n**: número de casos posibles o realizados

**Variable:** Es una característica de interés acerca de cada elemento de una población o una muestra. Las variables en las probabilidades se clasifican de la siguiente manera:

**Variable independiente:** es aquella propiedad de un fenómeno a la que se le va a evaluar su capacidad para influir, incidir o afectar a otras variables.

**Ejemplo:** El tiempo transcurrido en una hectárea de árboles en la zona de Las Delicias del Estado Aragua.

**Variable dependiente:** puede ser definida como los cambios sufridos por los sujetos como consecuencia de la manipulación de la variable independiente por parte del experimentador.

**Ejemplo:** La crecimiento de la población del gusano de palma en el Parque del Este en Caracas en los meses de Octubre- Diciembre del año 2003.

**Variable aleatoria:** variable que toma diferentes valores como resultado de un experimento aleatorio.

**Variable aleatoria continua:** variable aleatoria que puede tomar infinitos valores dentro de un rango cualquiera.

**Variable aleatoria discreta:** variable que toma un número finito o infinito de valores numerables.

**¿Qué es un experimento?**

Es todo proceso que produce un resultado u observación.

**¿Qué es un Espacio Muestral?**

Es el conjunto de todos los resultados posibles de un evento ò proceso se simboliza por letras mayúsculas, S = {árboles, ríos, casas,}, también se pueden representar en cuadros o formatos ya elaborados.

**¿Qué es un Evento?**

Es cualquier subconjunto del espacio muestral.

**Ejemplo:** En la UBV, se clasificó a cada estudiante de acuerdo con años y sexo. Los resultados se resumen en la siguiente tabla:

| ***Años*** \ *varones y hembras* | **Varones (V)** | **Hembras (H)** | **TOTAL** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 2002 (1er año) | 100 | 80 | **180** |
| 2003 (2do año) | 70 | 50 | **120** |
| 2004 (3er año) | 50 | 40 | **90** |
| **TOTAL** | **220** | **170** | **390** |

Determine la probabilidad de que al seleccionar el estudiante al azar sea:

1. De segundo año **$P_1$**
2. Hembra **$P_2$**
3. De tercer año **$P_3$**
4. Varón **$P_4$**

**Solución:**
Se procede de la siguiente manera:

Los eventos son: **f:** el estudiante elegido sea de segundo año.
**n:** el estudiante elegido sea hembra.

Se calcula las cuatro probabilidades a la vez**:**

$p_1 = \frac{120}{390} = 0,3076$ **$P_1$= 0,3076; $P_1$= 30,76%**

$P_2 = \frac{170}{390} = 0,4358$ **$P_2$ = 0,4358; $P_2$ = 43,58%**

$P_3 = \frac{90}{390} = 0,2307$ **$P_3$ = 0,2307; $P_2$ = 23,07%**

$P_4 = \frac{220}{390} = 0,5641$ **$P_4$ = 0,5641; $P_4$ = 56,41%**

**Interpretación:** Esto quiere decir que:

1. Existe el 0,3076 ó 30,76 % de probabilidad de que sea de segundo año,
2. El 0,4358 ó 43,58% de probabilidad de que sea hembra.
3. Existe el 0,2307 ó 23,07 % de probabilidad de que sea de tercer año,
4. El 0,5641 ó 56,41% de probabilidad de que sea varón.

**PROPIEDADES DE LA PROBABILIDAD:**

Si el evento no puede ocurrir, su probabilidad es igual a cero.
Si el evento ocurre, entonces su probabilidad es igual a uno.

Esto es, $\mathbf{0 \leq P(A) \leq 1}$

En otra forma $\mathbf{P(A) = 0}$
$\mathbf{P(A) = 1}$

**Ejemplo:** En los archivos de una clínica médica se han clasificado pacientes por su sexo y tipo de diabetes (I o II). Los grupos se exhiben a continuación. El cuadro indica el número de pacientes en cada clase.

| *sexo* \ tipos de diabetes | Tipo I | Tipo II | TOTAL |
|---|---|---|---|
| **Masculino** | 25 | 20 | **45** |
| **Femenino** | 35 | 20 | **55** |
| **TOTAL** | **55** | **40** | **95** |

Si se selecciona un archivo aleatoriamente, determine las probabilidades de que el individuo seleccionado:
a.- sea de sexo femenino
b.- tenga diabetes del tipo II

**Solución:** Se procede de la siguiente manera:

Los eventos son: **f:** Los pacientes de diabetes tipo II.
**n:** Los pacientes femeninos de diabetes tipo II.

Se calcula las cuatro probabilidades a la vez**:**

$P_1 = \frac{55}{95} = 0{,}578$ **$P_1$= 0,578 ; $P_1$= 57,8%**

$P_2 = \frac{40}{95} = 0{,}010$ **$P_2$ = 0,010; $P_2$ = 0,10%**

$P_3 = \frac{55}{95} =$ **$P_3$ = ; $P_2$ = %**

$P_4 = \frac{55}{95} = 0$ **$P_4$ = ; $P_4$ = %**

**Interpretación:** Esto quiere decir que:

1.- Existe el 0,578 ó 57,8 % de probabilidad de que sea de tipo II.
2.- El 0,010 ó 0,10% de probabilidad de que sea femenino.

3.- Existe el 0,2307 ó 23,07 % de probabilidad de que sea de tercer año,
4.- El 0,5641 ó 56,41% de probabilidad de que sea varón.

**REGLAS DE PROBABILIDAD PARA LA ADICIÓN Y EL PRODUCTO**

Si se tienen dos eventos cualesquiera, identifiquémoslo con las letras mayúsculas A y B, entonces

Para la adición: **P(A o B) = P(A) + P (B) – P(A –B)**
Y también, **P(A o B) = P(A) + P (B)**
Para eventos independientes:
Probabilidad condicional: **P(A/B) = P(A y B) / P (B)**

Para el Producto: **P(A y B) = P(A) . P(B /A)**
O bien **P(A y B) = P(B) . P(A /B)**

**P(A y B) = P(A) . P(B)**

**1.1.4. TIPOS DE PROBABILIDADES:**

**1. Probabilidad simple:** Probabilidad de que el dato escogido tenga una característica.
**Ejemplo:** Cuando se trata de persona; es hombre o mujer.

**2. Probabilidad conjunta:** Probabilidad de escoger un dato con dos (o más) características específicas.
**Ejemplo:** Cuando se trata de persona; es hombre o mujer, la talla y peso; son más de dos características.

Para el estudio de este tipo de probabilidad conjunta lo detallaremos de la manera siguiente:

- ***Probabilidad Conjunta de eventos mutuamente excluyentes:*** Son aquellos eventos definidos de manera que la ocurrencia de uno es imposible la ocurrencia de los demás, (brevemente, si alguno de ellos sucede, los restantes no pueden suceder). Se denotan en la fórmula con la letra ó.
- ***Probabilidad Conjunta de eventos solapados:*** Dos o más eventos son solapados si tienen puntos muestrales comunes, estos puntos maestrales forman una intersección entre ellos. Se denotan en la fórmula con la letra y.

- ***Probabilidad Conjunta de eventos complementarios:*** Dos eventos son complementarios si el segundo eventos tiene todos los puntos muestrales que no están en el primer evento.

- ***Probabilidad Conjunta de eventos independientes:*** Dos eventos son independientes cuando la ocurrencia o no ocurrencia de uno de ellos en una prueba, no afecta la probabilidad del otro en cualquier otra prueba.
- ***Probabilidad Conjunta de eventos dependientes:*** Dos o más eventos son dependientes cuando el conocimiento de la verificación de uno de ellos, altera la probabilidad de verificación del o de los otros.

**Ejemplo:** Una compañía desea probar un producto en una zona comercial seleccionada aleatoriamente. Las áreas pueden ser clasificadas con base en su ubicación y densidad de población. A continuación en la tabla siguiente se presenta el número de mercados en cada categoría:

| **DENSIDAD DE POBLACIÓN** | | | |
|---|---|---|---|
| **UBICACIÓN** | **Urbana(U)** | **Rural (R)** | **TOTAL** |
| Este (E) | 25 | 50 | **75** |
| Oeste (O) | 20 | 30 | **50** |
| **TOTAL** | **45** | **80** | **125** |

¿Cuál es la probabilidad de que el mercado seleccionado para la prueba esté en el este P(E)?
¿Cuál es la probabilidad de que el mercado seleccionado para la prueba esté en el oeste P(O)?
¿Cuál es la probabilidad de que esté localizado en un área urbana P(U)?
¿Cuál es la probabilidad de que esté localizado en un área rural P(R)?
¿Cuál es la probabilidad de que el mercado este en un área rural al oeste, P(R y O)?
¿Cuál es la probabilidad de que esté al este o dentro de un área urbana, P(E ó U)?
¿Cuál es la probabilidad de que si está en el este, esté localizado en un área urbana, P(U/ E)?
¿Son independientes la “ubicación” y la “densidad de población”?

**Solución:** Las primeras cuatro probabilidades, P(E), P(O), P(U), y P(R) representan preguntas del tipo “o”; esto quiere decir que los componentes son mutuamente excluyentes, estas probabilidades se pueden resolver aplicando la fórmula conocida, obteniendo las probabilidades sumando cada caso a través de las hileras o columnas del cuadro. En consecuencia los totales se encuentran en el total de columnas o hileras.

$P(E) = \frac{75}{125}$ (Total para el estudio dividido entre el número total de mercados)

$P(O) = \frac{50}{125}$ (Total para el oeste dividido entre el número total de mercados)

$P(U) = \frac{45}{125}$ (Total para la ubicación urbana dividido entre el número total de mercados)

$P(R) = \frac{80}{125}$ (Total para la ubicación rural dividido entre el número total de mercados)

Ahora se obtendrá P(O y R). Hay 30 mercados en el Oeste y 125 mercados en total. Así: $P(O \text{ y } R) = \frac{30}{125}$

Nótese que P(O)**.**P(R) **NO** proporciona la respuesta correcta $\left[\left(\frac{50}{125}\right)\left(\frac{80}{125}\right) = \frac{32}{125}\right]$. En consecuencia, "ubicación" y "densidad de población" son eventos dependientes.

La probabilidad P (E ó U) puede hallarse en varias formas. La más directa es examinando simplemente el cuadro y contando el número de mercados que satisfacen la condición de estar en el este o ser urbanos. El número obtenido es

95 = (25 + 50 + 20). Así:

$$P(E \text{ ó } U) = \frac{95}{125}$$

Nótese que los primeros 25 mercados están en el este y son urbanos; así E y U **no** son eventos mutuamente excluyentes.

Otra manera de encontrar P (E ó U) es utilizando la fórmula de la adición:

$P(E \text{ o } U) = P(E) + P(U) - P(E \text{ y } U)$, lo cual produce:

$$\frac{75}{125} + \frac{45}{125} - \frac{25}{125} = \frac{95}{125}$$

Una tercera forma de resolver el problema consiste en reconocer el complemento de (E ó U) es (O y R). Así P(E ó U) = 1 – P(O y R). Utilizando el cálculo anterior se obtiene $1 - \frac{30}{125} = \frac{95}{125}$

Finalmente se obtendrá P(U/E). Si se examina el cuadro anteriormente dado puede hacerse que hay 75 mercados en el Este. De estos, 25 son urbanos. Así:

$$P\left(U / E\right) = \frac{25}{75}$$

También puede utilizarse la fórmula de la probabilidad condicional:

$$P\left(U /E\right) = \frac{P\,(U \text{ y } E)}{P(E)} \qquad \Rightarrow \frac{\frac{25}{125}}{\frac{75}{125}} = \frac{25}{75}$$

"Ubicación" y "densidad de población" **NO** son eventos independientes (para este caso). Son dependientes, esto significa que loa probabilidad de estos eventos resulta afectada por la ocurrencia del otro.

**3. Probabilidad marginal (al margen de la tabla):** No es más que la probabilidad simple, vista con otro enfoque; o sea, mientras que la probabilidad simple es un concepto singular, la probabilidad marginal es esencialmente una suma de probabilidades conjuntas.

**Ejemplo:** Se presentó en el ejemplo de la tabla del primer ejemplo.

## 2.1. DISTRIBUCIÓN.

### 2.1.1. ¿Qué es la Distribución?

La distribución probabilística es esencialmente una explicación del comportamiento de un determinado fenómeno, es una herramienta imprescindible para tomar decisiones en aspectos donde de alguna forma intervenga la incertidumbre.

Existen varios tipos de distribuciones, las cuales son:

### 2.2.1. Distribución Binomial

Una distribución sigue la ley binomial siempre y cuando se cumplan las siguientes hipótesis:

1.-Un experimento es repetido varias veces, siendo sus resultados independientes.

2.-Los resultados de cada experimento se pueden clasificar en dos categorías mutuamente excluyentes, llamadas "éxito" o "fracaso".

3.-Las probabilidades de “éxito” o “fracaso” en una sola prueba, designadas respectivamente por p y q, donde q =1 – p, son invariables en todas las pruebas o experimentos.
4.-En cualquier experimento, el centro de interés estriba en si los resultados esperados ocurren o no.
5.- El experimento se realiza en las mismas condiciones un número fijo de pruebas “n”.
6.- La distribución es asimétrica negativa si; p > 0,5 ó p >1/2
7.- La distribución es asimétrica positiva si; p< 0,5 ó p < 1/2
8.- Es simétrica la distribución cuando p = 0,5 ó p = 1/2
9.- La distribución binomial se plica cuando la muestra proviene de una población infinita o cuando es extraída de una población finita con remplazamiento.

Su expresión matemática es:

$$P=(k,n,p)=\binom{n}{k}\cdot p^{k}\cdot q^{n-k}$$

**Ejemplo:** El 60% de las historias clínicas de un hospital de Caracas corresponden a adolescentes. Si se seleccionan 5 historias clínicas, ¿Cuál es la probabilidad de que ellas correspondan a adolescentes?

**Solución:** 1.- La ecuación es

$$P=(k,n,p)=\binom{n}{k}\cdot p^{k}\cdot q^{n-k}$$

2.- k es la cantidad de éxitos esperados: k = 3
3.- n es la muestra seleccionada: n = 5
4.- p es la probabilidad de éxito: p = 60/100 = 3/5
5.- q es la probabilidad de no éxito: q = 1 – p = 1 3/5 = 2/5
6.- Se desarrolla el número combinatorio:

$$\binom{n}{k}=\frac{n!}{(n-k)!\,.\,k!} \qquad \binom{5}{3}=\frac{5!}{(5-3)!\,.\,3!}=\frac{5\,.\,4\,.\,3!}{(5-3)!\,.\,3!}=$$

donde; $= \frac{20}{2\,.1}=10$

7.- Se sustituyen los valores en la fórmula:

**$P(3,\ 5,\ 3/5) = 10\ .\ (3/5)^{3}\ .\ (2/5)^{5-3} = 10\ .(27/125).\ (4/25) = 1080/3125 = 0{,}3456.$**

**Interpretación:** Si el 60% de las historias clínicas de un hospital son de adolescentes y del grupo se eligen 5 al azar, existen una probabilidad de 0.3456 de que 3 de ellas sean de adolescentes.

### 2.2.2. Media y Desviación Estándar de la Distribución Binomial

La **media** y la **desviación estándar** de una distribución binomial pueden obtenerse utilizando las dos fórmulas siguientes:

Esperanza matemática: $\mu = n.p$

Varianza: $\sigma^2_x = n.\ p.\ q$

Desviación Típica: $\sigma_x = \sqrt{}\ n.\ p.\ q$

Donde; $\mu$: Esperanza matemática
**n :** número de puntos muestrales
**p:** que indica éxito
**q: (1 – p):** que indica fracaso

**Ejemplo:**Determinar la media y la desviación estándar de la distribución binomial donde **n = 20** y **p =1/5.** Recuerde que esta distribución tiene o se compone de un cuadro sus valores con sus respectivas probabilidades, en este caso 21 valores y 21 probabilidades correspondientes.

**Solución:** Ahora utilizando la fórmula de la media y desviación estándar de esta distribución son iguales a:

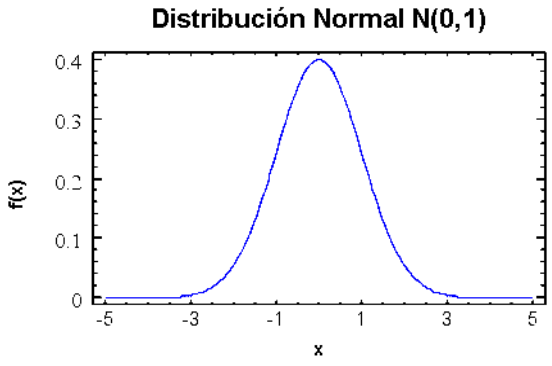
$\mu$ = n.p = (20)(1/5)= 4.0; es el valor medio de la variable aleatoria x.

$$\sigma_x = \sqrt{n.\,p.q} = \sqrt{(20)\,(1/5)\,(4/5)} = \sqrt{80/25} = \frac{4\sqrt{5}}{5} = \mathbf{1{,}79}$$

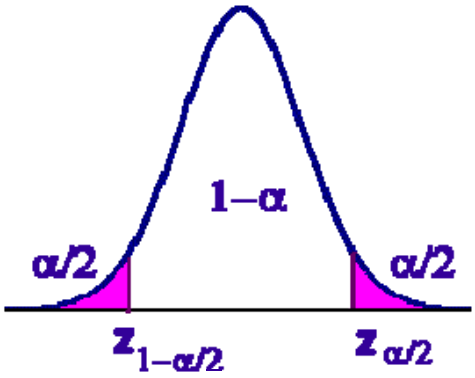
$\sigma_x$ = 1.79 es la desviación estándar de la variable aleatoria x. Además como lo dice en el enunciado, el dicho cuadro con sus 21 valores de x y sus 21 probabilidades correspondientes tienden a un decrecimiento en sus valores de x.

## 3. DISTRIBUCIÓN NORMAL

La distribución normal se presenta con un enfoque más práctico representándose con una gráfica o **Curva Normal** o **De Campana** utilizando una escala aproximada, como se indica en la figura siguiente:



En la siguiente figura se muestra la porcentaje-proporción que están relacionados, por lo general se utiliza el porcentaje, por ejemplo en una población, con la posibilidad de que el evento que se estudia tome un valor entre ciertos límites



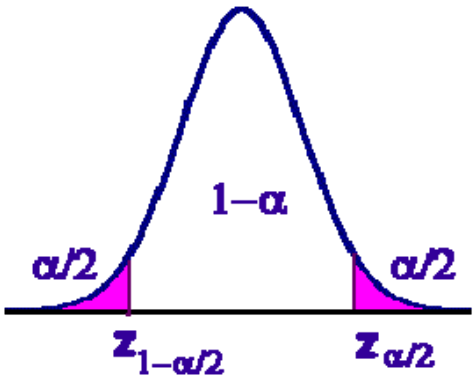

Una variable aleatoria tiene una distribución normal general, si es continua, si existen parámetros: "μ" (letra griega, miu) con un valor entre -∞ y +∞, y "σ" (letra griega, sigma) con valor mayor que cero; y si su función de densidad "f(x)" es de la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad \begin{cases} e = 2{,}71828 \\ \pi = 3{,}14159 \end{cases}$$

La representación gráfica de esta expresión es una curva simétrica respecto a la ordenada máxima.

Para indicar que una variable aleatoria sigue una distribución normal, se utiliza la notación:

$$X \sim N(\mu, \sigma)$$



## PROPIEDADES DE LA DISTRIBUCIÓN NORMAL

1.- Tiene como parámetros a: "$\mu$" y "$\sigma$", N ($\mu$ ,$\sigma$).
2.- La curva de la distribución normal es asintótica, es decir, las colas de la curva nunca llegan a tocar el eje de las abscisas.
3.- La distribución normal es simétrica con respecto a la ordenada máxima, siendo por lo tanto, las medidas de tendencia central iguales entre sí, es decir, $\overline{X} = Xd = Xo$.
4.- Si "X" está normalmente distribuida con "$\mu$" y "$\sigma$", entonces, ***z*** = (X - $\mu$)/$\sigma$, estará también normalmente distribuida.

Esta transformación de "X" a ***"z"*** (tipificación o estandarización) tiene el efecto de reducir "X" a unidades en términos de desviación típica. Es decir, dado un valor "X", el correspondiente valor de ***"z",*** nos dice en qué sentido y a que distancia se encuentra "X" de su "$\mu$" (media aritmética) en términos de desviaciones típicas.

Esta propiedad nos permite transformar el modelo normal general en el modelo normal tipificado o estandarizado:

### 3.1. DISTRIBUCIÓN NORMAL TIPIFICADO ó ESTANDÁR

Se dice que una distribución normal es de la forma estándar si su media aritmética es cero y su varianza es uno, y por lo tanto, su desviación típica es la unidad. Para esta distribución vamos a trabajar con el **valor *z,*** las medidas están asociadas a una variable ***x*** que esta determinada por su posición relativa con respecto a la media y la desviación estándar de la distribución, donde el valor de ***z*** esta definido de la siguiente manera:

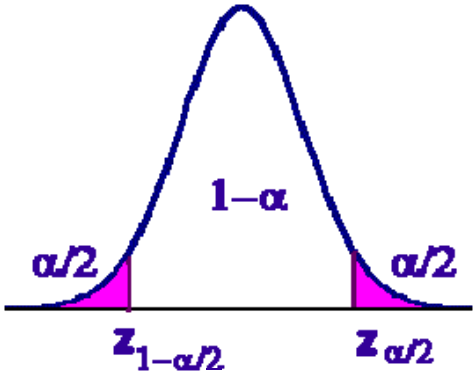
$$Z = \frac{X - \mu}{\sigma}$$

El valor de ***z*** es considerado como una variable "estandarizada" ya que sus unidades son desviaciones estándares, es decir que todas las probabilidades están asociadas a intervalos centrados en la media para los valores específicos a el valor ***z***, esto lo llamamos también representación de la probabilidad mediante un área.

Donde el valor de ***z*** se representa en el intervalo;

$P$ ($-\infty < z < +\infty$ ); o valor de $\sigma$ esta en valores positivos y negativos, y viene dada por:

$$f(z) = \frac{1}{2\sqrt{2\pi}} \cdot e^{-\frac{z}{2}}$$ ; la cual gráficamente representa una curva simétrica con respecto a "***z*** =0".




Cuando los valores viene dados por la "tabla de la distribución normal", y con valores o números reales (positivos y negativos), para tipificar la puntuación, esto es, tomar en cuenta la simetría de **f(z)** con respecto a **z,** se proporciona el cálculo y su uso de la tabla de la distribución normal, como se sigue la fórmula

$$Z_1 = \frac{a - \mu}{\sigma} \quad ; \quad Z_2 = \frac{b - \mu}{\sigma} \ldots$$ En general

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

**Ejemplo:** Al aplicar una encuesta de habilidades numéricas al sector "Los Maguitos" de la parroquia San Juan, se tomó una muestra a 300 personas de dicho sector, donde se obtuvo una distribución normal con una $\mu$ = 36 puntos y $\sigma_x$ = 5. Se desea saber:

a.- La amplitud intercuartil.

b.- ¿Cuál es la probabilidad de obtener una puntuación igual o inferior a X= 32?

c.- ¿Cuántas personas de la muestra tienen un puntaje igual o mayor que X =34?

**Solución:** Existen "**tres maneras"** de resolver este ejercicio:

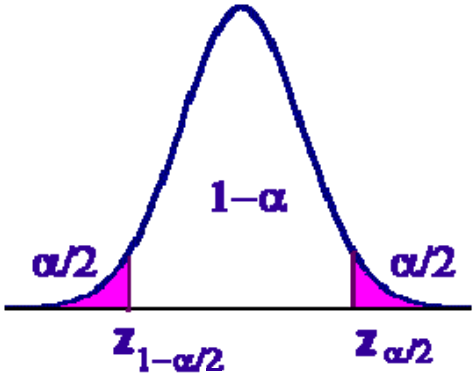
**1$^{era}$ manera:**

1.- Para calcular la amplitud es necesario conocer los cuarteles primero y tercero ($Q_1$ y $Q_3$). Esta amplitud es la diferencia de los dos cuarteles.

2.- Como la distribución es normal, bastará con buscar cuáles son los puntos de la distribución que separan el 25% inferior a la media aritmética y el 25% superior. Para ello, obtenemos las puntuaciones **z** correspondiente.

3.- Una vez conseguidos los valores de **z,** buscamos en la tabla de áreas bajo la curva normal con los porcentajes anteriores, se despeja de la ecuación de **z,** la incógnita (X = puntuación que corresponderían a los cuarteles $Q_1$ y $Q_2$).

$$Z_1 = \frac{X_1 - 36}{5} = -0,67\,; \quad Z_2 = \frac{X_2 - 36}{5} = 0,67$$




$X_1 = Q_1 = 5\ (-0,67) + 36 = 32,65.$
$X_2 = Q_2 = 5\ (0,67) + 36 = 39,35.$

Se calcula la amplitud intercuartíl: Q

$$Q = Q_2 - Q_1 = 39,35 - 32,65 = 6,7.$$

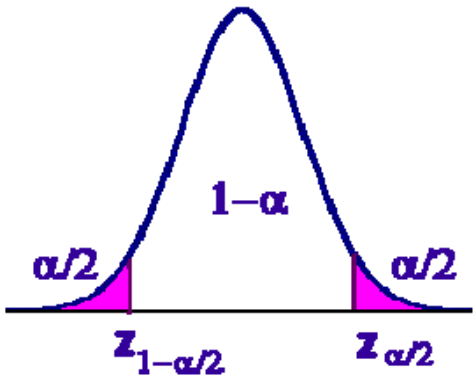
**2da manera:**

1.- La probabilidad de obtener una puntuación igual o inferior a 32, es la proporción del área de la curva normal existente por debajo de esa puntuación bruta o directa.

2.- Se tipifica la puntuación:

$$\mathbf{Z} = \frac{X - \mu}{\sigma_X} = \frac{32 - 36}{5} = -0,80$$

3.-la puntuación tipificada (z= -0,80) se busca en la tabla de áreas bajo la curva normal y corresponde a un 28,81%. Por lo tanto, por debajo de dicha puntuación se encontrará un (50 -28,81) =21,19%.

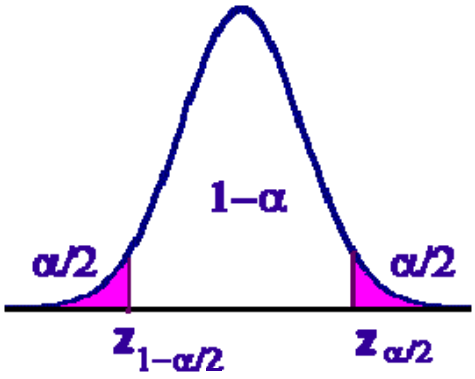


**3$^{era}$ manera:**

1.- Como la distribución es normal, los alumnos que tengan una puntuación igual o mayor que 34, corresponderá al porcentaje de la curva normal que supere esa puntuación.

2.- Se obtiene la puntuación típica:

$$\mathbf{Z} = \frac{X - \mu}{\sigma_X} = \frac{34 - 36}{5} = -0,40$$




3.- Se busca en la tabla de áreas para la curva normal, el valor (z= -0,40), obteniéndose un porcentaje igual a 15,54%.

4.- Como z= -0,40, es negativa, el porcentaje de alumnos, ubicados por encima de esa puntuación, será (15,54% + 50) = 65,54%.

5.-Si la muestra tiene 3000 alumnos, la cantidad de los que le superen esa puntuación (X= 34), es el 65,54% de 300, es decir (300)(65,54)/100 = 197 alumnos.

## NORMALIZACIÓN

**Escala T:** Los puntajes obtenidos en una distribución cualquiera pueden llevarse a puntos equivalentes dentro de una distribución normal.

Los puntajes **T** son puntajes estándar normalizados, convertidos en una distribución cuya media aritmética es 50 y alejándose en -5$\sigma$ de la media, le corresponde 0, mientras que el se aleja 5$\sigma$ de la media, tiene 100 puntos. Para obtener un puntaje **T**, se utiliza la siguiente fórmula:

$$\mathbf{T = 50 + 10 .z}$$

**Ejemplo:** Transformar las puntuaciones de la prueba de Biología de UBV, a la escala de veinte valores, utilizando la escala T.

**Solución:**
1.-Se obtienen las frecuencias ajustadas con la siguiente ecuación:

**F ajustada = F inferior = ½ . f**

Es decir, para calcular la frecuencia ajustada de una determinada casilla o intervalo de clase en una distribución de frecuencias, se le suma a la frecuencia acumulada inferior, la mitad de la frecuencia absoluta ordinaria de la casilla con que se esté trabajando, así:

Para el intervalo 10: (72 - 76) 3 50
9: (67 - 71) 7 47

F ajustada = 47 +1/2 . 3 = 47 + 1,5 = 48,5

Para el intervalo 9: (67 – 71) 7 47
8: (62 – 66) 1 40

F ajustada = 40 + (7/2) = 40 + 3,5 = 43,5

2.- Se obtienen los porcentajes acumulado (P) de esas frecuencias ajustadas (F ajustada), en base a la mayor frecuencia acumulada (no ajustada), mediante la fórmula:

$$\text{P acumulada} = \frac{\text{F ajust.}}{\text{F máx. no ajust.}} . 100$$

Para el intervalo 10: (72 – 76 3 50)

$$\text{P acum.}_{10} = \frac{48,5}{50} .100 = 97$$

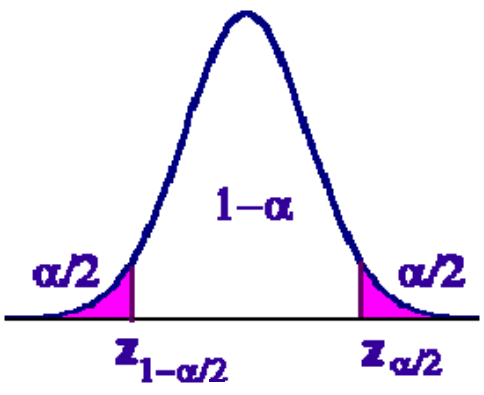
3.- Se obtiene en la curva normal el puntaje típico que corresponde a cada uno de os porcentajes obtenidos. Z = 1,88.

4.-Se obtienen los puntajes normalizados T, mediante la ecuación:

**T = 50 + 10 . z**

Para el intervalo 10: T = 50 + 10(1,88) = 68,8 = 69

Para el intervalo 9: T = 50 + 10 (1,13) = 63.



5.- Los puntajes T, pueden ser convertidos a nuestra escala tradicional de veinte valores mediante la relación:

$$E_1 \cdot 20 = (T/ Tmáx) \cdot 20$$

**Tabla de operaciones**

| Nº | $X_i - X_s$ | f | F | F ajust. | P | z | T | $E_1 \cdot 20$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 72 – 76 | 3 | 50 | 48,5 | 97 | 1,88 | 69 | 20 |
| 9 | 67 – 71 | 7 | 47 | 43,5 | 87 | 1,13 | 63 | 18 |
| 8 | 62 – 66 | 1 | 40 | 39,5 | 79 | 0,81 | 58 | 17 |
| 7 | 57 – 61 | 9 | 39 | 34,5 | 69 | 0,50 | 55 | 16 |
| 6 | 52 – 56 | 2 | 30 | 29 | 58 | 0,20 | 52 | 15 |
| 5 | 47 – 51 | 5 | 28 | 25,5 | 51 | 0,02 | 50 | 14 |
| 4 | 42 – 46 | 4 | 23 | 21 | 42 | - 0,20 | 48 | 14 |
| 3 | 37 – 41 | 4 | 19 | 17 | 34 | - 0,41 | 46 | 13 |
| 2 | 32 – 36 | 6 | 15 | 12 | 24 | - 0,71 | 43 | 12 |
| 1 | 27 – 31 | 9 | 9 | 4,5 | 9 | - 1,34 | 37 | 11 |
| | Σ | 50 | | | | | | |

## 1.11. ACTIVIDADES

Individual

- Lea con cuidado los contenidos presentados en este modulo y consulte la bibliografía a fin de ampliar sus conocimientos y considerar la opinión de otros autores sobre el tema.

Grupal Cooperativa

- Los empleados de una universidad fueron clasificados de acuerdo con su edad y adscripción a la administración, cuerpo docente o personal de apoyo.

| Clasificación/ *grupo* de edad | **20 – 30** | **31 – 40** | **41 – 50** | **51 o mayor** | **TOTAL** |
|---|---|---|---|---|---|
| **Administración** | 2 | 24 | 16 | 17 | **59** |
| **Cuerpo Docente** | 1 | 40 | 36 | 28 | **105** |
| **Personal de Apoyo** | 16 | 20 | 14 | 2 | **52** |
| **TOTAL** | **19** | **84** | **66** | **47** | **216** |

Considerando que se selecciona un empleado en forma aleatoria, obtenga la probabilidad de que el elegido:
a.- esté en la administración o tenga 51 años o más.
b.- no sea miembro del cuerpo docente.
c.- sea miembro del cuerpo docente dado que el individuo tiene 41 años o más.

- Suponga que cierta característica oftálmica está asociada al color de los ojos. Se estudiaron 3000 personas seleccionadas aleatoriamente con los siguientes resultados:

| **características** \ *color* de los ojos | **Azul** | **Café** | **Otro** | **TOTAL** |
|---|---|---|---|---|
| **Si** | 70 | 30 | 20 | **120** |
| **No** | 20 | 110 | 50 | **180** |
| **TOTAL** | **90** | **140** | **70** | **300** |

a.-¿Cuál es la probabilidad de que una persona seleccionada al azar tenga los ojos azules?
b.- ¿Cuál es la probabilidad de que una persona seleccionada al azar si tenga la característica?
c.-¿Son independientes los eventos A (tiene los ojos azules) y B (tiene la característica)? Justifique la respuesta.
d.- ¿Cómo están relacionados los eventos A (tiene ojos azules) y C (tiene ojos cafés) (independientes, mutuamente excluyentes)? Explique por qué cada término puede ser aplicado o no.

- Los pesos de sandías maduras cultivadas en una granja están distribuidos normalmente con una desviación estándar de 2.8 Kgrs. Obtenga el peso medio de las sandía maduras si sólo 3% pesa menos de 15 Kgrs.

- Se supone que un medicamento nuevo es 85% efectivo en el tratamiento de cierta enfermedad. (Es decir, el 85% de los pacientes con esta enfermedad responden favorablemente al medicamento). Sea ***z*** el número de pacientes de cada grupo de 50 que responden favorablemente. Utilice el método de aproximación normal para evaluar las siguientes probabilidades;
a.- P(x > 45)  b.- P (40< x < 50)  c.- P(x < 35)

- Una escuela primaria ha programado cuatro fechas de reunión al año con los padres de familia. Los registros de la escuela indican que la probabilidad de que los padres de un niño (uno o ambos) asistan desde 0 hasta 4 de las reuniones son las indicadas en el cuadro siguiente:

| **Número de Reuniones a las que asisten (x)** | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **Probabilidad** | 0.12 | 0.38 | 0.30 | 0.12 | 0.08 |

a.- ¿Es ésta una distribución de probabilidades? Explique.
b.- ¿Cuál es la probabilidad de que los padres de un niño en particular asistan al menos a una de esas reuniones?
c.- Calcule la media y la desviación estándar para esta distribución.
- Se ha aplicado una prueba de aptitudes sensoriales a 200 alumnos de un Liceo Capitalino, obteniéndose una media aritmética de 20 y una desviación típica de 5. Suponiendo normalidad:
a)¿Cuál es la probabilidad de obtener una puntuación mayor que 15 y menor que 18?
b)¿Cuál es la puntuación bruta que supera el 15% superior de la distribución?

Comunitario

## BIBLIOGRAFÍA BÁSICA


- Macchi, R.L. 2001. Introducción a la Estadística en Ciencias de la Salud. Editorial Médica Panamericana. Argentina.
- Milton, J.S. y Tsocos, J.O. 1991. Estadística para Biología y Ciencias de la Salud. McGraw-Hill, Inc.
- Puertas L., E.; Urbina, J.; Blanck, M.E.; Granadillo, D.; Blanchard, M.; García, J.A.; Vargas V.; P. & Chiquito, A. 1998. Bioestadística, Herramienta de la Investigación. Ediciones del Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Carabobo, Venezuela.
- SALAMA, D. 1987. *Estadística: Metodología y aplicaciones.* Editora Principios, Caracas, Venezuela. 308 p.
- SEGNINI, S. 2003. *Apuntes de Estadística para Biólogos*. Dirección de Publicaciones ULA, Mérida, Venezuela.
- Spiegel, M.R. y Stephens, L.J. 2002. Estadística. Serie Schaum. 3era edición. McGraw-Hill, Inc.


## COMPLEMENTARIA


- Sokal, R.R. y Rohlf, F.J. 1995. Biometry, the principles and practice of statistics in biological research. 3era edición. W.H. Freeman and Company. USA.


## PÁGINAS WEB


http://www.hrc.es/bioest/M_docente.html

http://www.e-biometria.com/ebiometria/conceptos_basicos/estimacion_estadistica.htm

Dr. Hossein Arsham
http://home.ubalt.edu/ntsbarsh/Business-stat/opre504S.htm#rqualestiunbsuff

# CAPITULO II: PRUEBA DE HIPÓTESIS E INTERVALOS DE CONFIANZA

## TEMA 4. ERROR ESTÁNDAR

**Competencias:**

Conoce la medida en la que se alejan los datos de la media poblacional, es decir, la diferencia entre el valor estimado y el valor real.

**Contenidos:**

Introducción a la Estadística Inferencial.
Definición y cálculos del error estándar.
Usos del error estándar.
Tamaño muestral.

### Introducción a la Estadística Inferencial

En este capítulo trabajaremos con las técnicas de la Estadística Inferencial, a través de las cuales se busca llegar a conclusiones valederas sobre poblaciones, tomando como base la información obtenida en una muestra. La única forma de conocer la información exacta sería realizando todas las observaciones posibles de todo el universo, lo cual suele ser difícil y poco práctico en función del costo y del tiempo. De allí surge la inferencia estadística la cual permite asumir o estimar las características de la población a partir de las muestras.

Los dos tipos de problemas que resuelven las técnicas estadísticas son: estimación y contraste de hipótesis. En ambos casos se trata de generalizar la información obtenida en una muestra a una población. Estas técnicas exigen que la muestra sea aleatoria. En la práctica rara vez se dispone de muestras aleatorias, por la tanto la situación habitual es la que se esquematiza en la Figura 1.

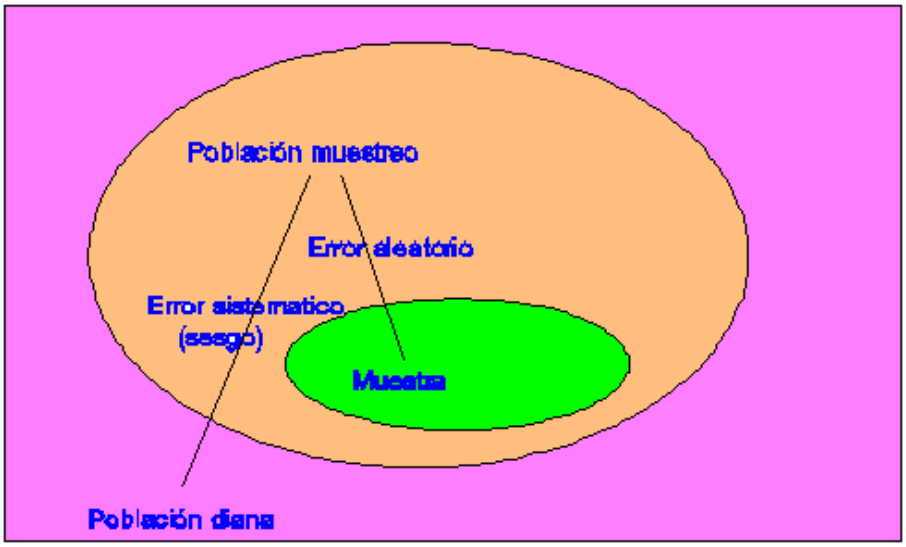


Figura 1. Diferenciación entre la población y la muestra.

Entre la muestra con la que se trabaja y la población de interés, o población diana, aparece la denominada *población de muestreo*: población (la mayor parte de las veces no definida con precisión) de la cual nuestra muestra es una muestra aleatoria. En consecuencia, la generalización está amenazada por dos posibles tipos de errores: *error aleatorio* que es el que las técnicas estadísticas permiten cuantificar y críticamente dependiente del tamaño muestral, pero también de la variabilidad de la variable a estudiar y el *error sistemático* que tiene que ver con la diferencia entre la población de muestreo y la población diana y que sólo puede ser controlado por el diseño del estudio.

Cabe recordar que la información de las muestras se trabaja a partir de estadísticos. Un estadístico es una variable aleatoria cuyos valores pueden ser determinados a partir de la observación de una muestra aleatoria. Este muestra una distribución de probabilidades propias, la cual es conocida como distribución muestral de un estadístico. Así puede tenerse una distribución muestral de medias, cuando el estadístico es la media aritmética; distribución muestral de proporciones, cuando el estadístico es una proporción o porcentaje, y así sucesivamente.

Las medidas obtenidas en una muestra (estadísticos) frecuentemente son diferentes al parámetro de la población. A la diferencia de estas dos medidas se les denomina error. Determinar el tamaño de ese error sólo sería posible si se conociera el parámetro de la población, pero este por lo general se desconoce.

Sin embargo, el error es posible estimarlo siguiendo un modelo estadístico. En el caso que el estadístico sea la media aritmética tenemos:

$X_i = \mu + E_i$

$X_i$ : es el valor de la variable.

$\mu$ : es la media poblacional.

$E_i$ : es el error.

Se pueden graficar las frecuencias de $E_i$, cuya distribución de frecuencias se comporta como una distribución normal, con media cero y varianza $s^2$ o $\sigma^2$. Los errores pueden ocurrir por exceso o por defecto, lo que significa que $E_i$ tendrá valores positivos y negativos. Cuando se estudia un gran número de valores de $E_i$, el promedio de ellos es cero.

En una distribución de frecuencias la medida de dispersión es la desviación estándar, y en la distribución muestral es el error estándar, el cual no es más que el promedio de los errores muestrales. Ello significa que las distintas medias de la distribución muestral contienen una fracción de error en sus estimaciones con respecto a la media poblacional.

Todo estadístico de una variable continua tiene una distribución muestral, donde:

1. La media de las medias de las muestras es igual a la media de la población.
2. La varianza de las medias de las muestras es igual a la varianza de la población, dividida entre el tamaño de las muestras (*n*).
3. La distribución de las medias muestrales tiene forma de curva normal.

**Definición y Cálculo del Error Estándar**

El error estándar puede definirse como la diferencia que existe entre el valor estimado en la muestra (estadístico) y el verdadero valor representativo de la población (parámetro), por lo tanto, mientras menor sea el error estándar mayor será la aproximación del estadístico al parámetro.

Al error estándar también se le conoce como error por muestreo o error típico, y puede indicarse que la magnitud del error es directamente proporcional a la dispersión de la población de origen de la muestra e inversamente proporcional al tamaño de ésta. Mientras mayor sea la muestra, menor es la magnitud del error estándar.

Cálculo del error estándar

Dado que el error estándar es una desviación estándar en una distribución muestral, este representa una medida de la dispersión de la distribución de los valores de las medias de muestras tomadas de una población, de la misma manera que la desviación estándar lo es para la dispersión de los datos originales.

Para calcular el error estándar de la media se utiliza la formula:

$$S_x = \frac{s}{\sqrt{n}}$$

Donde s: desviación estándar de la muestra
n: tamaño de la muestra
$S_x$ : error estándar

Para calcular el error estándar de una proporción o porcentaje:

$$Sp = \frac{p \times q}{\sqrt{n}}$$

Donde *p*: porcentaje de sujetos con la característica de estudio.
*q*: (100-*p*) porcentaje de sujetos sin la característica de estudio.

*n*: tamaño de la muestra

Obsérvese que hay dos situaciones en las que la posibilidad de error es nula (error estándar igual a cero):

- Una de ellas se produce cuando en la población original no hay dispersión, es decir que todos sus datos son iguales. Al ser el numerador cero, el cociente es cero.
- La segunda se verifica cuando la muestra tomada es infinitamente grande, o cuando se evalúa la totalidad de la población, en este caso el denominador es infinito y el resultado de dividir cualquier valor por infinito es cero.

En la realidad de la investigación es poco probable que se den estas dos situaciones, ya que en los datos numéricos es casi imposible evitar la dispersión, porque no todos los individuos de una población se comportan exactamente igual o porque es casi imposible no cometer algún error en la recolección de datos. Por otro lado, las poblaciones de interés son de tamaño demasiado grande como para que sea posible trabajar con todos sus integrantes.

En este sentido, puede indicarse que de todas las muestras tomadas en forma aleatoria a partir de una población:

a) Alrededor del 68 % tiene valores de media aritmética entre $\mu \pm 1\sigma_{\bar{x}}$

b) Cerca del 95 % tiene valores de media aritmética entre $\mu \pm 2\sigma_{\bar{x}}$

c) Alrededor del 99 % tiene valores de media aritmética entre $\mu \pm 2{,}5\sigma_{\bar{x}}$

De la misma forma puede indicarse que al tomar una muestra al azar:

a) Es “poco probable” ($p$ < 0,05) que su media aritmética esté alejada de la media de la población más de dos errores estándar.
b) Es “muy poco probable” ($p$ < 0,01) que su media aritmética esté alejada de la media de la población más de dos y medio errores estándar.

**Ejemplo 4-1**. Si de una población con $\mu$ = 1000 y $\sigma$ = 40, se toman muestras con n = 25 y puede esperarse que el 95 % de ellas tenga valores para su estadístico media aritmética entre 984 y 1016. Esto es así porque el error estándar en esta situación es 8 (40 dividido entre $\sqrt{25}$ ) y dos veces 8 es 16.

$$S_{\bar{x}} = \frac{40}{\sqrt{25}} = \frac{40}{5} = 8 \qquad 2\ S_x = 2 \times 8 = 16$$

$\mu + 2\ S_x = 1000 + 16 = 1016$

$\mu - 2\ S_x = 1000 - 16 = 984$

**Muestras con datos nominales**

Al tomar muestras de poblaciones de datos nominales la situación es equivalente a la descrita para los datos numéricos. Sigamos un ejemplo.

**Ejemplo 4-2**. Considérese una población hipotética de 8 individuos, de los cuales 4 (p = 0,5 o 50%) están en la categoría "enfermos". Los resultados posibles al tomar muestras de tamaño 4 (n = 4) se muestran en la Tabla 1 (Macchi, 2001). Al estimar el parámetro con el valor del estadístico a veces se "acierta" y a veces se sobrestima o subestima, pero "en promedio" se estima bien.

También en este caso la magnitud del error posible en la estimación es inversamente proporcional al tamaño de la muestra, a mayor tamaño de la muestra menor error posible.

La diferencia estriba en que la distribución en este caso no es normal sino binomial y el valor del error estándar es la raíz cuadrada del valor obtenido de:

$$p(1-p)/n$$

Esto es la raíz cuadrada del resultado del producto de la proporción en una categoría (0,5 en la categoría "enfermos" en el ejemplo) por la que no está en la categoría ([1 – p] = 0,5 en el ejemplo) dividido por el tamaño de la muestra (4 en el ejemplo).

$$S_x = \sqrt{\frac{0{,}5\ \ (0{,}5)}{4}} = \sqrt{\frac{0{,}25}{4}} = \sqrt{0{,}0625} = 0{,}25$$

Tabla 1. Resultados en las muestras tomadas de una población hipotética.
Población: enfermos: 4 sanos: 4 *p* = 0,5 50%

| MUESTRA | % ENFERMOS |
|---|---|
| A<br>4 enfermos<br>0 sanos | 100,0 |
| B<br>3 enfermos<br>1 sano | 75,0 |
| C<br>2 enfermos<br>2 sanos | 50,0 |
| D<br>1 enfermo<br>3 sanos | 25,0 |
| E<br>0 enfermo<br>4 sanos | 0,0 |
| Suma<br>% promedio | 250,0<br>50,0 |

**Usos del error estándar**

En función del error estándar y de las propiedades de la distribución muestral, es posible:

a) Estimar los valores representativos de una población.
b) Tomar decisiones en función de pruebas de hipótesis.
c) Calcular el tamaño de una muestra, cuando se espera una determinada precisión del estadístico y el parámetro.

**Ejemplo 4-3**. En una población de adultos sin manifestaciones de presencia de cálculos sobre sus superficies dentales, el contenido de calcio en saliva tiene un valor de media aritmética de 5,6mg/100ml con una desviación estándar de 0,9 mg/100ml.

a) ¿Es "poco probable" ($p$ > 0,05) o no que la media aritmética de una muestra de tamaño 100 tenga un valor de 5,3 mg/100ml? Es poco probable, ya que este valor está alejado de la media de la población, 0,30 más de dos errores estándar. El error estándar de este caso es 0,09 (0,9/$\sqrt{100}$ ) que multiplicado por dos es 0,18.

$\mu$ = 5,6mg/100ml
$\sigma$ = 0,9 mg/100ml
n = 100

$$S_x = \frac{s}{\sqrt{n}} \qquad S_x = \frac{0,9}{\sqrt{100}} = \frac{0,9}{10} = 0,09$$

$$2\ S_x = 2 \text{ x } 0,09 = 0,18$$

$$\mu - 2\ S_x = 5,6 \text{ mg/100ml} - 0,18 \text{ mg/100ml} = 5,42 \text{ mg/100ml}$$

Es poco probable. 5,42 > 5,3

b) ¿Y si la muestra hubiera tenido un tamaño igual a 20? El valor obtenido no sería poco probable, ya que en este caso el error estándar es de 0,20 y (0,9/$\sqrt{100}$ ) multiplicado por 2 es 0,40, valor menor que 0,30.

Sí n = 20

$$S_x = \frac{0,9}{\sqrt{20}} = \frac{0,9}{4,47} = 0,20$$

$2\ S_x = 2 \text{ x } 0,20 = 0,40$

$$\mu - 2\ S_x = 5,6 \text{ mg/100ml} - 0,4 \text{ mg/100ml} = 5,2 \text{ mg/100ml}$$

Es probable. 5,2 < 5,3

**Ejemplo 4-4**. En una población de adultos jóvenes la estatura media (media aritmética) es de 1,70 m y la desviación estándar 0,24 m. ¿Menor o mayor de

qué valor debe ser la media aritmética de una muestra de tamaño 64, tomada de esa población para poder considerar que se está frente a una situación poco probable ($p$ > 0,05)? El error estándar de la distribución de las medias de las muestras de ese tamaño tomadas de esa población es de 0,03 (0,24 / $\sqrt{64}$ ). Los valores 1,64 y 1,76 están dos errores estándar alejados de la media. Por lo tanto, cuando la media de la muestra obtenida sea menor o mayor, respectivamente, que esos dos valores, se estará frente a una situación "poco probable".

$\mu$ = 1,70 m

$\sigma$ = 0,24 m

$$S_x = \frac{0,24}{\sqrt{64}} = \frac{0,24}{8} = 0,03$$

n = 64

$$2\ S_x = 2 \times 0,03 = 0,06$$

$\mu \pm 2\ S_x$ : $\mu + 2\ S_x = 1,70 \text{ m} + 0,06 \text{ m} = 1,76 \text{ m}$

$\mu - 2\ S_x = 1,70 \text{ m} - 0,06 \text{ m} = 1,64 \text{ m}$

**Tamaño muestral**

El tamaño muestral juega el mismo papel en estadística que el aumento de la lente en microscopía: si no se ve una bacteria al microscopio, puede ocurrir que:

- la preparación no la contenga
- el aumento de la lente sea insuficiente.

Para decidir el aumento adecuado hay que tener una idea del tamaño del objeto. Del mismo modo, para decidir el tamaño muestral:

i) en un problema de estimación hay que *tener una idea* de la magnitud a estimar y del error aceptable.

ii) en un contraste de hipótesis hay que saber el *tamaño del efecto* que se quiere ver.

Generalmente, se considera que el tamaño de la muestra debe estar en función del tamaño de la población, y se dice que debe ser proporcional a este. Sin embargo, cuando la población es muy extensa, no es indispensable que la muestra sea tan numerosa; es cuestión de determinar la cantidad apropiada, a fin de que el error muestral no afecte los resultados, y su vez no se derrochen recursos, al utilizar una muestra de mayor tamaño que la requerida.

Existen fórmulas que permiten calcular el tamaño adecuado de una muestra cuando se espera una determinada precisión en los resultados. Las fórmulas a utilizar dependen de la información disponible (Puertas y col., 1998).

1) Cuando se conoce el tamaño de la población (N), se puede aplicar la siguiente fórmula:

$$n = \frac{N}{1 + (N \times P^2)}$$

donde, $n$ : tamaño de la muestra
$N$ : número total de sujetos u objetos en la población o tamaño de la población
$P$ : precisión (error máximo permitido entre el parámetro y el estadístico), expresado en proporción.

2) Cuando se quiere estimar el promedio de una población y se conoce la desviación estándar de la población:

$$n = \frac{Z^2 \times s^2}{P^2}$$

donde, $n$ : tamaño de la muestra
$Z$ : 1,96 constante. Expresa el nivel de confianza
$s$ : desviación estándar (conocida o estimada) de la población
$P$ : precisión

3) Cuando se conoce la proporción o porcentaje la población que tiene la característica de interés:

$$n = \frac{p \times q}{P^2} \times Z^2$$

donde, $n$ : tamaño de la muestra
$Z$ : 1,96 constante. Expresa el nivel de confianza
$p$ : porcentaje de la población que tiene la característica de interés
$q$ : porcentaje de la población que NO tiene la característica de interés
($q = 1 - p$)
$P$ : precisión

**Ejemplo 4-5**. Se desea conocer las condiciones de las familias afectadas directamente por la inundación del Río Pao al Sur del Estado Anzoátegui. Se estima que el área de la cuenca afectada por la inundación abarca 20.000 familias. Se decide tomar una muestra en la cual el error máximo permitido en los resultados no sea mayor de un 5%. ¿Cuántas familias deben incluirse en la muestra?

$N$ = 20.000 familias
$P$ = 5% (0,05 expresado en proporción)

$$n = \frac{N}{1 + (N \times P^2)} = \frac{20000}{1 + \left[20000 \times \left(0{,}05^2\right)\right]} = 392{,}5$$

Por lo tanto, la muestra requerida debe ser de 393 familias.

**Ejemplo 4-6**. Un investigador necesita conocer el valor promedio de plomo en sangre venosa de los pacientes que asisten al hospital donde el trabaja en el

centro de Caracas. En la literatura revisada encuentra que el valor promedio de plomo en sangre es de 0,83 mg/100ml, con una desviación estándar de 0,05 mg/100ml, determinado con un método distinto al que él utilizará. Está dispuesto a tolerar 0,02 mg/100ml como error máximo entre el valor del universo y la muestra. ¿Cuántos pacientes deben conformar la muestra?

*s* = 0,05 mg/100ml
*P* = 0,02 mg/100ml
*Z* = 1,96

$$n = \frac{Z^2 \times s^2}{P^2} = \frac{1{,}96^2 \times 0{,}05^2}{0{,}02^2} = 24{,}01$$

La muestra debe estar conformada por 24 pacientes.

**Ejemplo 4-7**. Se desea realizar una investigación sobre el desarrollo de enfermedades respiratorias en una población cercana al botadero de basura La Bonanza (vía los Valles del Tuy, Edo. Miranda), en la que anteriormente se ha estimado la que el 20% de la población presenta este tipo de síntomas. Se desea saber cuántas familias deben constituir la muestra, sí el índice buscado varía en más de un 6% con respecto al universo.
*p* = 20%
*q* = (100 - 20) = 80%
*P* = 6%
*Z* = 1,96

$$n = \frac{p \times q}{P^2} \times Z^2 = \frac{20 \times 80}{6^2} \times 1{,}96^2 = 170{,}73$$

La muestra debe estar conformada por 171 familias.

**ACTIVIDADES**
GRUPALES

1. Dado que el error estándar muestra la desviación estándar de la distribución muestral de cualquier estadístico, investigue como se puede calcular el error estándar para otros estadísticos, como la mediana, desviación estándar, varianza y coeficiente de variación.

INDIVIDUALES

1. Una población consiste en cinco números 2, 3, 6, 8 y 11. Considere todas las muestras de tamaño igual a 2 que pueden obtenerse, con reemplazamiento, a partir de esta población. Calcule a) la media de la población, b) la desviación estándar de la población, c) la media de la distribución muestral de medias y d) la desviación estándar de la distribución muestral de medias (es decir, el error estándar de las medias).
2. Resuelva el problema anterior, pero considerando que el muestreo es sin reemplazamiento.
3. Suponga que el peso de 3000 estudiantes universitarios varones se distribuye normalmente, con una media de 68,0 Kg y una desviación estándar de 3,0 Kg. Si se obtienen 80 muestras de 25 estudiantes cada una; ¿cuáles serían la media y la desviación estándar esperadas de la distribución muestral de medias resultante sí los muestreos se hubieran hecho a) con reemplazamiento y b) sin reemplazamiento.
4. En la comunidad de Guaraunos Estado Sucre, se han presentado un gran número de casos de malaria. Determine cuántas personas debe estudiar un investigador, para demostrar la existencia de una endemia, cuando se ha estimado en trabajos anteriores, que la prevalencia de la enfermedad es del 10%. Espera que los resultados obtenidos en la muestra no varíen en más de un 2% con respecto a los valores reales de la población.
5. Se ha encontrado que el valor promedio de mercurio en músculo liso de peces de áreas cercanas a la Refinería El Palito (Estado Carabobo) es de 0,9 µg/100g con una desviación estándar de 0,01 µg/100g. Se desea realizar una investigación en la cual la muestra dé una media de mercurio no mayor del valor real en más de 0,005 µg/100ml. ¿Cuántos peces deben incluirse en la muestra?
6. En una comunidad constituida por 1146 familias se desea realizar un diagnóstico socioambiental, para lo cual se tomará una muestra representativa que admita un error máximo de 5%. ¿Cuántas familias deben constituir la muestra?

## BIBLIOGRAFÍA BÁSICA


- Macchi, R.L. 2001. Introducción a la Estadística en Ciencias de la Salud. Editorial Médica Panamericana. Argentina.
- Milton, J.S. y Tsocos, J.O. 1991. Estadística para Biología y Ciencias de la Salud. McGraw-Hill, Inc.

- Puertas L., E.; Urbina, J.; Blanck, M.E.; Granadillo, D.; Blanchard, M.; García, J.A.; Vargas V.; P. & Chiquito, A. 1998. Bioestadística, Herramienta de la Investigación. Ediciones del Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Carabobo, Venezuela.
- SALAMA, D. 1987. *Estadística: Metodología y aplicaciones.* Editora Principios, Caracas, Venezuela. 308 p.
- SEGNINI, S. 2003. *Apuntes de Estadística para Biólogos*. Dirección de Publicaciones ULA, Mérida, Venezuela.
- Spiegel, M.R. y Stephens, L.J. 2002. Estadística. Serie Schaum. 3era edición. McGraw-Hill, Inc.


**COMPLEMENTARIA**


- Sokal, R.R. y Rohlf, F.J. 1995. Biometry, the principles and practice of statistics in biological research. 3era edición. W.H. Freeman and Company. USA.


**PÁGINAS WEB**


http://www.hrc.es/bioest/M_docente.html

http://www.e-biometria.com/ebiometria/conceptos_basicos/estimacion_estadistica.htm

Dr. Hossein Arsham
http://home.ubalt.edu/ntsbarsh/Business-stat/opre504S.htm#rqualestiunbsuff


## TEMA 5 ESTIMACIÓN

**Competencias:**

Infiere las características de la población a partir de las características de la muestra.

**Contenidos:**

Definición de estimación, estadístico, parámetro, sesgo, estimación por puntos, intervalos de confianza, teorema del límite central.

### La Estimación

Como vimos en el tema anterior, la inferencia estadística permite hacer generalizaciones hacia la población a partir de la información obtenida en una muestra. En este sentido, mediante la inducción es posible obtener un valor representativo de la población, el cual se conoce con el nombre de estimador.

Los resultados de un estimador pueden ser expresados como un simple valor; entendido como una estimación en un punto, o un rango de valores, referido como un intervalo de confianza. Siempre que utilicemos la valoración de un punto, calculamos el margen de error asociado a la estimación de ese punto.

El estimador usual de la media poblacional es $\bar{x} = \Sigma x_i / n$, donde n es el tamaño de la muestra y $x_1$, $x_2$, $x_3$,.....,$x_n$ son los valores de la muestra. Si el valor del estimador en una muestra particular es 5, entonces 5 es la estimación del μ de la media de la población.

Sin embargo, al proceder de esta manera no es posible tener mucha "confianza" en la estimación realizada. Puede haberse tenido la "suerte" suficiente como para extraer de la población un subconjunto de sus integrantes (muestra) en el que se manifieste esa situación. A menos que en la población no haya dispersión o la muestra haya sido infinitamente grande, también puede haberse tenido "mala suerte" de que esos estadísticos sobrestimen o subestimen los parámetros de la población.

La situación podría asemejarse a la "confianza" que se puede tener de "ganar un sorteo" mediante la adquisición de uno de entre todos los números que se sortearán. Sí estos son 100 y tenemos en nuestro poder uno, podríamos indicar que tenemos una confianza de uno en cien (0,01 o 1%) de ganar el premio. Sí se consiguen dos o más esos números podemos duplicar o aumentar nuestra confianza, aunque para transformar esa confianza en seguridad de ganar sería necesario disponer de la totalidad de los números.

La estimación es un proceso mediante el cual, en una muestra se obtiene un determinado valor, denominado estadístico, para luego, en función de é l, calcular (estimar) su valor en la población correspondiente. Recordemos que este valor poblacional recibe el nombre de parámetro.

Parámetro : constante que puede ser calculada con ayuda del modelo de probabilidad de una variable aleatoria o población.

Estimación : procedimiento que permite determinar valores posibles de un parámetro desconocido, a partir de los datos suministrados por muestras extraídas al azar.

Un estimador es cualquier cantidad calculada de los datos de la muestra los cuales se utilizan para obtener información sobre una cantidad desconocida de la población. Por ejemplo, la media muestral es un estimador de la media poblacional.

**Cualidades de un buen Estimador**

Para que resulte de mayor utilidad un buen estimador debe tener: imparcialidad, consistencia, ausencia de sesgo y eficiencia.

1. **Imparcialidad:** Una estimación es imparcial con respecto a un parámetro cuando el valor esperado del estimador puede ser expresado igual al parámetro que ha sido estimado. Por ejemplo, la media de una muestra es una estimación imparcial de la media de la población de la cual la muestra fue obtenida. La imparcialidad es una buena cualidad para una estimación, puesto que, usando el promedio ponderado de varias estimaciones se obtendría una mejor estimación que de cada una de ellas por separado. Por lo tanto, la imparcialidad permite que actualicemos nuestras estimaciones. Por ejemplo, si sus estimaciones de la medias poblacional μ son, digamos 10, y 11,2 con respecto a dos muestras independientes de tamaños 20, y 30 respectivamente, la mejor estimación de la media poblacional μ basada en ambas muestras es [ 20 (10) + 30 (11,2) ] (20 + 30) = 10,75.

2. **Ausencia de sesgo**: Se dice que un estimador es insesgado si la media de la distribución de medias de las muestras, es igual al valor del parámetro estimado. La media $\overline{X}$ es un estimador insesgado de μ.

3. **Consistencia:** como estudiamos en el tema anterior, la desviación estándar de una estimación es llamada el error estándar de esa estimación. Mientras mas grande es el error estándar existirá más error en su estimación. La desviación estándar de una estimación es un índice comúnmente usado del error exigido al estimar un parámetro de la población basado en la información en una muestra de tamaño n escogida al azar de la población entera. Un estimador debe ser "consistente" si al aumentar el tamaño de la muestra se produce una estimación con un error estándar más pequeño. Por lo tanto, su estimación es "consistente" con el tamaño de la muestra. Es decir, realizando un esfuerzo mayor, se obtiene una muestra más grande que produce una mejor estimación. Un estimador consistente es aquel que tiende aproximarse al valor del parámetro de la población, en la medida que el tamaño de la muestra crece.

4. **Eficiencia:** Se refiere a la precisión con la cual tales medidas pueden estimar un parámetro. Una estimación eficiente es la que tiene el error estándar más pequeño entre todos los estimadores imparciales. El "mejor" estimador es el que está más cercano al parámetro de la población que es estimado, aquel que tenga menor error estándar.

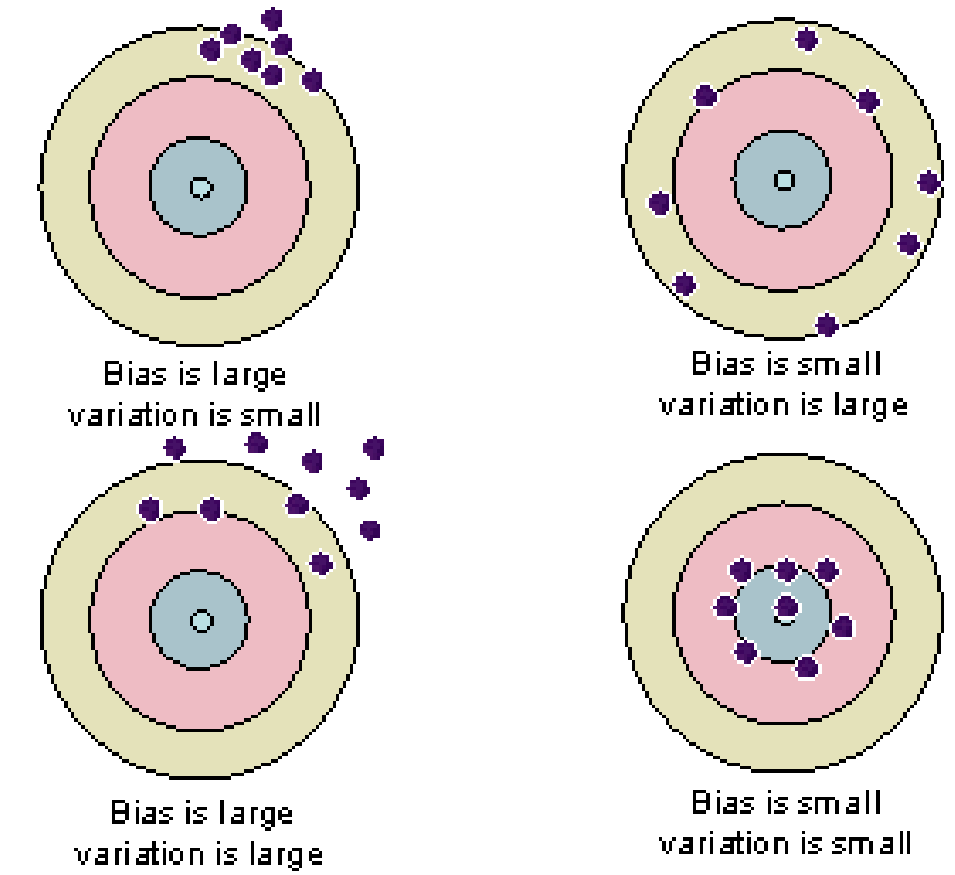



Figura 2. El Concepto de eficiencia para un estimador.

En la Figura 2 se ilustra el concepto de la proximidad por medias que tienen como objetivo el centro para la imparcialidad con varianza mínima. Cada tablero de dardos tiene varias muestras:

El primero tiene todos los tiros agrupados firmemente juntos, pero ningunos de ellos golpean el centro. El segundo tiene una extensión mas grande, pero alrededor del centro. El tercero es peor que los primeros dos. Sólo el último tiene un grupo apretado alrededor del centro, por lo tanto tiene buena eficiencia.

Si un estimador es imparcial, entonces su variabilidad determinará su confiabilidad. Si un perito es extremadamente variable, las estimaciones que produce pueden en promedio no estar tan cerca del parámetro poblacional como lo estaría un estimador parcializado con varianza más pequeña.

**Estimación de parámetros.**

La estimación de parámetros puede efectuarse por puntos o por intervalos. La estimación por puntos plantea un solo valor numérico como parámetro de la población, estimado a partir de una muestra.

Es probable que al considerar un solo punto como estimador de un parámetro se cometa un error, ya que la muestra no es más que una pequeña parte de un conjunto mucho más grande, por lo tanto es aventurado afirmar que el valor correspondiente a la población sea el mismo valor calculado para la muestra. Pero si el número de observaciones es suficientemente grande, se obtendrá una medida muy similar a la del parámetro. Sin embargo, con frecuencia hay limitaciones en cuanto a recurso y tiempo, por lo cual es necesario decidir sólo sobre la base de algunas observaciones, y determinar cuanta probabilidad existe que el valor estimado en la muestra coincida con el valor del parámetro. En este caso, no se estará utilizando el método de estimación puntual sino de intervalo.

Al considerar un estimador $\Theta$ de un parámetro poblacional $\theta$, la realización de una muestra aleatoria de tamaño n, $X_1$, $X_2$,..., $X_n$; suministra n datos, valores u observaciones, $x_1$, $x_2$,..., $x_n$, que determinan una **estimación puntual** del parámetro desconocido:

$\hat{\theta}$, estimación puntual de $\theta$, corresponde al valor del estimador $\Theta$ en $x_1, x_2, \ldots, x_n$.

Si pretendemos, por ejemplo, estimar puntualmente el valor medio $\mu$ con el estimador media muestral, extraeremos una muestra de la población, observaremos el valor de la variable en los **n** individuos de la muestra. En tal caso, los n datos obtenidos $x_1$, $x_2$,..., $x_n$, permiten calcular lo deseado:

$\bar{x}$, estimación puntual de $\mu$, que corresponde al valor del estimador $\bar{X}$ en $x_1, x_2, \ldots, x_n$ :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

La estimación por intervalos consiste en estimar dos valores numéricos extremos, los cuales permiten construir un intervalo, entre cuyos límites se considera está incluido el parámetro a estimar, según el nivel de confianza o de acierto, previamente establecido por el experimentador.

La **estimación por intervalos** de un parámetro $\theta$ consiste en la determinación de un intervalo, que contendrá el parámetro con una confianza $1-\alpha$, número entre 0 y 1, fijado por el experimentador. Para ello se requerirá lo siguiente:

- Una muestra aleatoria $X_1, X_2, \ldots, X_n$ de tamaño n extraída de la población **X.**
- Un estimador $\Theta$ del parámetro poblacional $\theta$, con distribución o ley de probabilidad conocida.

- El nivel de confianza **1-** ⟨, establecido a priori por el experimentador (los usuales son **0.95**, **0.90** y **0.99**).

Una estimación de intervalo de un parámetro, es un segmento en el continuo de la escala de números, donde en algún punto del cual se supone se encuentra el valor del parámetro considerado. Esto significa que en lugar de tener un solo punto como estimación de un parámetro, se tiene ahora todo un conjunto de puntos adyacentes, esto es, un intervalo entre cuyos puntos, probablemente alguno coincida con el valor del parámetro, con nivel de probabilidades de acierto conocido. Fijando de esta manera lo que se denomina un intervalo de confianza; el cual se obtienen mediante la formula:

*Estimador ± (valor crítico x error estándar)*

Ese intervalo numérico se calcula de tal forma que el investigador puede tener una confianza determinada, aunque no la seguridad de que el valor buscado se encuentra dentro de él.

**Estimación de la media de la población**

Para estimar este parámetro se requiere conocer la media aritmética de la muestra, así como su desviación estándar y fijar el nivel de confianza, el cual indica la probabilidad de que el valor del parámetro se encuentre dentro de los límites del intervalo establecido. La expresión matemática queda de la siguiente manera:

$$\overline{X} - Z\frac{s}{\sqrt{n}} > \mu > \overline{X} + Z\frac{s}{\sqrt{n}}$$

o de manera más sencilla:

*intervalo de confianza* = $\overline{X} \pm (Z \times S_x)$

Siendo

$$S_x = \frac{s}{\sqrt{n}}$$

donde:

$\overline{X}$ : media aritmética de la muestra.

*Z* : valor crítico o valor sigma. Se busca en la tabla de áreas de la curva normal, según el nivel de confianza establecido.

$S_x$ : error estándar.

*s* : desviación estándar de la muestra.

*n* : tamaño de la muestra.

**Ejemplo 5-1**. En una investigación acerca del estado nutricional de los escolares de primero a tercer grado, se encontró que los niveles de hemoglobina en ayunas se distribuyen en forma normal, con una media aritmética de 12.38gr%, y una desviación estándar de 0.87gr%. Se desea conocer, con el 95% de confianza, el valor promedio de hemoglobina para esa población de escolares, de donde se extrajo la muestra aleatoria de 144 niños (Puertas y col., 1998).

DATOS:
*X* = 12.38gr%
*s* = 0.87gr%
*n* = 144 niñeos
*nivel de confianza* = 95 % (α = 0.05), el cual equivale a 1.96 sigma (σ).

Aplicando la fórmula del *intervalo de confianza* = $\overline{X} \pm Z \frac{s}{\sqrt{n}}$

= $12.38gr\% \pm 1.96 \frac{087gr\%}{\sqrt{144}}$

= $12.38gr\% \pm 1.96 \text{x} 0.07$

= $12.38gr\% \pm 0.14$ $\left\{ \begin{array}{l} 12.52 \text{ gr\%} \\ 12.24 \text{ gr\%} \end{array} \right.$

Conclusión: En esa población de escolares, la media aritmética de hemoglobina no debe ser menor de 12.24gr%, ni mayor de 12.52gr%. Se hace tal afirmación con 95% de probabilidades de estar en lo cierto, (nivel de confianza) o con un 5 % de riesgo de no acertar o de equivocación (Nivel de significación).

**Ejemplo 5-2**. En una muestra de 350 mujeres se evaluó la edad en la que se presentaron los primeros síntomas de osteoporosis. Se obtuvieron los siguientes estadísticos de esa muestra: media aritmética 48,2 años y desviación estándar 10,2 años. ¿Qué estimación con 95% de confianza puede hacerse con respecto al parámetro media aritmética de la población a partir de estos datos? (Macchi, 2003)

= $48.2\text{años} \pm 1.96 \frac{10.2\text{años}}{\sqrt{350}}$

= $48.2 \text{ años} \pm 1.96 \text{x} 0.55$

= $48.2\text{años} \pm 1.07$ $\left\{ \begin{array}{l} 47.1 \text{ años} \\ 49.3 \text{ años} \end{array} \right.$

En resumen, puede estimarse con 95 % de confianza que el parámetro de la población está entre 47,1 y 49,3.

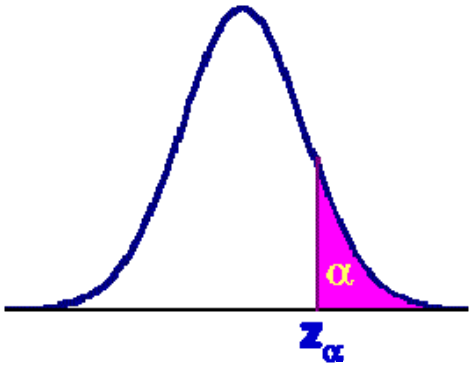
**Distribución muestral de medias**

Si tenemos una muestra aleatoria de una población N(μ,σ ), se sabe (*Teorema del límite central*) que la función de la distribución de la media muestral es también normal con media μ y varianza $\sigma^2/n$. Esto es exacto para poblaciones normales y aproximado (buena aproximación con n>30) para poblaciones cualesquiera. Es decir $\sigma/\sqrt{n}$ es el *error típico, o error estándar de la media*.

¿Cómo usamos esto en nuestro problema de estimación?
1º problema: No hay tablas para cualquier normal, sólo para la normal $\mu=0$ y $\sigma=1$ (la llamada *z*); pero haciendo la transformación (llamada *tipificación*)

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

una normal de media $\mu$ y desviación $\sigma$ se transforma en una *z*.

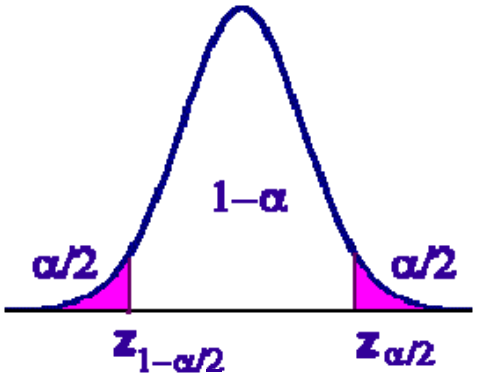



Llamando $z_\alpha$ al valor de una variable normal tipificada que deja a su derecha un área bajo la curva de $\alpha$, es decir, que la probabilidad que la variable sea mayor que ese valor es $\alpha$ (estos son los valores que ofrece la tabla de la normal)

podremos construir intervalos de la forma

$$z_{1-\alpha/2}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<z_{\alpha/2}$$

para los que la probabilidad es 1 - $\alpha$.




Teniendo en cuenta la simetría de la normal y manipulando algebraicamente

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

que también se puede escribir

$$\overline{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

o, haciendo énfasis en que $\sigma/\sqrt{n}$ es el error estándar de la media,

$$\bar{X} \pm z_{\alpha/2} EE(\bar{X})$$

Recuérdese que la probabilidad de que $\mu$ esté en este intervalo es 1 - $\alpha$. A un intervalo de este tipo se le denomina *intervalo de confianza* con un *nivel de confianza* del 100(1 - $\alpha$)%, o *nivel de significación* de 100$\alpha$%. El nivel de confianza habitual es el 95%, en cuyo caso $\alpha$=0,05 y $z_{\alpha/2}$=1,96. Al valor $\bar{X}$ se le denomina estimación puntual y se dice que $\bar{X}$ es un estimador de $\mu$.

**Ejemplo 5-3**. Si de una población normal con varianza 4 se extrae una muestra aleatoria de tamaño 20 en la que se calcula $\bar{X} = 5,3$ se puede decir que $\mu$ tiene una probabilidad de 0,95 de estar comprendida en el intervalo

$$5,3 \pm 1,96 \frac{2}{\sqrt{20}} = (4,42 \quad 6,18)$$

que sería el intervalo de confianza al 95% para $\mu$.

En general esto es poco útil, en los casos en que no se conoce $\mu$ tampoco suele conocerse $\sigma^2$; en el caso más realista de $\sigma^2$ desconocida los intervalos de confianza se construyen con la *t de Student* (otra función de la distribución de probabilidades continua para la que hay tablas) en lugar de la *z*.

$$\bar{X} \pm t_{\alpha n} \frac{s}{\sqrt{n}}$$

o, haciendo énfasis en que $s/\sqrt{n}$ es el error estándar estimado de la media,

$$\bar{X} \pm t_{\alpha n} \hat{EE}(\bar{X})$$

Esta manera de construir los intervalos de confianza sólo es válida si la variable es normal. Cuando *n* es grande (>30) se puede sustituir *t* por *z* sin mucho error.

### Estimación de proporciones

Sea *X* una variable binomial de parámetros *n* y *p* (una variable binomial es el número de *éxitos* en *n* ensayos; en cada ensayo la probabilidad de éxito (*p*) es la misma, por ejemplo: número de diabéticos en 2000 personas). Si *n* es grande y *p* no está próximo a 0 ó 1 ($np \geq 5$) *X* es aproximadamente normal con media *np* y varianza *npq* (siendo *q* = 1 - *p*) y se puede usar el estadístico

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

(proporción muestral), que es también aproximadamente normal, con *error típico* dado por $\sqrt{\frac{pq}{n}}$
en consecuencia, un IC para *p* al 100(1 - $\alpha$)% será

$$\hat{p} \pm z_{\alpha n} \sqrt{\frac{pq}{n}}$$

es decir, la misma estructura que antes:

$$\hat{\theta} \pm z_{\alpha/2} EE\left(\hat{\theta}\right)$$

Obsérvese que para construirlo, ¡se necesita conocer *p*!. Si *n* es grande (>30) se pueden substituir *p* y *q* por sus estimadores sin mucho error, en cualquier caso como $pq \leq 0{,}25$ si se substituye *pq* por 0,25 se obtiene un intervalo más conservador (más grande).

**Ejemplo 5-4**. En una muestra de 100 pacientes sometidos a un cierto tratamiento se obtienen 80 curaciones. Calcular el intervalo de confianza al 95% de la eficacia del tratamiento.

$$\hat{p} = 0{,}80 \qquad \hat{q} = 0{,}20 \qquad n = 100 \qquad z_{0{,}025} = 1{,}96$$

$$IC\,0{,}95\%: \quad 0{,}80 \pm 1{,}96 \sqrt{\frac{0{,}80 \times 0{,}20}{100}} = 0{,}80 \pm 0{,}0784$$

$$\begin{cases} 0{,}7216 \\ 0{,}8784 \end{cases}$$

¿Qué significa este intervalo? La verdadera proporción de curaciones está comprendida entre, aproximadamente, 72% y 88% con un 95% de probabilidad.
¿Es suficientemente preciso? Habrá que juzgarlo con criterios clínicos.

**¿Como se interpreta una confianza del 95%?**

Si llevamos a cabo un experimento 100 veces obtendríamos 100 distribuciones muestrales de datos y 100 intervalos de confianza. De estos 100 intervalos, 95 de ellos cubrirían el valor del verdadero parámetro poblacional. Desgraciada o afortunadamente, nosotros solo realizamos el experimento una sola vez. Con lo que nunca sabremos si nuestro intervalo es uno de esos 95 que contienen el parámetro de estudio.

Técnicamente, aunque esto suene a una sofisticación innecesaria, no podemos asociar el concepto de nivel de confianza con el concepto de probabilidad. Así no se puede establecer que tenemos una probabilidad del 95% de que el parámetro buscado este dentro de nuestro intervalo. Existe una relación entre el tamaño de muestra y el ancho del intervalo de la confianza,

aunado a esto, el intervalo de confianza calculado algunas veces no contiene al valor verdadero.

Digamos que se calcula un intervalo de confianza del 95% para una media μ. La manera de interpretar esto es imaginar un número infinito de muestras de la misma población, el 95% de los intervalos calculados contendrán la media μ de la población, y el 5% no. Sin embargo, es incorrecto indicar, "tengo el 95% de confianza de que la media μ de la población esta dentro del intervalo."

Una vez más la definición usual de un intervalo de confianza del 95% es un intervalo construido por un proceso tal que el intervalo contendrá el valor verdadero el 95% del tiempo. Esto significa que el "95%" es una característica del proceso, no el intervalo.

ACTIVIDADES

INDIVIDUALES

1. Para cada una de las variables presentadas en el siguiente cuadro, estime los respectivos parámetros de la población. Utilice niveles de confianza de 95 y 99%.

   Distribución de medias aritméticas y error estándar de las variables peso, frecuencia cardiaca y porcentaje de antecedentes (Puertas y col., 1998).

| Variables | En una clínica (n=50) | En el hogar (n=35) |
|---|---|---|
| Peso | 70.6 Kg ± 1.5 Kg | 81.3 Kg ±1.9 Kg |
| Frecuencia cardiaca (latidos por minuto) | 72.3 l/m ± 1.6 l/m | 82.9 l/m ±1.9 l/m |
| **Antecedentes familiares positivos** | 34.8 % ± 4.1 % | 41.7 % ± 6.8 % |

2. En una muestra de 400 personas se encontró que el peso promedio era de 67 Kg, con una desviación estándar de 2,5 Kg. Calcule los límites entre los cuales se encuentra el peso verdadero, con un nivel de confianza del 95 %.
3. En una investigación en una comunidad sobre niveles de proteínas totales en la sangre, se tomó una muestra de 15 individuos elegidos al azar en esta comunidad, cuyos resultados fueron: $\overline{X} = 5{,}64$ y $s = 0{,}72$. ¿Cuál será el verdadero promedio de proteínas totales en la sangre para la población donde fue extraída esa muestra? Con un 95 y 99 % de confianza.
4. Al examinar 9 muestras de agua se encontró una concentración de ión nitrato igual a 0,5 μg/ml. Se desea estimar mediante un intervalo de confianza del 95% la concentración promedio del nitrato en el agua, si se sabe que la desviación del método para este análisis es de 0,15 μg/ml.

## BIBLIOGRAFÍA

### BÁSICA


- Macchi, R.L. 2001. Introducción a la Estadística en Ciencias de la Salud. Editorial Médica Panamericana. Argentina.
- Milton, J.S. y Tsocos, J.O. 1991. Estadística para Biología y Ciencias de la Salud. McGraw-Hill, Inc.
- Puertas L., E.; Urbina, J.; Blanck, M.E.; Granadillo, D.; Blanchard, M.; García, J.A.; Vargas V.; P. & Chiquito, A. 1998. Bioestadística, Herramienta de la Investigación. Ediciones del Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Carabobo, Venezuela.
- SALAMA, D. 1987. *Estadística: Metodología y aplicaciones.* Editora Principios, Caracas, Venezuela. 308 p.
- SEGNINI, S. 2003. *Apuntes de Estadística para Biólogos*. Dirección de Publicaciones ULA, Mérida, Venezuela.
- Spiegel, M.R. y Stephens, L.J. 2002. Estadística. Serie Schaum. 3era edición. McGraw-Hill, Inc.


### COMPLEMENTARIA


- Sokal, R.R. y Rohlf, F.J. 1995. Biometry, the principles and practice of statistics in biological research. 3era edición. W.H. Freeman and Company. USA.


### PÁGINAS WEB


http://www.hrc.es/bioest/M_docente.html

http://www.e-biometria.com/ebiometria/conceptos_basicos/estimacion_estadistica.htm

Dr. Hossein Arsham
http://home.ubalt.edu/ntsbarsh/Business-stat/opre504S.htm#rqualestiunbsuff

## CAPITULO III ANALISIS DE VARIANZA

## TEMA 7 ANÁLISIS DE VARIANZA DE UN SOLO FACTOR

### COMPETENCIAS

Determina si una variable (dependiente) es afectada por distintos niveles de otra variable (factor)

### CONTENIDOS

-Introducción al análisis de la varianza
-Análisis de la varianza (ANOVA) de una sola clasificación o vía: Suma total de cuadrados, suma de cuadrados entre grupos, suma de cuadrados intra grupos, grados de libertad, Prueba de F, Pruebas *a posteriori*.

### Introducción al análisis de la varianza (ANOVA)

Muchos experimentos o tomas de muestras se realizan para determinar el efecto que tienen distintos niveles de algún factor de prueba sobre una variable de respuesta. El factor de contraste puede ser la temperatura, el fabricante de un producto, la dosis de fertilizante, el día de la semana, o cualquier otra cosa. Se desea investigar si una variable aleatoria sobre la que se toman muestras es afectada por distintos niveles de un factor, es decir, determinar si diferentes niveles del factor influyen sobre las diferencias en los valores de la variable. Se tomará así una decisión estadística relativa al efecto que tienen los niveles del factor contrastado sobre la variable de respuesta

El análisis de varianza de un factor es el modelo más simple: una única variable nominal independiente con tres o más niveles, explica una variable dependiente continua (existen análisis de varianzas donde se considera el efecto de más de un factor pero no los estudiaremos en este caso). Esta comparación podría realizarse comparando todas las posibles combinaciones de dos en dos de las medias de todos los subgrupos formados, es decir realizar pruebas de t de Student para cada par de medias. Esto trae varios problemas: a) se incrementa el riesgo de dar un resultado falso positivo, al realizar más de un análisis sobre un mismo conjunto de datos, es difícil interpretar la verdadera influencia de la variable que actúa como factor de clasificación porque genera diferentes niveles de significación (p), resultantes de las comparaciones entre sus subgrupos, y es un procedimiento muy largo y engorroso.

Con el análisis de varianza podemos contrastar mas de dos medias simultáneamente, lo que reduce los errores y facilita la comparación.
Mediante el ANOVA se analiza globalmente la influencia de cada variable independiente, generándose un único nivel de significación. (http://www.e-biometria.com/ebiometria/conceptos_basicos/ventajas_y_limitaciones_anova.ht

m)., además se realizan todas las comparaciones entre medias en un solo procedimiento.

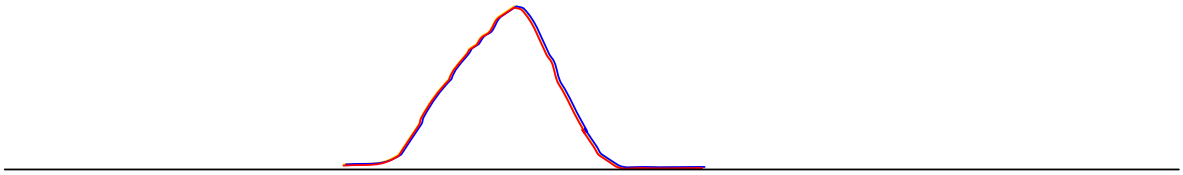
Por lo tanto el análisis de varianza se utiliza para contrastar hipótesis acerca de varias medias poblacionales pero a diferencia de la t de Student, en la prueba de ANOVA se realiza la comparación entre dos varianzas: "VARIANZA ENTRE" y "VARIANZA DENTRO". La varianza entre es la varianza entre grupos, tratamientos o niveles del factor y es la expresión de la variabilidad de los datos entre los grupos por efecto del tratamiento o niveles del factor, mientras que la varianza dentro es la varianza dentro de cada grupo, tratamiento o nivel del factor y es la expresión de la variabilidad de los datos debida a la variable o a la forma en que estos fueron tomados.

Si existe un efecto del tratamiento, se espera que la variabilidad de los datos entre los grupos o niveles del tratamiento sea mayor que la variabilidad dentro de cada grupo, lo que implica que la "varianza entre" debe ser mayor que la "varianza dentro".

En resumen, en un ANOVA la hipótesis nula a probar es:

$$H0: \mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5$$

que gráficamente puede representarse así:



$$\boldsymbol{\mu_1 = \mu_2 = \mu_3}$$

Mientras que las hipótesis alternativas son:

$H_1: \mu_1 = \mu_2$
$H_2: \mu_1 = \mu_3$
$H_3: \mu_1 = \mu_4$
$H_4: \mu_1 = \mu_5$
$H_5: \mu_2 = \mu_3$

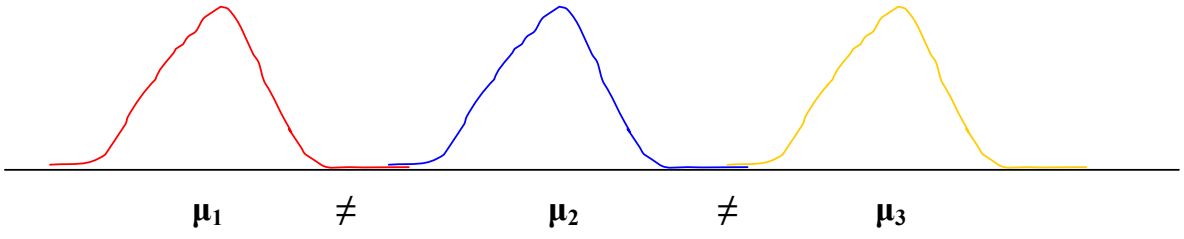
$H_6$:$\mu_2 = \mu_4$
$H_7$:$\mu_2 = \mu_5$
$H_8$:$\mu_3 = \mu_4$
$H_9$:$\mu_3 = \mu_5$
$H_{10}$:$\mu_4 = \mu_5$

Para probar la hipótesis nula $H_0$ sobre la igualdad de las cinco medias se tendría que contrastar cada una de estas 10 hipótesis con la técnica inicial para dos medias. El rechazo de cualquiera de ellas implicará rechazar las hipótesis de igualdad de las cinco medias. El no rechazo de las diez hipótesis acerca de parejas de medias, tendrá como consecuencia el no rechazo de la hipótesis nula principal. Supóngase que se contrastó una hipótesis sobre varias medias contrastando todas las parejas posibles de medias; el error tipo I global sería mucho mayor que el valor α asociado a una sola prueba. Las técnicas del ANOVA permiten contrastar la hipótesis nula (todas las medias son iguales) contra la alternativa (al menos un valor medio es distinto) con un valor de α especificado.

Por lo tanto la hipótesis alternativa para un ANOVA puede resumirse como:

$$H_0: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$$

que gráficamente puede representarse así:




**☞Ejemplo**

Suponga que Ud. desea comparar si los pepinos cosechados en tres localidades bajo distintas condiciones de fertilización (fertilizante químico, abonos verdes, compost) tienen el mismo peso. Para ello Ud. toma una muestra de 10 pepinos en cada una de las cosechas de las localidades y procede a comparar las 3 medias muestrales del peso de pepinos en las 3 localidades.

Si tuviéramos sólo 2 localidades y por ende 2 medias muestrales aplicaríamos una prueba de diferencia de medias (t de student). Pero como tenemos más de dos medias muestrales (en este caso 3) debemos aplicar una prueba de ANOVA.

En primer lugar deben organizarse los datos recolectados en una tabla similar a la Tabla 1 (es recomendable utilizar una hoja de cálculo EXCEL para elaborar la tabla, ya que más adelante deben realizarse algunos cálculos que pueden trabajarse fácilmente en EXCEL).

**TABLA 1 PESO DE LOS PEPINOS EN COSECHAS BAJO DIFERENTES ABONOS**

| N° Pepino | Peso (g) de los pepinos bajo fertilizante químico | Peso (g) de los pepinos bajo abono verde | Peso (g) de los pepinos bajo compost |
|---|---|---|---|
| 1 | 300 | 350 | 360 |
| 2 | 310 | 345 | 365 |
| 3 | 320 | 350 | 360 |
| 4 | 295 | 350 | 365 |
| 5 | 300 | 350 | 355 |
| 6 | 325 | 345 | 360 |
| 7 | 290 | 345 | 350 |
| 8 | 310 | 340 | 355 |
| 9 | 300 | 350 | 360 |
| 10 | 320 | 340 | 365 |

☞El ANOVA requiere que se cumplan los siguientes supuestos o condiciones:

1.- Los efectos debidos al azar así como los factores no contrastados están distribuidos en forma normal y la varianza originada por estos efectos es constante a lo largo del experimento. La variable tiene distribución normal.

2.-Igualdad de la varianza interna en todos los grupos $s^2_1 = s^2_2 = s^2_3 = \ldots\ldots = s^2_a = s^2$ (homocedasticidad)

3.- Independencia de las observaciones: NO debe haber ni autocorrelación entre los valores , ni grupos pareados. La independencia significa que los resultados de una observación del experimento no afectan los resultados de

cualquier otra observación.

Fuente: http://www.fvet.edu.uy/estadis/anova.htm

Antes de realizar una prueba de de ANOVA debe probarse si estos supuestos se cumplen o no. Si se cumplen se procede a realizar la prueba de ANOVA, si no se cumplen debe aplicarse una prueba no paramétrica similar como la prueba de Kruskall-Wallis.

**¿cómo se prueba cada uno de estos supuestos?**

1.-Para probar si una variable tiene o se ajusta a una distribución normal existen varios procedimientos posibles. En las direcciones WEB que se señalan a continuación (anexo) se explican algunas de ellas.

.-http://www.seh-lelha.org/intervalref.htm
.-http://www.seh-lelha.org/noparame.htm
.-
http://descartes.cnice.mecd.es/Bach_HCS_2/distribuciones_probabilidad/aplic_normal.htm

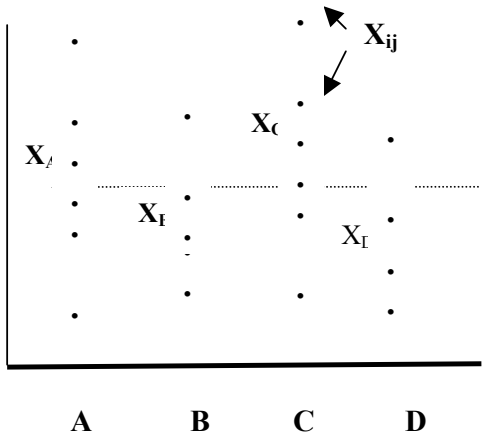
2.-Para probar la homogeneidad (igualdad) de las varianzas de los distintos grupos se realiza la prueba de F (que se explicará mas adelante) entre la mayor y la menor varianza (o desviación estándar) de los grupos. Si se prueba que la mayor y la menor varianza (o desviación estándar) no son estadísticamente diferentes, es decir, son homogéneas, entonces eso implica que todas las demás varianzas (cuyos valores se encuentran entre estas dos) también son homogéneas y se puede aplicar el ANOVA.

3.-La independencia de las observaciones: a la hora de diseñar un experimento o de tomar muestras a las cuales se les desee aplicar una prueba de ANOVA, debe procurarse que los datos de los distintos grupos sean independientes, es decir que los resultados o datos del experimento no afecten los resultados de cualquier otra observación o dato.

## 🕮 Análisis de la varianza: Desarrollo de la prueba

Como se dijo anteriormente en un **ANOVA**:

$$\mathbf{Ho}: m_1 = m_2 = m_3 = m_d = m$$




donde:

$X_{ij}$ es la **j-ésima observación del i-ésimo grupo (cada dato u observación)**

$X_i$ **($X_A$, $X_B$, $X_C$, $X_D$)** es la **media de cada grupo** y

$X_0$ es **la media general de todas las observaciones (gran media)**

Por lo tanto elevando al cuadrado:

$$(X_{ij} - X_0)^2 = (X_{ij} - X_i)^2 + (X_i - X_0)^2 + 2(X_{ij} - X_i)(X_i - X_0)$$

de donde:

$\Sigma (X_i - X_0)^2 = \Sigma (X_{ij} - X_i)^2 + \Sigma (X_i - X_0)^2$, porque la suma de los dobles productos se anula si los grupos son independientes.

Estudiemos cada una de estas sumas de cuadrados cuyas fórmulas son explicadas en el Cuadro 1:

**CUADRO 1**

| | |
|---|---|
| $\Sigma (x_{ij} - x_0)^2$ | **suma de cuadrados total (SCT)** es la suma de los cuadrados de los desvíos de todas las observaciones respecto a su media general. Si la dividimos por **n-1** obtendremos una estimación de la varianza de las observaciones **($S^2_{total}$ )** |
| $\Sigma (x_{ij} - x_i)^2$ | **suma de cuadrados entre grupos (SC entre)** es la suma de los cuadrados de los desvíos de las medias de los grupos respecto a la media general. Si se divide por **a-1** se obtiene otra estimación de la varianza de ***x*** **($S^2_{entre}$ )** |
| $\Sigma (x_{ij} - x_i)^2$ | **suma de cuadrados dentro de grupos (SC dentro)** es la suma de los cuadrados de los desvíos de las observaciones respecto a la media de su respectivo grupo. Al dividirla por **n-a** se obtiene una tercera estimación de la varianza de ***x*** **($S^2_{dentro}$)** |

Debe tenerse en cuenta que las estimaciones $S^2_{entre}$ y $S^2_{dentro}$ son independientes entre sí, pero si bien $S^2_{dentro}$ es **siempre un estimador insesgado**, $S^2_{entre}$ solamente lo será **si la Ho es verdadera**, es decir si las medias son iguales.

**El test se reduce por lo tanto a un ensayo de igualdad de varianzas**

$$F_{experimental} = S^2 \text{ entre} / S^2 \text{ dentro}$$

Se compara este $F_{experimental}$ con el $F_{tabulado}$ con $(\nu_1\text{-}1)$ y $(\nu_2\text{-}1)$ grados de libertad. Si $F_{experimental} > F_{tabulado}$ **se rechaza Ho, lo que implica que hay diferencias estadísticamente significativas entre las medias de los grupos.**

Fuente: http://www.fvet.edu.uy/estadis/anova.htm

Utilizando el ejemplo de la Tabla 1:

xij

xij

xij

**TABLA 1 PESO DE LOS PEPINOS EN COSECHAS CON DIFERENTES ABONOS**

| N° pepino | Peso (g) de los pepinos bajo fertilizante quíimico | Peso (g) de los pepinos bajo abono verde | Peso (g) de los pepinos bajo compost |
|---|---|---|---|
| 1 | 300 | 350 | 360 |
| 2 | 310 | 345 | 365 |
| 3 | 320 | 350 | 360 |
| 4 | 295 | 350 | 365 |
| 5 | 300 | 350 | 355 |
| 6 | 325 | 345 | 360 |
| 7 | 290 | 345 | 350 |
| 8 | 310 | 340 | 355 |
| 9 | 300 | 350 | 360 |
| 10 | 320 | 340 | 365 |

J=10

( $\overline{X}_{i=1}$) ( $\overline{X}_{i=2}$)

( $\overline{X}_{i=3}$)

$(\overline{X}_0)$

**i=** tratamientos o niveles del factor (en este caso 3)
**j=** datos de cada tratamiento (en este caso hay 10 datos en cada tratamiento o nivel)

Fuente: http://www.fvet.edu.uy/estadis/anova.htm

**CUADRO 2 RESUMEN DE LAS FUENTES DE VARIACIÓN, VARIANZAS Y GRADOS DE LIBERTAD EN UN ANOVA**

| FUENTE DE VARIACIÓN | SUMA DE CUADRADOS (SC) | GRADOS DE LIBERTAD (gl) | MEDIA DE CUADRADOS (MC) | $F_{calculado}$ |
|---|---|---|---|---|
| ENTRE GRUPOS | $SC_{ENTRE}$ | a - 1 (*) | $SC_{Entre}$ / a-1 | $MC_{Entre}/MC_{Dentro}$ |
| DENTRO DE GRUPOS | $SC_{DENTRO}$ | n – a | $SC_{Dentro}$/ n-a | |
| TOTAL | $SC_{TOTAL}$ | n – 1 | | |

Fuente: http://www.fvet.edu.uy/estadis/anova.htm

**Prueba de F**

La prueba de F permite establecer si dos varianzas muestrales estiman o no una misma varianza poblacional. Para ello se realiza el cociente de la varianza muestral mayor sobre la varianza muestral menor. El valor obtenido se compara con los valores de la distribución F representados en una tabla de F (Tabla 4 en la sección de "anexos") en la cual se ingresa con los grados de libertad de la varianza del numerador (v1), los grados de libertad de la varianza del denominador (v2) y el nivel de significancia $\alpha$, para obtener un F crítico. Si el valor obtenido en el cociente de las dos varianzas (Fcalculado) es mayor que el valor crítico obtenido en la Tabla de F (Fcrìtico), se cae en la zona de rechazo de $H_0$, por lo que rechazo la hipótesis nula.

El cociente de variables independientes (cociente entre dos $X^2_1 / X^2_2$) cada una distribuida como $S^2$ y dividida por sus respectivos grados de libertad, se distribuye como **F (de Fisher)** con $\nu_1, \nu_2$ grados de libertad

$$F= X^2_1 / X^2_2$$

La distribución existe sólo para los valores NO negativos de **F**, presenta asimetría positiva y tiene dos parámetros :
$\nu_1$ grados de libertad del numerados y $\nu_2$ grados de libertad del denominador

Fuente:
http://www.fvet.edu.uy/estadis/anova.htm

En el caso de la prueba de ANOVA las dos varianzas que se comparan mediante la prueba de F son la "varianza entre grupos O VARIANZA **ENTRE**" y la "varianza intragrupos O VARIANZA **DENTRO**", pero como estas no se conocen se utilizan sus estimadores muestrales $S^2_{dentro}$ y $S^2_{entre}$ por lo tanto se divide la 1ra entre la 2da. Si el cociente es superior al F crítico hallado en la tabla se rechaza la hipótesis nula lo que indica que al menos una de las medias comparadas es diferente al resto.

Para realizar un análisis de varianza Ud. debe calcular el valor de F que resulta de dividir las desviaciones estándar: $S^2$ entre y $S^2$ dentro lo que indica que en primer lugar hay que hallar estas dos desviaciones.

Para hallar dichas desviaciones utilice las fórmulas indicadas en el Cuadro 1 para obtener finalmente $S^2$ entre y $S^2$ dentro. Luego debe realizar el cociente $S^2$ entre / $S^2$ dentro para hallar el valor de F.

Finalmente debe comparar este valor de Fcalculado con el valor de Fcrítico obtenido en la tabla de F tal como se indica en la sección Prueba de F. Si el valor de Fcalculado es superior al Fcrítico hallado en la tabla se rechaza la hipótesis nula lo que indica que al menos una de las medias comparadas es diferente al resto.

**Pruebas a posteriori**

Al rechazar la hipótesis nula en un ANOVA, debe realizarse una prueba más, con la finalidad de determinar cuales son las medias que son diferentes .Este tipo de pruebas se denominan pruebas *a posteriori*. Dentro de estas se encuentran las pruebas de Tukey, Duncan, Diferencia Mínima Significativa, entre otras.

**Prueba de Diferencia Mínima Significativa**

Es un procedimiento usado para comparar cada una de las medias de un conjunto con un tratamiento control.

El valor de **DMS** (**D**iferencia **M**ínima **S**ignificativa) es igual a:

**DMS** = $t.S^2_d$; donde $S_d^2 = \frac{Si^2}{n} + \frac{Si'^2}{n'}$ siendo $Si^2$ y $Si'^2$ las varianzas

estimadas de los experimentos que reciban los tratamientos i e i´ respectivamente, r y ri´ son los números de unidades experimentales que reciben los tratamientos i e i´, respectivamente, t es el valor de t de Student al nivel de significación α escogido y con f grados de libertad asociados con la desviación estándar de la media.

Todas las diferencias entre las medias (que se calculan restándole la menor media a la mayor) son comparadas con la **DMS** calculada. Si la

diferencia excede la **DMS**, se dice que las medias provienen de poblaciones con medias distintas.

A continuación realizaremos un ejemplo de cómo realizar una prueba de ANOVA utilizando los datos de la Tabla 1.

En primer lugar los datos de la Tabla 1 fueron introducidos en una hoja de cálculo de EXCEL. Luego se comprobó que se cumplían todas las condiciones para poder aplicar la prueba de ANOVA.

Finalmente se procedió a realizar en la hoja de cálculo todas las operaciones necesarias.

**EJEMPLO**

| N° PEPINO | Peso (g) pepinos fertilizante quíimico | Peso (g) pepinos abono verde | Peso (g) pepinos compost |
|---|---|---|---|
| 1 | 290 | 350 | 385 |
| 2 | 295 | 345 | 385 |
| 3 | 290 | 350 | 380 |
| 4 | 290 | 350 | 385 |
| 5 | 290 | 350 | 380 |
| 6 | 290 | 345 | 385 |
| 7 | 290 | 345 | 380 |
| 8 | 295 | 340 | 385 |
| 9 | 295 | 350 | 380 |
| 10 | 295 | 340 | 385 |
| | $\overline{X}$ **292** | $\overline{X}$ **346.5** | $\overline{X}$ **383** |
| | **$S^2_1$=6,66** | **$S^2_2$=16,94** | **$S^2_3$=6,67** |

$\overline{X}$ 0 = **340.50**

| **Σ (xij – xo)2** | **Σ (xi – xo)2** | **Σ (xij – xi)2** |
|---|---|---|
| 2550.25 | 2352.25 | 4.00 |
| 2070.25 | 36.00 | 9.00 |
| 2550.25 | 1806.25 | 4.00 |
| 2550.25 | 4194.50 | 4.00 |
| 2550.25 | dividido entre a-1 (3-1) | 4.00 |
| 2550.25 | **2097.25** | 4.00 |
| 2550.25 | **este valor es $S^2$ entre** | 4.00 |
| 2070.25 | | 9.00 |
| 2070.25 | | 9.00 |

| | |
|---|---|
| 2070.25 | 9.00 |
| 90.25 | 12.25 |
| 20.25 | 2.25 |
| 90.25 | 12.25 |
| 90.25 | 12.25 |
| 90.25 | 12.25 |
| 20.25 | 2.25 |
| 20.25 | 2.25 |
| 0.25 | 42.25 |
| 90.25 | 12.25 |
| 0.25 | 42.25 |
| 1980.25 | 30.25 |
| 1980.25 | 30.25 |
| 1560.25 | 0.25 |
| 1980.25 | 30.25 |
| 1560.25 | 0.25 |
| 1980.25 | 30.25 |
| 1560.25 | 0.25 |
| 1980.25 | 30.25 |
| 1560.25 | 0.25 |
| 1980.25 | 30.25 |
| 42217.50 | 395.00 |
| dividido entre (n-1) x a | dividido entre n-a (10-3) |
| **1455.78** | **56.43** |
| **este valor es $S^2$ total** | **este valor es $S^2$ dentro** |

**$F=S^2$ entre / $S^2$ dentro**

***F=2097,25/56,43***

***$F_{calculado}$= 37.16***

***$F_{crìtico}$* (tabla)**

**para (a-1) y (n-1) x a**

**grados de libertad y $\alpha$=0.05**

***(a-1)=(3-1)=2 grados de libertad***

**(n-1) x a= (10-1) 3=27 gl**

***Fcrìtico= 18.6***

**(2,27,0.05)**

**Fcalculado > Fcrìtico**

⇒ **Rechazo Ho**

**lo que significa que al menos uno de los Fertilizantes (tratamientos) produce pepinos con pesos distintos a los otros dos**

⇓

**Entonces procedo a realizar la prueba a posteriori DMS para determinar cual o cuales fertilizantes (tratamientos) producen pepinos con pesos diferentes**

**PRUEBA a posteriori DMS**

| | | x1-x2 | x1-x3 | x2-x3 |
|---|---|---|---|---|
| | $Si^2$ | 4.44 | 4.44 | 4.44 |
| | $Sj^2$ | 28.69 | 4.44 | 28.69 |
| | $Sd^2$ $(Si^2+Sj^2)$ | 33.14 | 8.88 | 33.14 |
| | **Sd** | **5.76** | **2.98** | **5.76** |
| **tcrit(n-1, α)** | | | | |
| **tcrit(9, 0.025)** | **2.82** | | | |
| | **DMS=Sd x t crit** | 5,76 x 2,82 | 2,98 2,82 | 5,76 x 2,82 |
| | | **DMS (x1-x2)** | **DMS (x1-x3)** | **DMS (x2-x3)** |
| ***Diferencia de*** | ***medias*** | **93.43** | **25.05** | **93.47** |
| **(x1-x2)** | **54.5** | **N O** | | |
| **(x1-x3)** | **91** | | **SI** | |
| **(x2-x3)** | **36.5** | | | **NO** |

Estos resultados indican que las diferencias entre X1 y X2 y entre X2 y X3 no son estadísticamente significativas, mientras que la diferencia entre X1 y X3 si es significativa. En otras palabras, los pepinos cultivados con fertilizantes químicos y los cultivados con compost no pesan lo mismo, en promedio. Por otra parte, los pepinos cultivados con abono verde

| | | | | |
|---|---|---|---|---|
| tienen un peso cultivados con compost, y no | promedio que fertilizantes y se puede | es intermedio el de los diferenciar | entre el de cultivados de estos. | los con |
| Si observamos tratamiento diferencia los pepinos | los valores de (las medias) consiste en que cultivados con | peso podemos el peso compost. | promedio en deducir que promedio es | cada esta mayor en |

**Conclusiones:**

**1.-**La prueba de ANOVA indicó que se rechaza la Ho, es decir, las medias poblacionales (estimadas a través de las medias muestrales) de los pesos de los pepinos cultivadas bajo diferentes tratamientos (fertilizantes) no son iguales. En otras palabras, los pepinos cultivados bajo diferentes tratamientos no tienen el mismo peso, en promedio.

**2.-**La prueba a posteriori de **D**iferencia **M**ínima **S**ignificativa indicó que las medias poblacionales (estimadas por las medias muestrales) de peso de pepinos que son diferentes son X1 y X3: peso promedio de pepinos cultivados con fertilizantes y peso promedio de pepinos cultivados con compost, es decir, los pepinos cultivados con fertilizantes y compost no tienen el mismo peso promedio. Por otra parte, el peso de los pepinos cultivados con abono verde no se puede diferenciar estadísticamente del de los cultivados con fertilizantes ni del de los cultivados con compost, lo que permite suponer que el valor se encuentra entre estos dos valores.

**3.-**La observación de los valores promedio de peso de los pepinos cultivados con fertilizantes y con compost, indica que la diferencia entre ambos se debe a que los pepinos cultivados con compost tienen un mayor peso promedio que los pepinos cultivados con fertilizantes.

**4.-** Esta experiencia indica que en este caso particular, es preferible cultivar los pepinos con compost que con fertilizantes debido a que así se obtienen pepinos de mayor peso. Haciendo eso además estaríamos obteniendo una ganancia extra ya que se sabe que al cultivar con compost se minimiza la degradación de suelos y aguas y la afectación negativa de personas, animales, plantas y otros seres vivos que ocurre por el uso de fertilizantes.

**Actividad individual**

Señale en cual de estos casos podría utilizar un ANOVA y explique por que. Tome en cuenta el tipo de variable para esta consideración.

1) Caso: estudio de las cosechas de tomate en distintas localidades. Aunque es la misma especie en las tres localidades en cada una hay variaciones en el tamaño del fruto.

Tamaño del fruto

| | Variedad 1 | Variedad 2 | Variedad 3 |
|---|---|---|---|
| 1 | Grande | Mediano | Pequeño |
| 2 | Mediano | Mediano | Mediano |
| 3 | Grande | Mediano | Pequeño |
| 4 | Grande | Pequeño | Mediano |
| 5 | Mediano | Pequeño | Mediano |
| 6 | Grande | Grande | Pequeño |
| 7 | Mediano | Pequeño | Pequeño |
| 87 | Grande | Mediano | Pequeño |
| 8 | Mediano | Mediano | Mediano |
| 9 | Grande | Mediano | Pequeño |
| 10 | Grande | Mediano | Mediano |

2) Caso: estudio del tipo de suelo más adecuado para el establecimiento de un vivero para la producción de plantas para la reforestación.

Altura (cm) de la planta

| | Suelo tipo 1 | Suelo tipo 2 | Suelo tipo 3 |
|---|---|---|---|
| 1 | 50 | 66 | 50 |
| 2 | 55 | 63 | 55 |
| 3 | 54 | 65 | 52 |
| 4 | 50 | 66 | 50 |
| 5 | 53 | 64 | 51 |
| 6 | 52 | 65 | 49 |
| 7 | 57 | 60 | 50 |
| 87 | 54 | 63 | 48 |
| 8 | 53 | 66 | 51 |
| 9 | 55 | 64 | 50 |
| 10 | 53 | 68 | 49 |

**Actividad grupal**

En grupos de 3 realice una estimación de la cantidad de desechos sólidos generados en tres localidades diferentes. El procedimiento a realizar es el siguiente: seleccione en cada localidad 15 viviendas similares. En cada vivienda selecciones una bolsa plástica tipo “supermercado” o “abasto” y pésela. Al final de la experiencia debe tener 15 valores de peso de desechos,

cada uno correspondiente a una bolsa. En cada vivienda y a una multiplique el valor del peso de una bolsa de desechos por el número total de desechos por el número total de bolsas de desechos generadas en un día (la familia de esa vivienda puede decirle cuantas bolsas de desecho se generan en esa vivienda).

Finalmente una vez realizadas la experiencia en las 3 comunidades Ud. puede elaborar una tabla así:

**Peso (kg) de desechos generado por día**

| | Localidad 1 | Localidad 2 | Localidad 3 |
|---|---|---|---|
| **1** | | | |
| **2** | | | |
| **3** | | | |
| **4** | | | |
| **5** | | | |
| **6** | | | |
| **7** | | | |
| **8** | | | |
| **9** | | | |
| **10** | | | |
| **11** | | | |
| **12** | | | |
| **13** | | | |
| **14** | | | |
| **15** | | | |

Con estos datos verifique las condiciones para la aplicación de una prueba de ANOVA y en caso que estas condiciones se cumplan aplique la prueba y explique los resultados.

## BIBLIOGRAFÍA

### BÀSICA

Chacín, F. 2000. Diseño y análisis de experimentos. UCV-Vicerrectorado Académico. Venezuela.

Machi, R. 2003. Introducción a la Estadística en Ciencias de la Salud. Editorial Médica Panamericana. España.

Sokal, R. y J. Rohlf. 1980. Introducción a la Bioestadística. Editorial Reverté S.A. España.

### PÁGINAS WEB:

http://www.dim.uchile.cl/doc/MA34B/tablas_esta.pdf
http://www.fvet.edu.uy/estadis/anova.htm
http://www.seh-lelha.org/intervalref.htm
http://www.seh-lelha.org/noparame.htm

http://descartes.cnice.mecd.es/Bach_HCS_2/distribuciones_probabilidad/aplic_normal.htm
http://www.e-biometria.com/ebiometria/conceptos_basicos/ventajas_y_limitaciones_anova.htm

## CAPITULO IV REGRESIÓN Y CORRELACIÓN

**Objetivo:**

Establece la relación que existe y la forma o la ecuación mediante la cual se relacionan dos variables

### Tema 8 REGRESION: Estimar una variable a partir de otra.

### COMPETENCIA A LOGRAR:

- Conoce las técnicas para la formulación de las rectas de regresión
- Formulas rectas de regresión para datos agrupados y datos no agrupados
- Predice valores que asumirá una variable respecto de otra

### CONTENIDOS:

### 1. REGRESIÓN

**Que es Regresión?**

El análisis de regresión se utiliza con el propósito de predecir el comportamiento de una variable respecto de otra. El objetivo del análisis de regresión es formular un modelo o ecuación que nos permita predecir el valor de la variable dependiente o de respuesta a partir de los valores de una variable independiente.

### 2. CURVA DE AJUSTE

Según SCHILLER 2000, con frecuencia, en la práctica se encuentra que existe una relación entre dos o mas variables y uno desea expresar esta relación de manera matemática, planteando una ecuación que conecte las variables.

Para realizar esto, podemos ejecutar varios pasos:

- Primero recolectar los datos y mostrando los valores correspondiente a cada variable x, y

- Segundo representar en un sistema de coordenadas rectangulares los valores de los pares ordenados (x, y). Por ejemplo representamos los valores de las variables peso(x), altura (y). De esta representación obtenemos un conjunto de puntos que denominan *Diagrama de dispersión o Nubes de puntos*.

- Tercero a partir del diagrama de dispersión ó nube de puntos es posible visualizar una curva que se aproxima a los datos a la cual denominamos *Curva de aproximación.* Cuando el diagrama de dispersión muestra una aproximación a una línea recta se dice que existe una *relación lineal* entre las variables. Cuando no presenta una aproximación a una relación lineal se dice que presenta una *relación no lineal*

- Cuarto a partir de la relación existente y de los datos podemos encontrar las ecuaciones de curvas de aproximación que se ajusten al conjunto de datos que se llama *Curva de ajuste.*

Las ecuaciones de las curvas de ajustes más importantes son:

- Ecuación de la línea recta $y = a + bx$
- Ecuación de la parábola $y = a + bx + cx^2$
- Ecuación exponencial $y = ab^x$

Dentro de las ecuaciones de las curvas de ajustes más usadas tenemos la línea recta, la cual llamaremos recta de regresión

Ecuación de la línea recta: aquí se nos presentan dos casos:

1. Primer caso: Cuando y es la variable dependiente o valor de la variable a estimar. Con lo cual x será la variable independiente **$y = a + bx$**

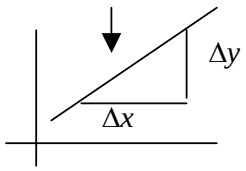
2. Segundo caso: Cuando x es la variable dependiente o valor de la variable a estimar. Con lo cual y será la variable independiente
   **$x = a + by$**

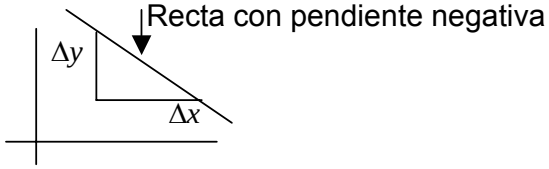
En cuanto **a** y **b** son los parámetros o valores indeterminados dentro de una ecuación, siendo a la intersección con el eje de las ordenadas mientras que b será la pendiente de la recta la cual representa el incremento que sufre la una variable con cada unidad de incremento de la otra.

Cuando b: (+) el incremento o pendiente será positivo
Cuando b: (-) el incremento o pendiente será negativo

$\Delta x = Diferencia(x_2, x_1)$
$\Delta y = Diferencia(y_{2,} y_1)$

Recta con pendiente positiva




3. **Aplicación del Método de los mínimos cuadrados para la recta de regresión: caso de datos no agrupados**

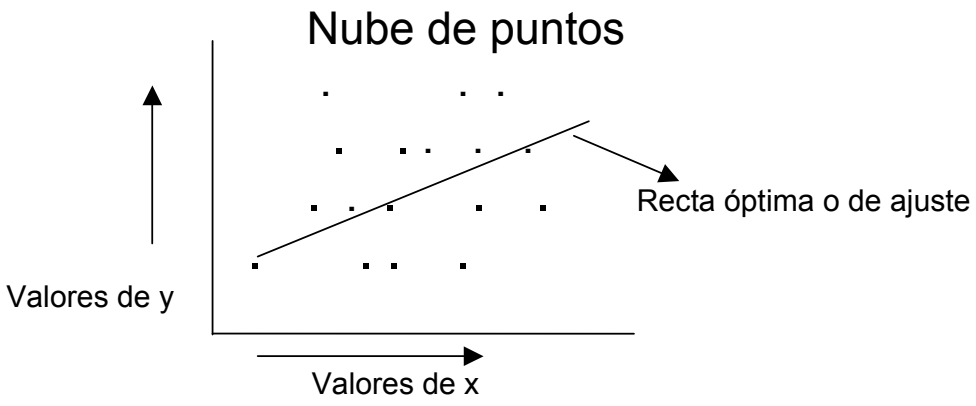
Según BERENSON 2001, para poder formular la ecuación de regresión es necesario determinar los coeficientes a y b de manera de encontrar la recta que

mejor ajusta los datos o que la diferencia entre las variables se hace mínima. La que minimiza la diferencia se conoce como el método de los mínimos cuadrados técnica matemática que nos ayuda a determinar los valores de a y b

SOTO 1982, se basa en la segunda propiedad de la media aritmética la cual dice: La suma de los cuadrados de las desviaciones respecto a la media aritmética es siempre un valor mínimo

$$\sum (X_i - X)^2 = mínimo\,.$$

El método consiste determinar dentro de la infinidad de líneas que existen en un plano, aquella línea recta optima o promedio que pase lo mas cerca posible de todos los puntos originados al representar las dos variables consideradas (nubes de puntos)




### 3.1. Rectas de regresión de y en x

Esta recta nos permite estimar valores de la variable y conocidos los de la variable x a través de la siguiente ecuación.

**y = a + bx**

Determinación de los valores a y b

Partimos de la ecuación y = a + bx, por existir dos parámetros tendremos que generar dos ecuaciones:
La primera se obtiene multiplicando la ecuación por $\sum$ (Sumatoria) de lo cual obtenemos: $\sum y = \sum a + b\sum x$

Donde $\sum a = Na$ , pues según la propiedad de las sumatorias: *La sumatoria de una constante es igual a N veces la constante*

de lo cual se obtiene $\sum y = Na + b\sum x$

La segunda ecuación se obtiene multiplicando la ecuación original por $\sum x$

$$\sum xy = a\sum x + b\sum x^2$$

Agrupamos las ecuaciones resultantes y tenemos:

$$\sum y = Na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

Como podemos ver tenemos dos ecuaciones con dos incógnitas por lo cual podemos aplicar cualquiera de los métodos conocidos, o podemos usar las siguientes formulas

axy= Parámetro a de la recta de regresión de y en x

bxy= Parámetro b de la recta de regresión de x en y

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

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

### 3.2. Calculo de la recta de regresión de x en y

Esta recta nos permite estimar valores de la variable x conocidos los de la variable y a través de la siguiente ecuación.

**x = a + by**

Determinación de los valores a y b

Partimos de la ecuación x= a + by, por existir dos parámetros tendremos que generar dos ecuaciones:
La primera se obtiene multiplicando la ecuación por $\sum$ de lo cual obtenemos:

$$\sum x = \sum a + b\sum y$$

Donde $\sum a = Na$ de los cual se obtiene $\sum x = Na + b\sum y$

La segunda ecuación se obtiene multiplicando la ecuación original por $\sum y$

$$\sum xy = a\sum y + b\sum y^2$$

Agrupamos las ecuaciones resultantes y tenemos:

$$\sum x = Na + b\sum y$$
$$\sum xy = a\sum y + b\sum y^2$$

Como podemos ver tenemos dos ecuaciones con dos incógnitas por lo cual podemos aplicar cualquiera de los método conocidos, o podemos usar las siguientes formulas

axy= Parámetro a de la recta de regresión de x en y

bxy= Parámetro b de la recta de regresión de x en y

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

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

### 3.3. Cálculo del error de estimación

Para el calculo del error de estimación de la recta de regresión x en y

$$Exy = \sqrt{\frac{\sum (X_o - X_c)^2}{N}}$$

Ex.y = Error de estimación de la recta de regresión de x en y
$X_o$= Valores observados de la variable x
$X_c$= Valores calculados de la variable x
N = Nº de pares de variables
Otro método

$$Exy = \sqrt{\frac{\sum X^2 - a\sum X - b\sum XY}{N}}$$

Podemos calcular el error de estimación si conocemos el coeficiente de correlación.

$$Exy = S_x \sqrt{1 - r^2}$$

Calculos del error de estimación de y en x

$$Exy = \sqrt{\frac{\sum (Y_o - Y_c)^2}{N}}$$

Eyx = Error de estimación de la recta de regresión de y en x
$Y_o$= Valores observados de la variable y
$Y_c$= Valores calculados de la variable y
N = Nº de pares de variables
Otro método

$$Eyx = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum YX}{N}}$$

Podemos calcular el error de estimación si conocemos el coeficiente de correlación.

$$Eyx = S_y \sqrt{1 - r^2}$$

### 3.4. Coeficiente de Regresión

Se denominan de esta manera a los parámetros b de cada una de las rectas de regresión, es decir bxy y byx los cuales nos indica el numero unidades que se modifica la variable dependiente por cada una de variación en la variable independiente.

$$r = \sqrt{|bxy.byx|}$$

## 4. Rectas de regresión para datos agrupados en clases

### 4.1. Recta de regresión de x en y

**x = a + by**

Se usará una formula basada en el Coeficiente de Correlación r de manera de obtener mayor utilidad de la tabla de doble entrada

$$X - \mathbf{X} = \frac{rSx}{Sy}(Y - \mathbf{Y})$$

Donde
X = Es la variable dependiente
**X** = Es la media aritmética
r = Coeficiente de correlación
Sx y Sy son la Desviación Estándar de cada una de las variables

Para el calculo de la media aritmética usaremos la formula

$$\mathbf{X} = \mathbf{X}a + \frac{\sum fxdx}{\sum fx} . Ic_x$$

Donde **X**a es el valor de una media arbitraria la asumirá el valor del punto medio donde la diferencia o desvió unitario sea cero. $Ic_x$ , $Ic_y$ son los intervalos de clase de cada variable

$$\mathbf{Y} = \mathbf{Y}a + \frac{\sum fydy}{\sum fy} . Ic_y$$

Para el calculo de la Sx

$$Sx = Ic_x \left( \sqrt{\frac{\sum fxd^2x}{\sum fx} - \left(\frac{\sum fxdx}{\sum fx}\right)} \right)$$

Para el calculo Sy

$$Sy = Ic_y \left( \sqrt{\frac{\sum fyd^2y}{\sum fy} - \left(\frac{\sum fydy}{\sum fy}\right)} \right)$$

Para el calculo de la recta de regresión de y en x **y = a + bx**

$$Y - \mathbf{Y} = \frac{rSy}{Sx}(X - \mathbf{X})$$

Y procederemos igual sustituyendo los valores respectivos

## 5. Actividades:

### 5.1. Individuales:

1. Lea detenidamente la guía y consulte otra Bibliografía recomendada
2. Escriba un ensayo sobre la correlación y su utilidad. Consulte la bibliografía disponible e Internet.
3. Dentro de su familia mas cercana recopile los datos de altura y peso y calcule la recta de regresión respectiva.

### 5.2. Grupal Cooperativo

- Los siguientes datos corresponden a las precipitación promedio mensual y las temperaturas promedio mensual en Caracas durante algunos meses 2004.

| mm. | °T |
|---|---|
| 75 | 20 |
| 90 | 21 |
| 70 | 24 |
| 103 | 24,5 |
| 178 | 25 |
| 215 | 26 |
| 345 | 25,5 |

Calcule: La ecuación de las rectas de regresión, Coeficiente de regresión, Error de estimación.

- Los siguientes datos corresponden a las superficies plantadas y su rendimiento de la producción agrícola en Venezuela para 1999.

| Rubro | Superficie (Ha) | Rendimiento (Kg/Ha) |
|---|---|---|
| Arroz | 149.480 | 4.482 |
| Sorgo | 163.232 | 2.461 |
| Coco | 18.046 | 5.795 |
| Plátano | 64.744 | 8.509 |
| Ajo | 12.560 | 7.189 |
| Tomate | 9.147 | 20.538 |
| Mango | 8.650 | 15.050 |

Calcule: Calcule: La ecuación de las rectas de regresión, Coeficiente de regresión, Error de estimación.

## 6. BIBLIOGRAFÍA

### BÁSICA


BERENSON, M. LEVINE, D. & KREHBIEL, T. Estadística para Administración. *2da Ed. Pearson Prentice Hall*, México.734 p.

FUENLABRADA, IRMA. 2002. Probabilidad y Estadística. *1ra. Ed. McGraw Hill*, México 399 p.

SOTO, ARMANDO. 1982. Iniciación a la Estadística. *Editorial José Marti*. Caracas 395 p.

SPIEGEL, M. SCHILLER, J. & SRINIVASAN, R. 2001. Probabilidad y Estadística. *2da. Ed. McGraw Hill*, Bogota 399 p.

**Tema 9 CORRELACION: Nivel de relación entre las variables.**

**COMPETENCIA A LOGRAR:**

- Conoce las técnicas para formular y calcular el Coeficiente de Correlación entre dos variables
- Formula y calcula el coeficiente de Correlación para datos agrupados y datos no agrupados.
- Establece la relación o correlación existente entre dos variables consideradas.

**CONTENIDOS**:

## 1. CORRELACION

### ¿Qué es Correlación?

Según SOTO, 1982, es el grado de relación, asociación o dependencia que pueda existir entre dos o más variables.
Es frecuente encontrar fenómenos íntimamente ligados ó variables relacionadas con alguna forma de dependencia tales como:

- Entre el tiempo que transcurre para que una persona se adapte a la oscuridad y el nivel de azúcar en su sangre.
- Entre el peso de una persona, su edad y hábitos que lo predisponen a contraer una enfermedad.
- Entre la longitud de la circunferencia y sus radios, cuya relación se expresa mediante la ecuación $C= 2 \pi r$.

Esta relación o dependencia que es de naturaleza cuantitativa puede deberse a diferentes tipos de factores.

## 2. Tipos de relación entre variables:

### a. Relación causal

Cuando los movimientos que experimenta una de las variables dependen de una causa o son efecto del movimiento de la otra variable.
Por ejemplo al presentarse una aumento de nutrientes en una masa de agua aumenta las poblaciones de las especies.
De igual manera al aumentar la población de depredadores de una especie el aumento de la especie depredada se ve limitado.

**b. Relación circunstancial**

En otros casos tenemos que la relación entre las variables depende de una circunstancia común, por ejemplo un proceso social.
Por ejemplo las edades de los contrayentes en el matrimonio (a mayor edad el hombre y a menor edad la mujer), no son un efecto de la variación de la edad del esposo, sino que la correlación es debida a una proceso social que tiene por tendencia el casarse en edades con diferencias no muy significativas.

**c. Relación casuística o aleatoria**

Cuando la relación entre las variables se deben al azar, casualidades o simples coincidencias.
Por ejemplo es una relación casual que al aumentar la producción de hierro en el país aumente también la cantidad de accidentes de transito en todo el país. De esto podemos concluir que no hay lógica alguna en la asociación de los hechos.

Hay que ser muy cuidadoso en la selección adecuada de las variables que tratemos de relacionar, para evitar un mal uso de los tipos de relaciones, descartando aquellas que se presentan contrarias al sentido común según el nivel de nuestro conocimiento de la realidad.
El estudio de la correlación tiene la importancia de permitirnos sintetizar el nivel relación en un solo valor: *El Coeficiente de Correlación.*

Tres son los aspectos principales en el estudio de dos o más variables:

a. La relación o dependencia que pueda existir entre las variables en estudio.
b. La dirección o tipo de relación que hay entre ellas.
c. El nivel de intensidad entre ellas

## 3. Clasificación de la Correlación

**De acuerdo al número de variables consideradas**

a. **Correlación Simple:** Aquella que considera la posible relación entre dos variables.

b. **Correlación múltiple:** Aquella que considera la posible relación entre mas de dos variables.

**De acuerdo a la tendencia de los datos de las variables bajo estudio**

a. **Correlación rectilínea:** Cuando los datos de las variables consideradas presentan una tendencia de una línea recta.

b. **Correlación curvilínea:** Cuando los datos de las variables investigadas tiene una tendencia distinta a una línea recta.

**4. Correlación simple y rectilínea**

Es aquella que estudia la relación, asociación o de pendencia entre dos variables cuyas magnitudes presentan una tendencia en forma de una línea recta.

**5. Coeficiente de Correlación ( r ) para datos no agrupados**

Se define como el indicador cuantitativo de tipo adimensional el cual no indica el tipo y nivel de relación entre dos ó mas variables.

La formula del Coeficiente de correlación simple y rectilínea para datos no agrupados.

**Método de Pearson:**

$$\mathbf{r} = \frac{\sum(dxdy)}{\sqrt{\sum(d^2x)(\sum(d^2y)}}$$

donde dx = x – X ; dy= y – Y

Otra manera de expresar la formula del Coeficiente de correlación según el metodo de Pearson seria:

$$\mathbf{r} = \frac{\sum(dxdy)}{NSxSy}$$

Donde N= numero de pares variables
Sx= Desviación Estándar de los datos de la variable x
Sy= Desviación Estándar de los datos de la variable y

**Covarianza:** Se define como media aritmética del producto de las desviaciones de cada variable con respecto a sus respectivas medias X Y, y que expresaremos mediante el símbolo Sxy

$$Sxy = \frac{\sum(dxdy)}{N}$$

Con lo cual pudiéramos construir una nueva expresión de la formula de Pearson

$$\mathbf{r} = \frac{Sxy}{SxSy}$$

**Tipos de Correlación en cuanto al signo del Coeficiente r**

**Correlación Positiva o directamente proporcional**
Cuando r= (+ ) nos indica que existe una relación directa, esto quiere decir que al modificarse una variable en una dirección la otra se modifica en la misma dirección.
Por ejemplo al aumentar la altura mayor numero del calzado

**Correlación Negativa o inversamente proporcional**
Cuando r= ( - ) nos indica que existe una relación inversa, esto quiere decir que al modificarse una variable en una dirección la otra se modifica en la dirección opuesta.
Por ejemplo al aumentar la altura desde la superficie terrestre menor es la temperatura del ambiente

**Incorrelación**
Cuando el Coeficiente de Correlación es igual a cero r= 0 se dice que no existe relación ó asociación alguna entre las dos variables consideradas. Es decir son carente de relación o dependencia lineal.

**Limites de Variación del Coeficiente de Correlación**

El coeficiente de correlación r puede variar entre (+1) y (-1)
o sea $-1 \leq r \geq +1$

- Cuando r= -1 se dice que existe una correlación negativa perfecta o inversamente proporcional; o sea al cambiar una variable en un sentido la otra la hace en el sentido contrario.

- Cuando r= +1 se dice que tiene una correlación positiva perfecta ó directamente proporcional; o sea al cambiar una variable en un sentido la otra cambia en el mismo sentido.

- Cuando r= 0 se dice que las variables son incorrelacionadas ó con ausencia total de relación, asociación o dependencia entre ellas.ç

El nivel de intensidad del Coeficiente de correlación será más fuerte, mientras mas se aleje r del valor cero
De lo anterior y a escala general podemos decir:

- Si r > 0,30 el coeficiente de Correlación es débil

- Si 0,30< r $\leq$ 0,50 el coeficiente de Correlación es mediano

- Si 0,50< r $\leq$ 0,80 el coeficiente de Correlación es apreciable

- Si 0,80< r ≤ 1,00 el coeficiente de Correlación es fuerte

**Ejemplo de aplicación:**

De la evaluación de las plantas de un sector se encontró cinco especies de árboles (Saman, Caobas, Cedros, Puy y Zapatero) se a cada ejemplar de cada una de las especies se calculo la altura y su diámetro a 1 metro del suelo resultando los siguientes datos promedio para cada especie.

| **Especie** | **Altura promedio** | **Diámetro** |
|---|---|---|
| | m | m |
| Saman | 19 | 1,35 |
| Caobas | 15 | 0.85 |
| Cedro | 18 | 0,95 |
| Puy | 13 | 0,75 |
| Zapatero | 21 | 0,50 |

Se desea conocer:

a. Coeficiente de Correlación por el método de Pearson
b. Coeficiente de correlación mediante el uso de las Desviaciones Estándar
c. Coeficiente de correlación mediante el uso de la Covarianza
d. Interpretar el coeficiente obtenido.

Solución:

a. Calculo del Coeficiente de Correlación mediante uso del método de Pearson

| **m** **Altura** X | **m** **Diámetro** Y | **dx** X-**X** | **dy** Y-**Y** | dxdy | $d^2x$ | $d^2y$ |
|---|---|---|---|---|---|---|
| 19 | 1,35 | 1,8 | 0,47 | 0,846 | 3,240 | 0,220 |
| 15 | 0.85 | - 2,20 | 0,00 | 0,000 | 4,840 | 0,000 |
| 18 | 0,95 | 0,80 | 0,07 | 0,056 | 0,005 | 0,003 |
| 13 | 0,75 | -4,20 | - 0,13 | 0,546 | 0,017 | 0,298 |
| 21 | 0,50 | 2,80 | - 0,38 | - 1,064 | 0,144 | 1,132 |
| $\sum X$ 86 | $\sum Y$ 4,40 | | | $\sum(dxdy)$ 0,384 | $\sum d^2x = 8,246$ | $\sum d^2x = 1,653$ |

Calculo de las medias aritméticas

**X =** $\frac{\sum x_i}{N} = \frac{86}{5} = 17,20$ **; X =** 17,20 **Y =** $\frac{\sum y_i}{N} = \frac{4,40}{5} = 0,88$ **; Y =** 0,88

Aplicamos la formula de Pearson para el calculo del Coeficiente de Correlación

**r =** $\frac{\sum(dxdy)}{\sqrt{\sum(d^2x)(\sum(d^2y)}}$ **;** **r =** $\frac{0,384}{\sqrt{(8,246)(1,653)}}$ ; **r =** 0,104

b. Calculo del Coeficiente mediante el uso de las Desviaciones Estándar

Calculamos las Desviaciones estándar **S** para cada variable

**Sx** = $\sqrt{\frac{\sum (X - X)}{N}}$ ; **Sx** = $\sqrt{\frac{\sum d^2 x}{N}}$ ; **Sx** = $\sqrt{\frac{8,246}{5}}$ ; **Sx** = 1,28

**Sy** = $\sqrt{\frac{\sum (Y - Y)}{N}}$ ; **Sy** = $\sqrt{\frac{\sum d^2 y}{N}}$ ; **Sy**= 0,575

**r =** $\frac{\sum (dxdy)}{NSxSy}$**;** **r =** $\frac{0,384}{(5)(1,28)(0,575)}$**;** **r =** 0,104

c. Determinamos r en función de la Covarianza

Determinamos la Covarianza

**Sxy=** $\frac{\sum (dxdy)}{N}$; **Sxy=** $\frac{0,384}{5}$; **Sxy** = 0,0768

**r =** $\frac{Sxy}{SxSy}$ **; r =** $\frac{0,0768}{(1,28)(0,575)}$**;** **r =** 0,104

d. Como se observa, por cualquiera de las modalidades del método Pearson el resultado es el mismo **r =** + 0,104 lo que indicaría una débil (muy débil) correlación positiva o directamente proporcional entre la altura de los árboles con su diámetro a 1 metro del suelo.

### a. Coeficiente de Correlación ( r ) para datos agrupados

Según Soto 1982, para determinar el coeficiente de correlación de datos agrupados es conveniente construir la Tabla de Correlacion de doble entrada, denominada así por presentar una entrada para la variable x y otra para la variable y.

La variable x se ubica en forma horizontal y la variable de manera vertical en cada caso se incrementan los valores desde el origen hacia los extremos de cada eje.

En la primera fila horizontal se colocará cada clase de la variable x con su respectivo punto medio entre paréntesis ordenadas de menor a mayor ( hacia el extremo)
El la primera columna se colocará cada una de las clases de la variable y con sus respectivos puntos medios.

La formula para calcular el Coeficiente de Correlación r para datos agrupados es:

$$r = \frac{N\sum fxdxdy - (\sum fxdx)(\sum fydy)}{\sqrt{\left[N\sum fxd^2x - (\sum fxdx)^2\right]\left[N\sum fyd^2y - (\sum fydy)^2\right]}}$$

Para obtener los elementos faltantes de la formula tenemos que construir 6 filas y 6 columnas adicionales a la tabla de doble entrada

En el mismo eje de la variable x (eje horizontal) localizaremos a fy , dy, $d^2y$, fydy, $fyd^2y$, $\sum fdydx$

En el mismo eje de la variable y (eje vertical) localizaremos a fx , dx, $d^2x$, fxdx, $fxd^2x$, $\sum fdydx$

fx = Frecuencia absoluta de la variable x

dx = diferencia respecto a la media (desvíos unitarios) de la variable x

$d^2x$ = el cuadrado de la diferencia respecto a la media (desvíos unitarios) de la variable x
fydy = el resultado del producto entre fx y dx
$fxd^2x$ = el resultado del producto entre fx y $d^2x$

$\sum fdydx$ = Sumando de los valores localizado en las cedillas (cuadros) pequeños (Estos valores son el producto entre la f por dx y dy manteniendo el signo resultante.

## 7. ACTIVIDADES PROPUESTAS:

**Individual:**

- Lea con cuidado los contenidos presentados con relación a la correlación y el calculo de su coeficiente
- Escriba un ensayo sobre la correlación y su utilidad. Consulte la bibliografía disponible e Internet.
- Dentro de su familia más cercana recopile los datos de altura y peso y calcule la ecuación recta de regresión respectiva.

**Grupal Cooperativo:**

- Los siguientes datos corresponden a las precipitación promedio mensual y las temperaturas promedio mensual en Caracas durante algunos meses 2004.

| **mm.** | **°T** |
|---|---|
| 76 | 20 |
| 91 | 21 |
| 71 | 24 |
| 104 | 24,5 |
| 178 | 25 |
| 215 | 26 |
| 345 | 25,5 |

**Calcule:** Coeficiente de Correlación por el método de Pearson, Coeficiente de correlación mediante el uso de las Desviaciones Estándar, Coeficiente de correlación mediante el uso de la Covarianza, Interprete el coeficiente de Correlación obtenido.

- Los siguientes datos corresponden a las superficies plantadas y su rendimiento de la producción agrícola en Venezuela para 1999.

| **Rubro** | **Superficie (Ha)** | **Rendimiento (Kg./Ha)** |
|---|---|---|
| Arroz | 149.480 | 4.482 |
| Sorgo | 163.232 | 2.461 |
| Coco | 18.046 | 5.795 |
| Plátano | 64.744 | 8.509 |
| Ajo | 12.560 | 7.189 |
| Tomate | 9.147 | 20.538 |
| Mango | 8.650 | 15.050 |

**Calcule:** Coeficiente de Correlación por el método de Pearson, Coeficiente de correlación mediante el uso de las Desviaciones Estándar, Coeficiente de correlación mediante el uso de la Covarianza, Interprete el coeficiente de Correlación obtenido.

## a. BIBLIOGRAFÍA

### BÁSICA


BERENSON, M. LEVINE, D. & KREHBIEL, T. Estadística para Administración. *2da Ed. Pearson Prentice Hall*, México.734 p.

FUENLABRADA, IRMA. 2002. Probabilidad y Estadística. *1ra. Ed. McGraw Hill*, México 399 p.

SOTO, ARMANDO. 1982. Iniciación a la Estadística. *Editorial José Marti*. Caracas 395 p.

SPIEGEL, M. SCHILLER, J. & SRINIVASAN, R. 2001. Probabilidad y Estadística. *2da. Ed. McGraw Hill*, Bogota 399 p.

## CAPITULO V PRUEBAS NO PARAMETRICAS

**Objetivo:**

Suministrar herramientas prácticas para el tratamientos de datos no numéricos

## Tema 10 PROCEDIMIENTOS PARA PRUEBAS NO PARAMETRICAS

**COMPETENCIA A LOGRAR:**

- Conoce las pruebas no paramètricas
- Aplica las pruebas y métodos para facilitar el tratamiento de datos no numéricos.

**CONTENIDOS:**

En este tema se describen procedimientos para tratamiento de datos no numéricos o sea cualitativos (categorías) llamados métodos no paramétricos donde no hacemos suposiciones de sobre la distribución de la población estudiada.. Estos métodos requieren un tamaño de muestra mayor que los métodos paramétricos y son de gran utilidad cuando la distribución de la muestra estudiada no es próxima a la normal, según MONTGOMERY, 2003 estos métodos son ideales para gran cantidad de casos pues los datos cualitativos en el común de los casos no satisfacen la suposición de normalidad.

### 2. La prueba del Signo

La prueba del signo se utiliza para probar hipótesis sobre la mediana Med. de una distribución continua. Cuando la distribución normal es simétrica la media y la mediana coinciden.
La prueba del signo se puede emplear para probar la hipótesis sobre la media de una distribución normal μ
Supóngase las hipótesis:
**$H_0$:** $\mu = \mu_0$

**$H_1$:** $\mu < \mu_0$

El procedimiento en esta prueba es el siguiente

Formamos las diferencias $X_i - \mu_{0,}$ donde i= 1,2,...n
Si la hipótesis nula **$H_0$:** $\mu = \mu_0$ es verdadera cualquier diferencia $X_i - \mu_{0,}$ tiene la misma probabilidad de ser positiva o negativa, entonces el estadístico de el numero de diferencias positivas $R^{+.}$ La prueba de hipótesis nula es en realidad una prueba de que el número de signos mas es un valor de una variable aleatoria binomial con parámetro p=1/2.. Podemos calcular un valor P para el número observado de signos más $r^+$ directamente de la distribución binomial.

Podemos rechazar la hipótesis $H_0$ a favor de $H_1$, solo si la proporción de signos mas $R^{+.}$ es suficientemente menor que ½ ó equivalente cada vez que el numero observado de signos mas $r^+$ es muy pequeño. Por lo tanto si el valor p calculado p = p ($R^+ < r^+$ cuando p=1/2) es menor o igual que el nivel de significancía **α** , entonces la hipótesis $H_0$ se rechaza y se concluye que $H_1$ es verdadera,

También puede probarse la alternativa bilateral
**$H_0$:** $\mu = \mu_0$
**$H_1$:** $\mu \neq \mu_0$
Se rechaza $H_0$: $\mu = \mu_0$ si la proporción de signos mas difiere significativamente de ½ (0 sea por arriba o por debajo)

**Ejemplo:**

1 Se evalúa la altura de árboles de un bosque de la cual se toma una muestra de 20 árboles seleccionados al azar. Se desea probar que la hipótesis de que la media de la altura es de 20 m, utilizando un nivel de significancia de α= 0.05

Procedimiento:

1.- El parámetro de interés es la media de la distribución de la altura de los árboles

2.- **$H_0$:** $\mu = 20$ m

3.- **$H_1$:** $\mu \neq 20$ m

4.- $\alpha = 0.05$

5.- La tabla de datos

| Observación (i) | Altura de árboles ($x_i$) | Diferencias ($x_i$ - 20) | Signo |
|---|---|---|---|
| 1 | 21,58 | + 1,58 | + |
| 2 | 16,78 | - 3,21 | - |
| 3 | 23,16 | + 3,16 | + |
| 4 | 20,61 | + 0,61 | + |
| 5 | 22,07 | - 2,07 | - |
| 6 | 17,08 | - 2,91 | - |
| 7 | 17,84 | - 2,15 | - |
| 8 | 25,75 | + 5,75 | + |
| 9 | 23,57 | + 3,57 | + |
| 10 | 22,56 | + 2,56 | + |
| 11 | 21,65 | + 1,56 | + |
| 12 | 23,99 | + 3,99 | + |
| 13 | 17,79 | - 2,20 | - |
| 14 | 23,36 | + 3,36 | + |
| 15 | 17,65 | - 2,34 | - |
| 16 | 20,53 | + 0,53 | + |
| 17 | 24,14 | + 4,14 | + |
| 18 | 22,00 | + 2,00 | + |
| 19 | 26,54 | + 6,54 | + |
| 20 | 17,53 | + 2,46 | + |

6.- El numero observado de diferencia en es la distribución $r^+ = 14$

Calculamos el valor de p puesto que $r^+ = 14$ es mayor que n/2 = 20/2 = 10 el valor de p se calcula

$P = 2P\ (R^+ \geq 14$ cuando $p=1/2$

P= 2P(R+(14 cuando p=1/2)

$$P=2\sum_{r=14}^{20}\binom{20}{r}(0,5)^{r}(0,5)^{20-r}$$

P= 0,1153

Conclusión: Como P= 0,1153 es menor que $\alpha = 0.05$, no es posible rechazar la hipótesis nula de que la media de altura de los árboles es de 20 m

2 Prueba del signo para muestras pareadas

Es una prueba sencilla para el caso de muestras pareadas. Esta prueba consiste en tomar la diferencia entre los datos pareados y escribir solamente el signo de la diferencia.

3 Prueba de Rango con signo de Wilcoxon

La prueba del signo solo emplea los signos mas y menos de las diferencias entre las observaciones y la mediana ( o los signos mas y menos de las diferencias entre las observaciones en el caso de pareado). La prueba no toma en cuenta la magnitud de estas diferencias.

Wilcoxon propuso un procedimiento de prueba que toma en cuenta el signo (el sentido) y la magnitud (la diferencias).La prueba de rango de Wilcoxon se utiliza para el caso de distribuciones continuas simétricas por lo tanto es teste caso la media es igual a la mediana

4 Aproximación para grandes muestras

Si el tamaño de la muestra es moderadamente grande ( n >20) entonces puede demostrarse que $W^+$ tiene aproximadamente una distribución normal con media

$$\mu_{w+} = (n(n+1))/4$$

y con varianza

$$\sigma^2_{w+} = (n(n+1)(2n+1))/24$$

Por lo tanto una prueba de $H_0 : \mu = \mu_0$ puede basarse en el estadistico

$Z_0 = ((W^+ - n(n+1))/4)/\sqrt{(n(n+1)(2n+1))/24}$

**Ejercicios:**

Considere los datos del ejercicio anterior y suponga que la distribución es simétrica y continua. Utilice la prueba con rango de Wilcoxon con $\alpha = 0.05$

Para probar la hipótesis de $\mathbf{H_0\ \mu = 7}$ **contra** $\mathbf{H_1\ \mu \neq 7}$

**Bibliografía**

MONTGOMERY, D. & RUNGER, GEORGE. 2003. Probabilidad y estadística : Aplicadas a la Ingeniería. Mc Graw – Hill, Mexico895 p.

**ANEXO 1**

**PRUEBAS PARA EL AJUSTE DE UNA VARIABLE A UNA DISTRIBUCIÓN NORMAL**

1.-http://www.seh-elha.org/noparame.htm

Asociación de la Sociedad Española de Hipertensión
Liga Española para la lucha contra la Hipertensión Arterial

## ¿Y si los datos no siguen una distribución normal?...

### Bondad de ajuste a una normal.
### Transformaciones.
### Pruebas no paramètricas

*Preparado por Luis M. Molinero (Alce Ingeniería)*

CorreoE: **bioestadistica alceingenieria.net**

*Julio 2003*

Artículo en formato PDF

www.seh-lelha.org/stat1.htm

Cuando se analizan datos medidos por una variable cuantitativa continua, las pruebas estadísticas de estimación y contraste frecuentemente empleadas se basan en suponer que se ha obtenido una muestra aleatoria de una distribución de probabilidad de tipo normal o de Gauss. Pero en muchas ocasiones esta suposición no resulta válida, y en otras la sospecha de que no sea adecuada no resulta fácil de comprobar, por tratarse de muestras pequeñas. En estos casos disponemos de dos posibles mecanismos: los datos se pueden **transformar** de tal manera que sigan una distribución normal, o bien se puede acudir a pruebas estadísticas que no se basan en ninguna suposición en cuanto a la distribución de probabilidad a partir de la que fueron obtenidos los datos, y por ello se denominan **pruebas no paramétricas** (*distribution free*), mientras que las pruebas que suponen una distribución de probabilidad determinada para los datos se denominan pruebas paramétricas.

Dentro de las pruebas paramétricas, las más habituales se basan en la **distribución de probabilidad normal**, y al estimar los parámetros del modelo se supone que los datos constituyen una muestra aleatoria de esa distribución, por lo que la elección del estimador y el cálculo de la precisión de la estimación, elementos básicos para construir intervalos de confianza y contrastar hipótesis, dependen del modelo probabilístico supuesto.

Cuando un procedimiento estadístico es poco sensible a alteraciones en el modelo probabilístico supuesto, es decir que los resultados obtenidos son aproximadamente válidos cuando éste varía, se dice que es un procedimiento **robusto**.

Las inferencias en cuanto a las medias son en general robustas, por lo que si el tamaño de muestra es grande, los intervalos de confianza y contrastes basados en la *t de Student* son aproximadamente válidos, con independencia de la verdadera distribución de probabilidad de los datos; pero si ésta distribución no es normal, los resultados de la estimación serán poco precisos.

## Procedimientos para verificar el ajuste a una distribución de probabilidad

Existen diferentes pruebas para verificar el ajuste de nuestros datos a una distribución de probabilidad. Las dos más utilizadas son el contraste $\chi^2$ **de Pearson**, y la **prueba de Kolmogorov-Smirnov**.

### Contraste $\chi^2$ de Pearson

La idea del contraste de Pearson es muy sencilla: se agrupan los datos en $k$ clases ($k \geq 5$), como si fuéramos a construir un histograma, cubriendo todo el rango posible de valores, siendo deseable disponer, aproximadamente, del mismo número de datos en cada clase y al menos de tres datos en cada una.

Llamamos $Oi$ al número de datos observado en la clase $i$. Mediante el modelo de probabilidad que se desea verificar se calcula la probabilidad $Pi$ asignada a cada clase, y por lo tanto, para una muestra de $n$ datos, la frecuencia esperada según ese modelo de probabilidad es $Ei=n.Pi$.

Se calcula entonces el siguiente índice de discrepancia entre las frecuencias observadas y las que era previsible encontrar si el modelo fuera el adecuado:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

que se distribuye aproximadamente como una $\chi^2$ si el modelo es correcto.

Si el modelo se especifica de forma completa con las probabilidades $Pi$, conocidas antes de tomar los datos, el número de grados de libertad es $k-1$. Pero si se han estimado $r$ parámetros del modelo a partir de los datos, entonces los grados de libertad son $k-r-1$.

### Prueba de Kolmogorov-Smirnov

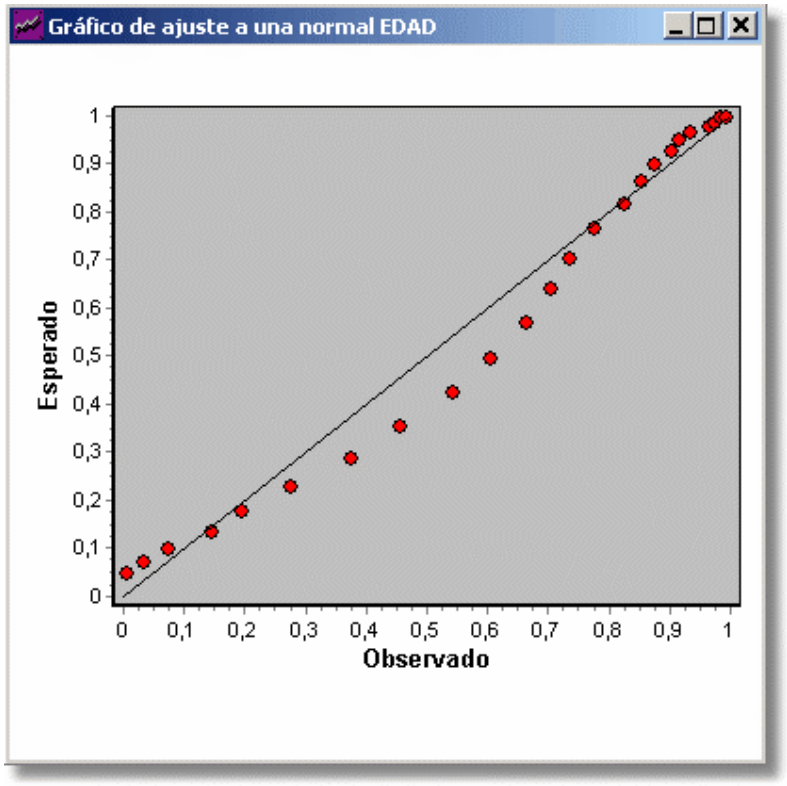
Este contraste, que es válido únicamente para variables continuas, compara la función de distribución (probabilidad acumulada) teórica con la observada, y calcula un valor de discrepancia, representado habitualmente como $D$, que corresponde a la discrepancia máxima en valor absoluto entre la distribución observada y la distribución teórica, proporcionando asimismo un valor de probabilidad $P$, que corresponde, si estamos verificando un ajuste a la distribución normal, a la probabilidad de obtener una distribución que discrepe tanto como la observada si verdaderamente se hubiera obtenido una muestra aleatoria, de tamaño $n$, de una distribución normal. Si esa probabilidad es grande no habrá por tanto razones estadísticas para suponer que nuestros

datos no proceden de una distribución, mientras que si es muy pequeña, no será aceptable suponer ese modelo probabilístico para los datos.

- **Prueba de Shapiro-Wilks**

Aunque esta prueba es menos conocida es la que **se recomienda** para contrastar el ajuste de nuestros datos a una distribución normal, sobre todo cuando la muestra es pequeña ($n<30$).

Mide el ajuste de la muestra a una recta, al dibujarla en papel probabilístico normal. Este tipo de representación también lo proporcionan algunos programas de estadística, de tal manera que nos permite además apreciar el ajuste o desajuste de forma visual:




En escala probabilística normal se representa en el eje horizontal, para cada valor observado en nuestros datos, la función de distribución o probabilidad acumulada observada, y en el eje vertical la prevista por el modelo de distribución normal. Si el ajuste es bueno, los puntos se deben distribuir aproximadamente según una recta a 45º. En la imagen vemos que en este ejemplo existe cierta discrepancia.

En cualquier caso siempre es adecuado efectuar una representación gráfica de tipo histograma de los datos, y comparar el valor de la media y la mediana, así como evaluar el coeficiente de asimetría y apuntamiento, además de llevar a cabo una representación en escala probabilística de la distribución de probabilidad esperada versus observada, como la de la figura.

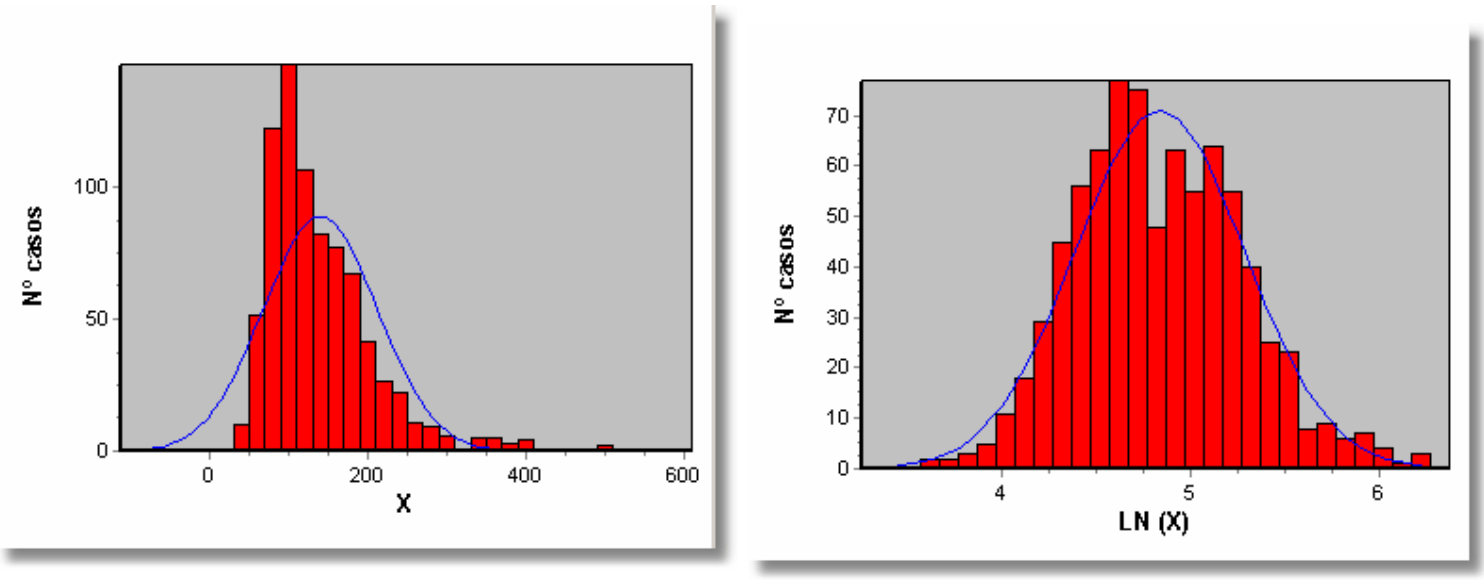
## Posibles soluciones cuando se rechaza la hipótesis de normalidad

Si rechazamos o dudamos de la normalidad de nuestros datos, existen varias soluciones posibles:

- Si la distribución es más apuntada que la normal (mayor parte de los valores agrupados en torno de la media y colas más largas en los extremos), se debe investigar la presencia de heterogeneidad en los datos y de posibles valores atípicos o errores en los datos. La solución puede ser emplear pruebas no paramétricas.
- Si la distribución es unimodal y asimétrica, la solución más simple y efectiva suele ser utilizar una transformación para convertir los datos en normales.
- Cuando la distribución no es unimodal hay que investigar la presencia de heterogeneidad, ya que en estos casos la utilización de transformaciones no es adecuada y los métodos no paramétricos pueden también no serlo.
- Una alternativa muy interesante a los métodos paramétricos y a las pruebas no paramétricas clásicas, la constituye la metodología de **estimación autosuficiente** ya esbozada en otro artículo de esta serie.

## Transformaciones para conseguir datos normales

La utilización de transformaciones para lograr que los datos se ajusten a una distribución normal es en muchas ocasiones la solución más natural, ya que existen gran cantidad de parámetros biológicos que tienen una distribución asimétrica como la de la figura de la izquierda, y que se convierten en aproximadamente simétricas al transformarlas mediante el **logaritmo**.




Tenemos problemas con la transformación logarítmica ln(x) si la variable puede tomar el valor 0, por lo que en esos casos, o incluso si existen valores muy pequeños, será adecuado emplear la transformación ln(x+1). Cuando la desviación típica de los datos es proporcional a la media o cuando el efecto de los factores es multiplicativo, en lugar de aditivo, está indicado el uso de la transformación logarítmica.

Otra transformación posible es $\sqrt{x}$, que es aplicable cuando las varianzas son proporcionales a la media, lo que ocurre a menudo cuando los datos provienen de una distribución de Poisson (recuentos).

Otra transformación habitualmente empleada es $1/x$, que también precisa que sumemos una cantidad a cada valor si existen ceros.

Estas tres transformaciones comprimen los valores altos de los datos y expanden los bajos, en sentido creciente en el siguiente orden: $\sqrt{x}$ (la que menos), $\ln x$, $1/x$.

Si la concentración de datos está, a diferencia de la figura anterior, en el lado de la derecha y la cola en la izquierda, se puede utilizar la transformación $x^2$, que comprime la escala para valores pequeños y la expande para valores altos.

Cuando los datos son proporciones o porcentajes de una distribución binomial, las diferencias con una distribución normal son más acusadas para valores pequeños o grandes de las proporciones, utilizándose entonces transformaciones basadas en $arcsen\sqrt{p}$.

En todos los casos para los cálculos estadísticos basados en la teoría normal, se utilizarán los valores transformados, pero después para la presentación de los resultados se efectuará la transformación inversa para presentarlos en su escala de medida natural.

Más abajo se proporcionan algunos enlaces sobre el tema de las transformaciones, de fácil lectura.

## Pruebas no paramétricas

Se denominan pruebas no paramétricas aquellas que no presuponen una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre (*distribution free*). En la mayor parte de ellas los resultados estadísticos se derivan únicamente a partir de procedimientos de ordenación y recuento, por lo que su base lógica es de fácil comprensión. Cuando trabajamos con muestras pequeñas ($n \leq 10$) en las que se desconoce si es válido suponer la normalidad de los datos, conviene utilizar pruebas no paramétricas, al menos para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal.

En estos casos se emplea como parámetro de centralización la **mediana**, que es aquel punto para el que el valor de X está el 50% de las veces por debajo y el 50% por encima.

Vamos a comentar la filosofía de alguna de las pruebas no paramétricas y en los enlaces se puede aumentar esta información.

### Prueba de Wilcoxon de los rangos con signo

Esta prueba nos permite comparar nuestros datos con una mediana teórica (por ejemplo un valor publicado en un artículo).

Llamemos $M_0$ a la mediana frente a la que vamos a contrastar nuestros datos, y sea *X1, X2 .. Xn* los valores observados. Se calcula las diferencias $X1-M_0$, $X2-M_0$, ..., $Xn-M_0$. Si la hipótesis nula fuera cierta estas diferencias se distribuirían de forma simétrica en torno a cero.

Para efectuar esta prueba se calculan las diferencias en valor absoluto $|Xi-M_0|$ y se ordenan de menor a mayor, asignándoles su rango (número de orden). Si hubiera dos o más diferencias con igual valor (empates), se les asigna el rango medio (es decir que si tenemos un empate en las posiciones 2 y 3 se les asigna el valor 2.5 a ambas). Ahora calculamos *R+* la suma de todos los rangos de las diferencias positivas, aquellas en las que *Xi* es mayor que $M_0$ y *R-* la suma de todos los rangos correspondientes a las diferencias negativas. Si la hipótesis nula es cierta ambos estadísticos deberán ser parecidos, mientras que si nuestros datos tienen a ser más altos que la mediana $M_0$, se reflejará en un valor mayor de *R+*, y al contrario si son más bajos. Se trata de contrastar si la menor de las sumas de rangos es excesivamente pequeña para ser atribuida al azar, o, lo que es equivalente, si la mayor de las dos sumas de rangos es excesivamente grande.

### Prueba de Wilcoxon para contrastar datos pareados

El mismo razonamiento lo podemos aplicar cuando tenemos una muestra de parejas de valores, por ejemplo antes y después del tratamiento, que podemos denominar *(X1,Y1), (X2,Y2), ... ,(Xn,Yn)*. De la misma forma, ahora calcularemos las diferencias *X1-Y1, X2-Y2, ... , Xn-Yn* y las ordenaremos en valor absoluto, asignándoles el rango correspondiente. Calculamos *R+* la suma de rangos positivos (cuando *Xi* es mayor que *Yi*), y la suma de rangos negativos *R-*. Ahora la hipótesis nula es que esas diferencias proceden de una distribución simétrica en torno a cero y si fuera cierta los valores de *R+* y *R-* serán parecidos.

### Prueba de Mann-Whitney para muestras independientes

Si tenemos dos series de valores de una variable continua obtenidas en dos muestras independientes: *X1, X2, ... , Xn, Y1, Y2, ... , Ym*, procederemos a ordenar conjuntamente todos los valores en sentido creciente, asignándoles su rango, corrigiendo con el rango medio los empates. Calculamos luego la suma de rangos para las observaciones de la primera muestra *Sx*, y la suma de rangos de la segunda muestra *Sy*. Si los valores de la población de la que se extrajo la muestra aleatoria de X se localizan por debajo de los valores de Y, entonces la muestra de X tendrá probablemente rangos más bajos, lo que se reflejará en un valor menor de *Sx* del teóricamente probable. Si la menor de las sumas de rangos es excesivamente baja, muy improbable en el caso de que fuera cierta la hipótesis nula, ésta será rechazada.

Existen más pruebas no paramétricas de entre las que a continuación mencionamos las más habituales, remitiendo al lector interesado a cualquier libro básico de bioestadística:

- Prueba de Kruskal-Wallis para comparar K muestras
- Prueba de Friedman para comparar K muestras pareadas (bloques)
- Coeficiente de correlación de Spearman para rangos
- Prueba de rachas de Wald-Wolfowitz

## Enlaces de interés

- The normal distribution Douglas G Altman & J Martin Bland BMJ 1995;310:298 (4 Febrero)
- Transforming data J Martin Bland & Douglas G Altman BMJ 1996;312:770 (23 March)
- Transformations, means, and confidence intervals J Martin Bland & Douglas G Altman BMJ 1996;312:1079 (27 April)
- The use of transformation when comparing two means J Martin Bland & Douglas G Altman BMJ 1996;312:1153 (4 May)
- Algunas direcciones que permiten efectuar cálculos de pruebas no paramétricas:

  - Prueba de los signos
  - Contraste para la mediana de dos muestras independientes
  - Prueba de Wilcoxon para datos pareados
  - Prueba de Mann-Whitney para comparar dos muestras independientes
    Otro enlace para esta prueba
  - Prueba de rachas de Wald-Wolfowitz
  - Prueba de Kruskall-Wallis
  - Prueba de Friedman
  - Coeficiente de correlación de Spearman

## Bibliografía seleccionada


- J.S. Milton, J.O. Tsokos. **Estadística para biología y ciencias de la salud**. Madrid: Interamericana-McGraw Hill; 1989



Indice de artículos

Arriba


2.-http://www.seh-lelha.org/intervalref.htm

Asociación de la Sociedad Española de Hipertensión
Liga Española para la lucha contra la Hipertensión Arterial

## Estimación de intervalos de referencia de variables biológicas


*Preparado por Luis M. Molinero (Alce Ingeniería)*


Artículo en formato PDF

www.seh-lelha.org/stat1.htm

CorreoE: **bioestadistica alceingenieria.net**

*Febrero 2004*

## Introducción

Los gráficos y las tablas de percentiles constituyen una herramienta de uso común en la práctica clínica. Se denomina **intervalo de referencia** a una pareja de valores que corresponden a los límites de determinados percentiles de la distribución de probabilidad de los datos, y que son simétricos con respecto a la mediana.

Evidentemente para establecer intervalos de referencia es fundamental emplear una muestra adecuada, tanto desde el punto de vista de la representatividad de la población que se desea cuantificar, habiendo sido obtenida mediante algún procedimiento de muestreo aleatorio, como en cuanto al tamaño de la misma, que permita efectuar las estimaciones con una adecuada precisión.

En ocasiones también se habla de "*intervalo o rango de normalidad*", aunque esta terminología afortunadamente va cayendo en desuso por confusa e inapropiada, ya que en un intervalo del 95% obtenido a partir de una población sana, por definición, el 5% de los individuos estarán fuera de ese denominado *intervalo de normalidad*. Por otro lado nada impide el determinar intervalos de referencia para poblaciones de enfermos con una patología concreta, donde por tanto el término normal constituye en cierta medida un contrasentido. Además la palabra normal nos induce rápidamente a pensar en una distribución de probabilidad normal o gaussiana, cuando lo más habitual es que los datos que estamos midiendo no se ajusten en principio a ese tipo de distribución de probabilidad, sobre todo cuando se maneja determinaciones analíticas.

## Estimación de intervalos de referencia

A partir de una muestra de $n$ sujetos se trata de estimar un intervalo de referencia del $q$ %, donde frecuentemente $q$ es el $95\%$ o el $90\%$. Se trata de estimar los percentiles $(100-q)/2$ y $(100+q)/2$, que corresponden al $2.5\%$ y $97.5\%$ para un intervalo de referencia del $95\%$.

La forma más simple de estimar esos percentiles es calcularlos directamente a partir de la distribución de nuestros datos. Procedemos entonces a ordenar los valores en sentido creciente, y para un intervalo del $95\%$ el límite superior vendrá dado por aquel valor que deja por debajo el $97.5\%$ de los datos y por encima el $2.5\%$ restante. Si debido al número de observaciones ese punto no coincide exactamente con un valor de la serie, se obtiene por interpolación. El problema de este método radica en que produce estimaciones sesgadas, sobre todo si las muestras no son muy grandes, por lo que se prefiere utilizar otros procedimientos, disponiendo de dos alternativas: utilizar modelos paramétricos o utilizar técnicas no paramétricas.

En cuanto a las técnicas paramétricas las más empleadas se basan en suponer una distribución normal o de Gauss para los datos. Una vez estimada la media $m$ y la desviación estándar $s$, a partir de nuestros datos, los percentiles se estimarán a partir del modelo de distribución de probabilidad normal como $m+z_p s$, donde $z_p$ es el valor de la función de distribución normal correspondiente al percentil $p$. Así para el percentil *97.5%* el valor de $z_p$ es *1.96*

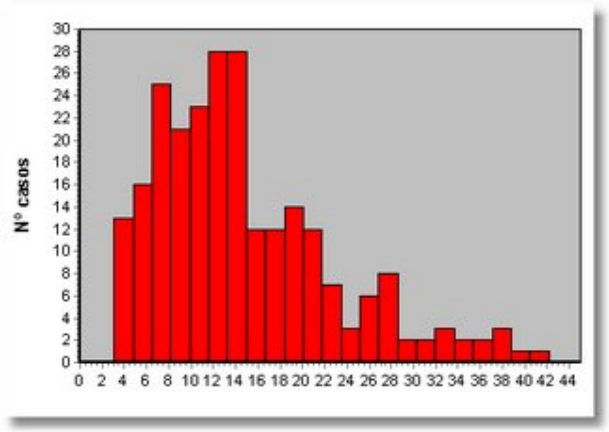



Figura 1

En la práctica no es adecuado utilizar directamente el modelo de probabilidad normal, ya que la mayoría de parámetros biológicos suelen alejarse de ese modelo (como en la figura 1), presentando asimetría (*skewness*) más o menos marcada, ya sea con colas hacia el lado izquierdo de la distribución (asimetría negativa) o hacia el lado derecho (asimetría positiva). En otras situaciones, aunque sí exista simetría, sin embargo la densidad de probabilidad de los datos es diferente de la que correspondería a una distribución normal, bien porque presenta una mayor agrupación de valores en torno a ese valor central (mayor apuntamiento), o porque, al contrario, presenta una distribución de probabilidad más "aplanada" o extendida. Esta característica, relativa a cómo se reparte la frecuencia entre el centro y los extremos de la distribución, se denomina **apuntamiento** o **curtosis**. A veces con una transformación sencilla es suficiente para lograr una variable modificada que sí sigue una distribución de probabilidad normal.

Cuando tenemos asimetría hacia el lado derecho, la transformación logarítmica puede ser adecuada. Otras transformaciones sencillas consisten en utilizar la raíz cuadrada, o la función inversa 1/x.

Un tipo de transformación muy empleado es la de Box-Cox, que tiene la siguiente expresión

$$Y = \frac{X^{\lambda} - 1}{\lambda}$$

cuando $\lambda$ =0 la transformación corresponde a ln(X). El parámetro $\lambda$ se estima por el procedimiento de máxima verosimilitud.

En ocasiones es necesario aplicar dos transformaciones: una para eliminar la asimetría y luego otra para eliminar la curtosis.

Para verificar si los datos siguen o no una distribución normal, se suele utilizar los gráficos de ajuste a una normal, y algún contraste específico para tal fin, como puede ser la *prueba de Kolmogorov-Smirnov*, *prueba de Anderson-Darling*, *prueba de Shapiro-Francia*, *prueba de Shapiro-Wilks*, la *prueba de chi²*, o la prueba de *Cramer-von Mises*.

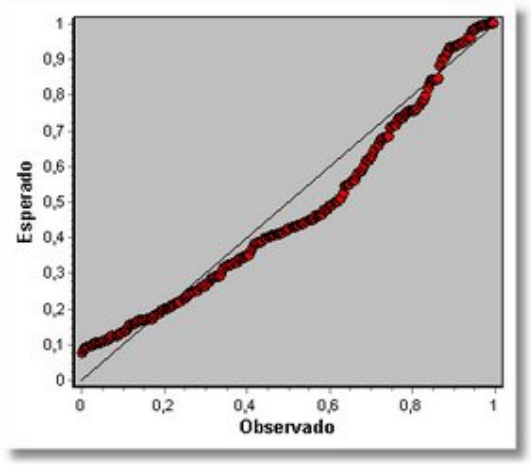



Figura 2. Gráfico de verificación de ajuste a una distribución normal

Una vez que ya se puede suponer que los datos transformados siguen una distribución de probabilidad normal de forma aceptable, se estimarán los percentiles mediante la distribución normal, y después se deshacen las transformaciones, para obtener los límites de referencia en las unidades de los datos.

## Estimación de intervalos de referencia mediante técnicas no paramétricas

Existen alternativas al método paramétrico de estimación de percentiles anteriormente descrito, que no vamos a comentar aquí por su complejidad teórica. Una de ellas se basa en estimar los percentiles aplicando "filtros" a los datos, de tal manera que se aplica mayor peso a los datos próximos al punto del percentil y menor a medida que nos alejamos (*kernel density estimation*).

Otra técnica se basa en simular muestras a partir de los datos, mediante técnicas de muestreo con reemplazamiento, y calcular en ellas los percentiles, de tal manera que la media de todos los percentiles calculados se utilizará como estimación. Estas técnicas, denominadas de *bootstraping*, ya se han comentado someramente en otro artículo.

## Estimación de intervalos de referencia en función de una variable

Una gran cantidad de variables biológicas van evolucionando con la edad, por lo que no es adecuado postular unos límites de referencia globales, sino que éstos deben ser determinados **en función de la edad**. Para resolver este cálculo se han propuesto diferentes procedimientos, en general bastante complejos.

Vamos a presentar aquí uno de los métodos más sencillos, planteado por Wright y Royston, consistente en aplicar transformaciones básicas y técnicas de regresión múltiple para modelar la media, la desviación estándar y la asimetría.

Ahora para cada sujeto tenemos dos datos: la variable estudiada $X$, y la *edad* que vamos a representar por $T$. Es muy posible que antes de nada convenga transformar la variable $X$ tomando logaritmos, para corregir la presencia de asimetría positiva (presencia de una cola alargada hacia el lado derecho, hacia los valores elevados de $X$) o si la dispersión de los datos aumenta con el valor medio (heterocedasticidad). Se supone que disponemos sólo de una observación por sujeto, para una edad determinada, ya que si no fuera así, tendríamos una situación de estudio longitudinal con medidas repetidas para el mismo sujeto, que requiere técnicas especiales de análisis, bastante más complejas.

Se utilizarán técnicas estándar de regresión múltiple para ajustar por el método de mínimos cuadrados ecuaciones polinómicas para la media y la desviación estándar, en función de la edad.

En primer lugar se procederá **modelar la media** $m_T$ en función de la edad. Para ello se comienza intentando ajustar un polinomio de orden tres:

$$a + b \cdot Edad + c \cdot Edad^2 + d \cdot Edad^3$$

Seguidamente comprobamos si con polinomio de orden inferior habría sido suficiente, si el coeficiente $d$ no es significativamente diferente de 0, en cuyo caso se ajusta un polinomio de segundo orden, y repetimos el mismo razonamiento para ver si es suficiente con una recta, o incluso puede ser que el parámetro no varíe con la edad.

Si los ajustes no son buenos y se comprueba que es necesario al menos un polinomio de orden 3, puede ser interesante probar a ajustar un *polinomio fraccional*, en cuyo caso las potencias de la variable *Edad* se escogen del conjunto {-2,-1,-0.5,0,0.5,1,2,3...}, donde 0 corresponde a la función logarítmica. La ventaja de este tipo de polinomio respecto a los polinomios de coeficientes enteros de orden 3 o superior, es que presentan un ajuste con mayor plausibilidad biológica, ya que no tienen las curvaturas artificiosas de los polinomios estándar.

Una vez elegido el modelo para la media, procedemos a **modelar la desviación estándar**. Para ello se calculan los *residuos absolutos escalados*:

$$A = 1{,}25 \cdot |X - m_T|$$

que corresponden a la diferencia en valor absoluto entre cada valor y la media estimada según el modelo para esa edad, multiplicada por 1.25.

Al igual que hicimos con la media, se analiza ahora si los residuos $A$ varían con la edad, y se busca de forma análoga un modelo para esa evolución. La desviación estándar suele requerir funciones menos complejas que la media, y habitualmente es suficiente con ajustar una recta ($a + b.Edad$). Si observamos que la dispersión no depende de la edad, es decir que se mantiene aproximadamente constante al variar ésta, se estimará

entonces a partir de la dispersión residual del modelo de regresión utilizado para modelar la media.

De esta forma tenemos dos ecuaciones: una para estimar la media $m_T$ y otra para la desviación estándar $s_T$ en función de la edad.

Seguidamente habrá que verificar si el modelo se ajusta a una distribución normal. Para ello calculamos los valores estandarizados:

$$Z = \frac{\text{Medida} - m_T}{s_T}$$

Lo primero que podemos hacer es construir un gráfico de ajuste a la normalidad, como el de la figura 2, y aplicar una prueba de bondad de ajuste a una distribución normal.

Otra herramienta gráfica consiste en representar los valores estandarizados en el eje de las Y, en función de la edad en el eje de las X, y éstos se deben distribuir de forma constante, de igual manera a ambos lados del valor cero para todo el rango de edades de la muestra.

Si comprobamos que es aceptable el ajuste a una normal de los valores estandarizados, podemos ya utilizar la fórmula:

$$\text{Percentil}_T = m_T + Z_p . s_T$$

donde $Z$ es el valor correspondiente de la distribución normal (1.96 para un intervalo de referencia de 95%, 1.28 para un intervalo del 90%, etc). Además habrá que tener en cuenta que si los datos se transformaron previamente, por ejemplo con la función logaritmo, ahora habrá que deshacer la transformación, calculando *exp(Percentil)*.

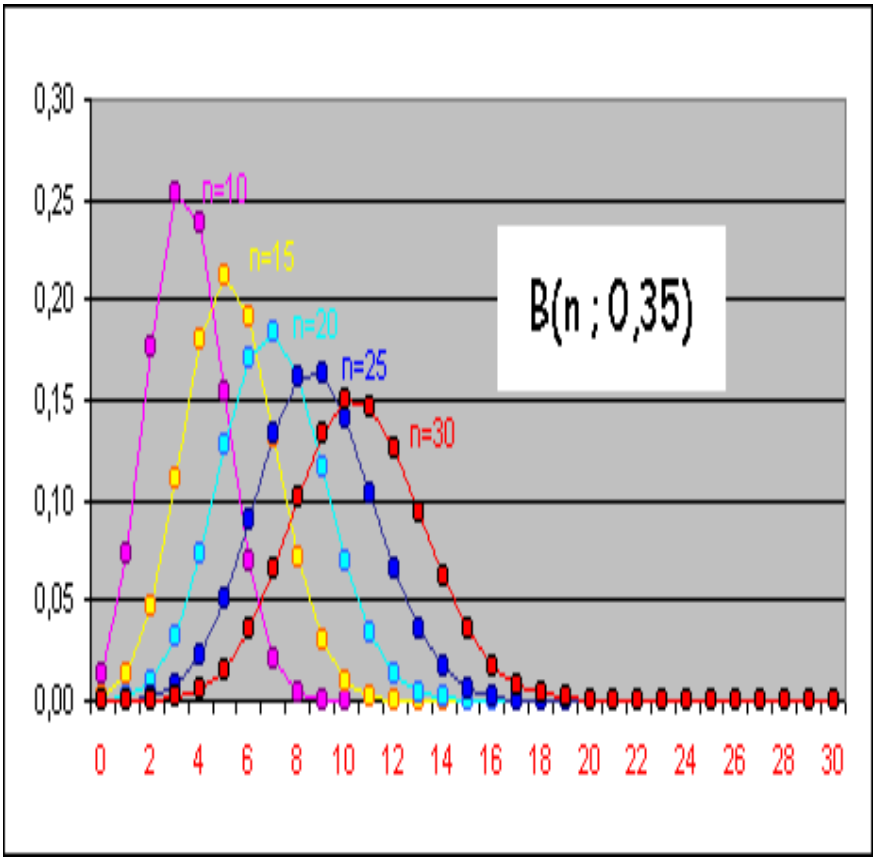
Si no fuera bueno el ajuste a una distribución normal, habrá que tener en cuenta en el modelo también la asimetría, pero para no aumentar la complejidad de esta exposición de momento no vamos a profundizar en ese aspecto, remitiendo al lector interesado al artículo Wright y Royston.

3.-María José García Cebián, Ministerio de Educación. 2001. http://descartes.cnice.mecd.es/Bach_HCS_2/distribuciones_probabilidad/aplic_normal.htm

Veamos a continuación cómo se puede emplear la distribución normal para aproximar una distribución binomial lo que facilita los cálculos en ésta, y por último un ejemplo de ajuste a una normal.



## 1. APROXIMACIÓN DE UNA DISTRIBUCIÓN BINOMIAL POR UNA NORMAL




Una distribución binomial **B(n,p)** se parece a una normal tanto más cuanto mayor es el producto np (o nq si q<p, siendo q=1-p). Cuando np y nq superan 5, la aproximación es casi perfecta, como se puede apreciar en la figura.

En estas condiciones:

**B(n,p)** se aproxima a $\mathbf{N(np, \sqrt{npq})}$

Podemos emplear la normal para calcular probabilidades en el caso de una distribución binomial, aunque hemos de tener en cuenta que la binomial es discreta y la normal continua, por lo que es necesario introducir un ajuste en el cálculo llamado *corrección de Yates*. Así:

$\mathbf{p(X \leq x) = p(X' \leq x+0,5)}$ $\mathbf{p(X < x) = p(X' \leq x-0.5)}$ $\mathbf{p(X = x) = p(x-0,5 \leq X' \leq x+0,5)}$

Veamos un par de ejemplos:

1) El 35% de una población está afectado por la gripe. Se eligen 30 personas al azar.

Se trata de una **B(30;0,35)** que aproximamos por **N(10,5;2,61)**

Calcula la probabilidad de que:

- haya exactamente 10 enfermos

$P(X=10) = P(9,5 \leq X' \leq 10,5)$

- haya más de 5 y menos de 12 enfermos

$P(5<X<12) = P(5,5 \leq X' \leq 11,5)$

*Cambia los valores de a, b y calcula con la tabla las probabilidades correspondientes*

2) Se lanza una moneda 200 veces, calcula la probabilidad de que aparezca cara al menos 100 veces.¿Cuál es la probabilidad de que aparezcan 90 caras?

*Utiliza la escena cambiando los valores de los datos.*

## 2. AJUSTE DE UN CONJUNTO DE DATOS A UNA NORMAL

Con frecuencia conviene saber si puede suponerse que una serie de datos obtenidos experimentalmente proceden de una población distribuida normalmente.

Recordemos que en una distribución normal:

- el 68% de los datos está en el intervalo (x-s, x+s)
- el 95% de los datos está en el intervalo (x-2s, x+2s)
- el 99% de los datos está en el intervalo (x-3s, x+3s)

Si calculadas la media x y la desviación típica s de nuestros datos, se cumplen aproximadamente estos porcentajes podemos considerar que la población de partida es normal.

b) Comparamos la distribución empírica con la normal N((x,s) en este caso con la N(171;7,5).

Veamos un ejemplo en el que seguimos un proceso un poco más elaborado:

## Ejemplo

La tabla adjunta muestra la altura en cm de 100 estudiantes. ¿Es razonable suponer que estos resultados proceden de una distribución normal?

a) Calculamos la media y la desviación típica de la distribución

- Recuerda cómo se calcula la media y la desviación típica.
- Dibujamos el histograma de la distribución.
- Observa que el perfil del histograma recuerda a la curva normal.

| $x_a$ | $x_b$ | $x_i$ | $f_i$ | $x_i f_i$ | $x_i^2 f_i$ |
|---|---|---|---|---|---|
| 155 | 160 | 157,5 | 8 | 1260 | 198450 |
| 160 | 165 | 162,5 | 14 | 2275 | 369687,5 |
| 165 | 170 | 167,5 | 22 | 3685 | 617237,5 |
| 170 | 175 | 172,5 | 28 | 4830 | 833175 |
| 175 | 180 | 177,5 | 16 | 2840 | 504100 |
| 180 | 185 | 182,5 | 8 | 1460 | 266450 |
| 185 | 190 | 187,5 | 4 | 750 | 140625 |
| | | | 100 | 17100 | 2929725 |
| media | | | | | x= 171 |
| desviación típica | | | | | s= 7,50 |