UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE MEDICINA
ESCUELA DE BIOANÁLISIS
CATEDRA DE MATEMATICA Y BIOESTADÏSTICA

GUIA TEORICO - PRÁCTICA DE BIOESTADÍSTICA I

Facultad de Medicina
Bioanálisis
UCV 1949

BIOESTADÍSTICA I

CARÁCTER: Teórico-Práctico
CONDICIÓN: Obligatoria
CRÉDITOS: 4 (3 Teóricos – 1 Práctico)
UBICACIÓN: I Semestre

PERSONAL DE LA CÁTEDRA QUE ELABORÓ LA GUÍA

Prof. María Rosaria Ruggiero
Prof. Yacelli Bustamante
Prof. Claudia Mark
Preparadora: Br. Delimar Recio

INDICE

CAPÍTULO I

¿Qué es la Estadística?

La palabra Estadística proviene del latín “status”. En la antigüedad chinos, egipcios, hebreos, griegos y romanos la practicaban en recuentos de población y riquezas. Con el tiempo se perfeccionó mediante métodos matemáticos y probabilísticos hasta generalizar su estudio y uso a cualquier actividad científica.

La Estadística es la ciencia o conjunto de métodos científicos que tienen por objeto la recolección, agrupación, presentación, análisis e interpretación de los datos obtenidos de una población o muestra, como medio para hacer estimaciones e inferencias para la toma de decisión ante diversas alternativas.

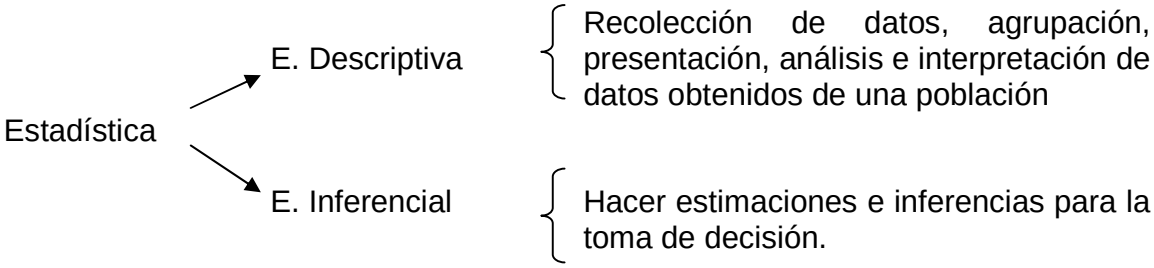
Concepto e importancia de la Bioestadística

La Bioestadística constituye el empleo de métodos estadísticos en la investigación de los hechos biológicos.

La comprensión de la Estadística aumentará la capacidad del profesional de la salud para interpretar datos, sea con el propósito de tratar a un paciente en particular o para obtener conclusiones generales de una investigación.

Partes de la Estadística

La Estadística se divide en dos partes:



CAPÍTULO II

Estadística Descriptiva

Como se expresó anteriormente, esta parte de la Estadística se caracteriza por la recolección de datos, agrupación, presentación, análisis e interpretación de datos obtenidos de una población o muestra. Estudiaremos los siguientes aspectos:

Métodos Estadísticos.
Medidas de Tendencia Central.
Medidas de Posición.
Medidas de Dispersión.
Medidas de Forma.

Analicemos cada uno de estos aspectos.

Métodos estadísticos

Está constituido por los siguientes puntos:

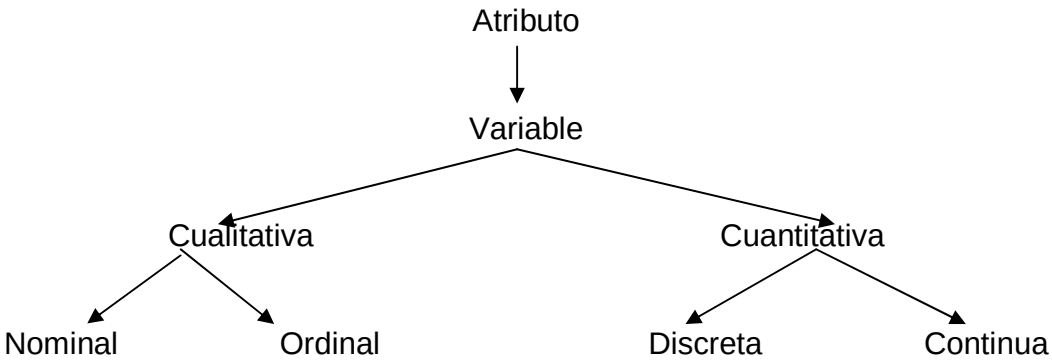
La fuente de datos

Para estudiar un determinado comportamiento o características existentes de un conjunto de elementos (datos) que integran una población (conjunto de individuos, objetos o acontecimientos definidos con relación a algún rasgo en común que los identifique). Puede considerarse un censo, en el que se investigan todos y cada uno de los elementos de la población o bien una muestra en el que se investiga un subconjunto de la población y se escogen al azar de modo tal que ellos sean representativos de la población. Estadísticamente hablando, el tamaño de la población se denota por *N* y el tamaño de la muestra por *n*.

Observación: En el momento de un estudio, es importante conocer si los datos provienen de una población o una muestra, permitiendo así determinar el lineamiento del estudio a realizar:

- Población ⟶ Estadística descriptiva
- Muestra ⟶ Estadística descriptiva ⟶ Estadística Inferencial

Características a las cuales se refieren los datos



El atributo constituye la característica a la que se refieren los datos. Como los atributos varían de miembro a miembro, se denominan variables.

Las variables son los símbolos que adquieren diferentes valores de una misma situación. Si en algún caso la variable toma el mismo valor en una situación, entonces su denominación es constante.

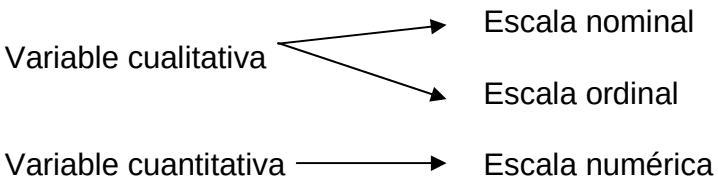
Según el comportamiento de la variable, se clasifican en:

- **Variable Cualitativa:** Son aquellas que únicamente se pueden describir, como por ejemplo el color, sabor, tipo de medicina, entre otras. Estas variables tienen dos tipos de escalas: una nominal en el que no se tiene un orden preestablecido (color del cabello) y otra ordinal en el que se tiene un orden preestablecido (clases sociales)

- **Variable Cuantitativa:** Son aquellas que se pueden contar y medir. Estas variables pueden ser discretas, que se caracterizan por oscilar únicamente entre valores enteros (número de hijos) y las continuas, que se caracterizan por ser susceptibles a subdividirse indefinidamente y pueden tomar cualquier valor (peso)

Formas de medición

Depende de la naturaleza y comportamiento de la variable.



Variable cuantitativa con escala numérica: Al asignar un número indica la propiedad del atributo que miden, además de poderse establecer diferencias entre ellas. Ejemplo: El peso.

Variable cualitativa con escala nominal: Clasifican la variable en categorías descriptivas mutuamente excluyentes y colectivamente exhaustivas, es decir, se les puede asignar un número para expresar clases diferentes y no para establecer relaciones de mayor a menor. Ejemplo: La moneda.

Variable cualitativa con escala ordinal: Se clasifican de una forma de interpretación jerárquica. Ejemplo: Clase social.

Formas de recolección de datos

Los datos se pueden recolectar mediante encuestas, cuestionarios, entrevistas, bibliografías, historias clínicas, entre otras.

Formas de representación de los datos

Se pueden presentar de dos formas, como una distribución de frecuencias y/o de forma gráfica.

Distribución de frecuencias

La organización tabular que contiene todas las variantes o clases de la variable y sus frecuencias respectivas, es llamada distribución de frecuencias, y constituye la forma práctica y clara de presentar la información numérica obtenida de una investigación.

Título

Encabezado	
Columna Matriz	Cuerpo
Total	

Fuente de datos

- Título: Coloque título de manera breve, concisa, completa (qué se estudia, como se estudia, dónde, cuándo)

- Encabezado: Indica a qué se refieren los datos y el contenido de cada columna.

- Columna matriz: Se asienta las diferentes escalas de clasificación usada.

- Cuerpo: Contiene las frecuencias y distintos valores a los que se refieren los datos.

- Total: Se indican los totales de las columnas.

- Fuente de datos: Indique fuente de los datos de manera precisa y completa.

Dentro de la distribución de frecuencias, se pueden visualizar las siguientes columnas:

- Frecuencia Absoluta: Corresponde al número de datos que caen en cada uno de los intervalos.

- Frecuencia Relativa: Corresponde al peso de la frecuencia absoluta de cada intervalo con respecto al total.

- Frecuencia Acumulada: Corresponde al número de elementos contenidos en la distribución a nivel de cada clase, o bien frecuencia absoluta acumulada hasta cada clase.

La distribución de frecuencias se puede presentar de dos formas:

- Datos sin agrupar: Cada uno de los datos aparece con sus frecuencias.

- Datos agrupados: Cuando el número de variantes o clases de la variable es muy grande, es preferible incluir en cada clase varias mediciones de la variable en vez de una sola.

La terminología utilizada en este tipo de agrupación es la siguiente:

- Clases: constituye cada grupo de variantes.

- Intervalos de clases: es el rango de los valores que determinan una clase. Se obtiene restando el límite superior del límite inferior de cada clase.

- Límites de una clase*:* son los valores inferiores y superiores que definen a cada clase. El inferior se llama límite inferior y el superior se llama límite superior. Los límites pueden darse de manera aparente o real:

    - *Límite aparente (L_{ai} - L_{as}):* El límite superior de la primera clase no coincide con el límite inferior de la segunda clase, y así sucesivamente. Si la variable es continua, se puede perder algún dato, por lo que buscamos los límites reales.
    - *Límite real (L_{ri} – L_{rs}):* Es el valor que dos clases contiguas comparten. Se obtiene de esta forma:

$$L_{S_1} = \frac{1}{2}\left(L_{S_1} + L_{I_2} \right)$$

Ejemplo:

Límites aparentes	Límites reales
60 – 62	[59.5 - 62.5)
63 – 65	[62.5 – 65.5)

Hay que tener cuidado de que los límites reales no coincidan con los valores observables, para evitar ambigüedades sobre la clase a la que corresponde una observación.

- Marca de clase: Es el valor central de cada grupo, se obtiene al sumar el límite superior con el límite inferior de la clase, y luego dividirlo entre dos, es decir:

$$X_i = \frac{L_i + L_S}{2}$$

❖ Ejemplo de construcción de una Distribución de Frecuencia para datos sin agrupar

Se tomó una muestra de 20 estudiantes de la Escuela de Bioanálisis para evaluar sus latidos del corazón (pul/min) después de acondicionamiento físico. Se obtuvieron los siguientes valores:

86 82 85 92 88 90 85 95 90 86
90 92 88 92 90 92 88 90 90 96

Se pide, construir la distribución de frecuencias de datos sin agrupar.

Distribución de Frecuencia correspondiente a latidos del corazón de 20 estudiantes de la Escuela de Bioanálisis después de acondicionamiento físico

X_i (pul/min)	f_i	$\%f_i$	f_a	$\%f_a$
82	1	5	1	5
85	1	5	2	10
86	2	10	4	20
88	3	15	7	35
90	6	30	13	65
92	4	21	17	85
95	2	10	19	95
96	1	5	20	100
Total	20	100		

Fuente de datos: Prof Mateo Rodríguez, Acondicionamiento Físico. Universidad Central de Venezuela. Caracas, 2000

- <u>Reglas generales para agrupación de datos.</u>

  - Reste el mayor de los datos con el menor de ellos. A esto se le llama rango y se denota por R. Esto indica los límites dentro de los cuales se presentan todos los datos considerados.

  - Calculamos el número de intervalos de clase. Hay dos formas; que lo den en el enunciado (lo conveniente es que oscile entre 5 y 20) o bien podemos calcularlo por medio de la Regla de Sturges:

    Número de Intervalos= 1+3.32log(n)

Observación: No redondee este número para calcular amplitud.

- Calculamos la amplitud:

$$A = \frac{R}{Número\,de\ \text{int}ervalos}$$

Este valor corresponde a la amplitud aparente. Para encontrar la amplitud real, dicho valor se redondeará por exceso de acuerdo a la unidad de la variable.

- *Ventajas del uso de límites reales.*

  1.- No se rompe la continuidad.
  2.- No existe la posibilidad de que un valor caiga en la frontera.
  3.- No se altera la marca de clase.

- *Desventajas del uso de límites reales*

  1.- Da impresión de continuidad.
  2.- Se trabaja con decimales.
  3.- No recomendable para variables discretas.

❖ Ejemplo de construcción de una Distribución de Frecuencia para datos agrupados

En el Laboratorio del Hospital Clínico Universitario, se escogió una muestra de 25 personas para analizar sus niveles de glicemia (mg/dl) y estos fueron los resultados:

75	*82*	*90*	*95*	*101*	*112*	*121*	*132*	*140*
97	*84*	*90*	*96*	*102*	*114*	*121*	*138*	*87*
91	*96*	*104*	*123*	*89*	*93*	*99*		

Se pide construir la distribución de frecuencias de datos agrupados.

En principio debemos calcular el rango:

$$Rango = 140 - 75 = 65\,mg/dl$$

Ahora calculamos el número de intervalos por la regla de Sturges:

$$Num.de\,Intervalos = 1 + 3.32\log(25) = 5.6411608$$

Finalmente, la amplitud es:

$$A = Amplitud = \frac{65}{5.6411608} = 11.52$$

Como los datos son enteros, entonces la amplitud aparente (A_a) es 11 y la amplitud real (A_r) es 12.

Distribución de Frecuencia de los Niveles de Glicemia de 25 personas del Hospital Clínico Universitario

L_{ai} - L_{as} (mg/dl)	L_{ri} – L_{rs} (mg/dl)	f_i	$\%f_i$	f_a	$\%f_a$	X_i
75 – 86	74.5 – 86.5	3	12	3	12	80.5
87 – 98	86.5 – 98.5	10	40	16	52	92.5
99 – 110	98.5 – 110.5	4	16	17	66	104.5
111 – 122	110.5 – 122.5	4	16	21	84	116.5
123 – 134	122.5 – 134-5	2	8	23	92	128.5
135 – 146	134.5 – 146.5	2	8	25	100	140.5
Total		25	100			

Fuente de datos: Laboratorio del Hospital Clínico Universitario. Caracas, 2000

Gráficas

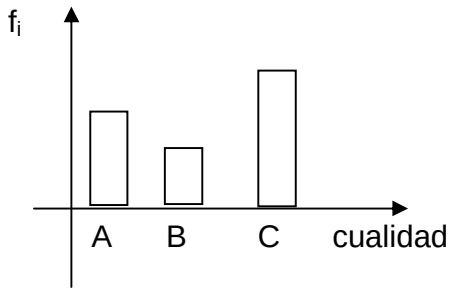
Representación de datos numéricos por medio de coordenadas o dibujos que hacen visible la relación o gradación que esos datos guardan entre sí. Permiten visualizar mejor las variaciones de las variables. Dividiremos los tipos de gráficos dependiendo de la forma como estén dados los datos.

Datos Cualitativos

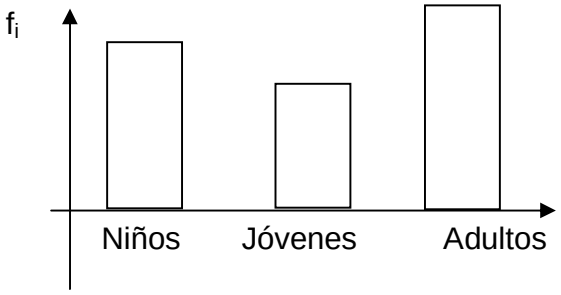
- Histograma de frecuencias: Se compone de barras rectangulares levantadas sobre el eje horizontal, en el cual se marcan, utilizando escalas adecuadas, los valores que asume la variable en la distribución de frecuencias. Si los datos son *nominales* se ordena por orden alfabético, en

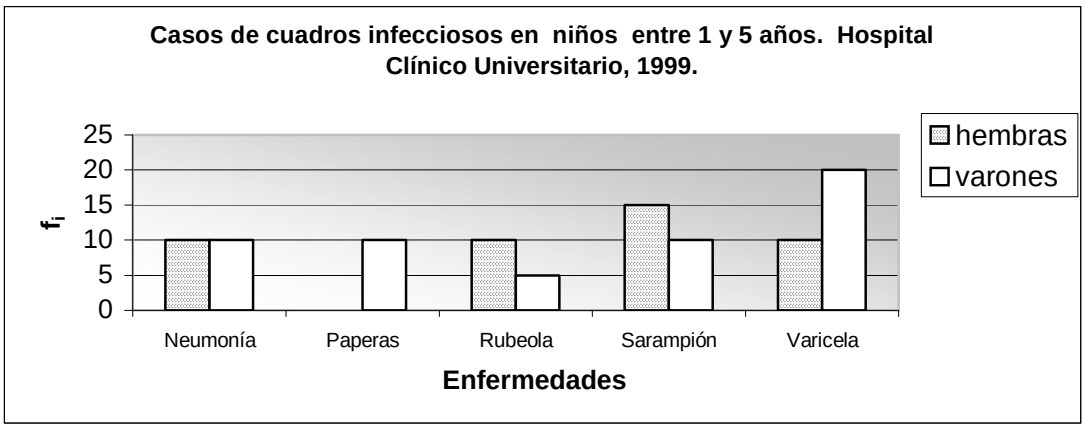
cambio que si los datos son *ordinales*, se colocará en orden jerárquico. Ejemplos:

Datos nominales

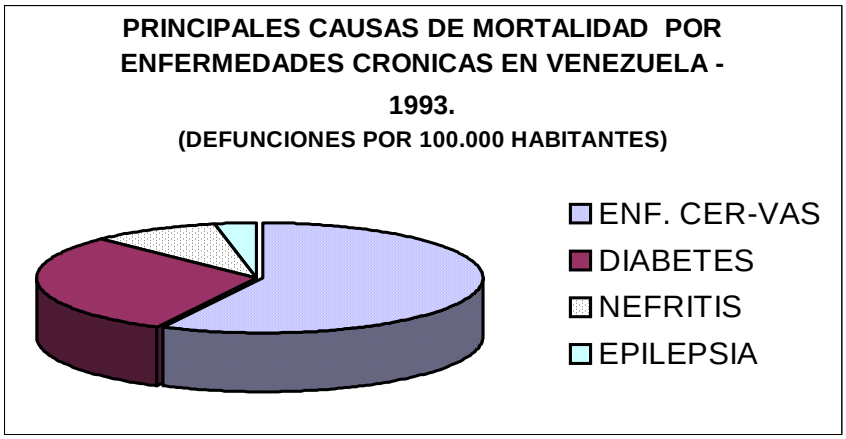


Datos ordinales





- Diagrama circular: Se divide el área de un círculo proporcionalmente a las frecuencias relativas de cada clase. Es necesario colocar la leyenda de datos. Ejemplo:



- Pictogramas: Se busca algún símbolo que represente la frecuencia absoluta de cada clase.

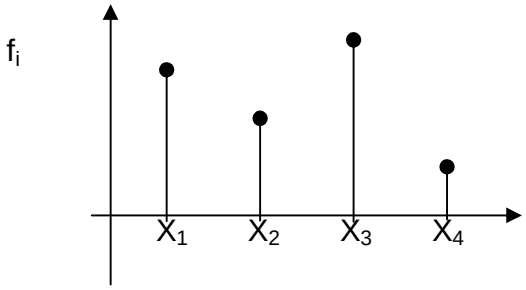
 Equivale a 10.000 habitantes por km^2

 Equivale a 30.000 habitantes por km^2

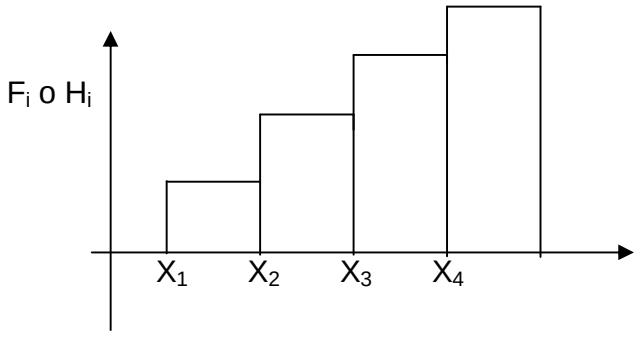
Datos Cuantitativos

❖ **Datos sin agrupar**

- Histograma de frecuencia (absolutas o relativas): Es este caso, como los datos son puntuales, solo se representa una línea.

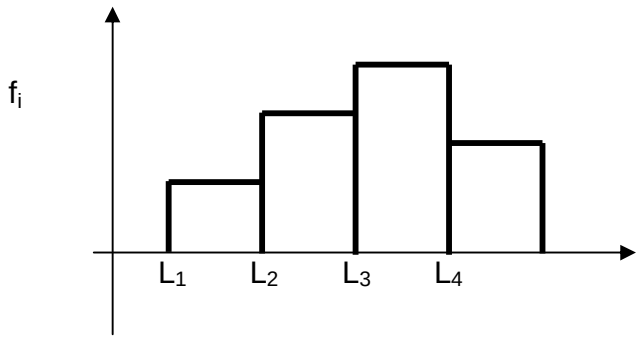


- Ojiva (frecuencia acumulada absoluta o relativa): Es una forma de escalera ya que entre un dato y dato no hay valores intermedios.

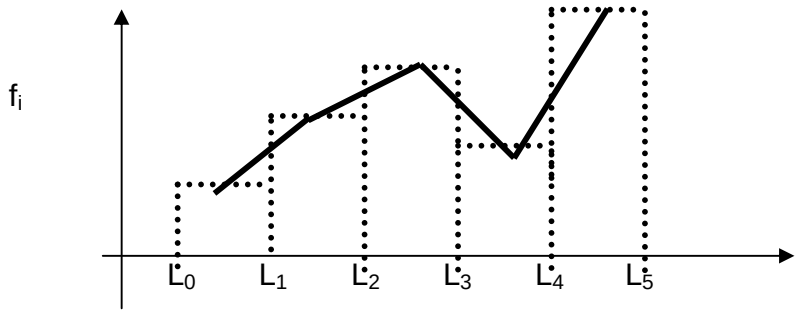


❖ **Datos agrupados**

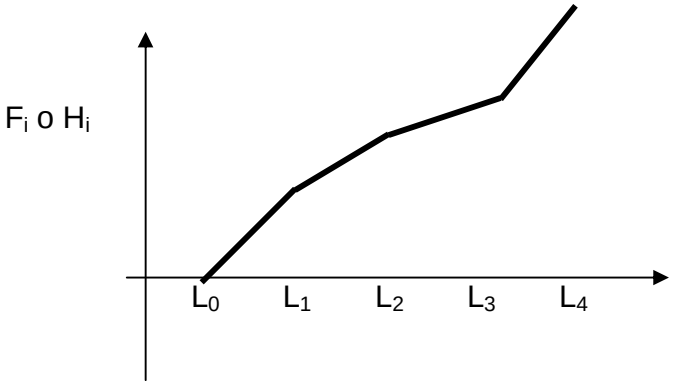
- Histograma de frecuencia: Similar al de los datos cualitativos solo que en el eje horizontal se coloca los límites reales, el ancho de los rectángulos es igual a la amplitud del intervalo de clase y el punto medio representa la marca de clase.



- Polígono de frecuencia: Es la unión de los puntos medios de las barras (marcas de clase). Se usan para compara dos distribuciones en una misma gráfica



- Ojiva: Es la línea quebrada que se traza por los puntos de intersección de las coordenadas que corresponden a los límites reales de cada clase y sus respectivas frecuencias acumuladas.



Medidas de Tendencia Central

Son las medidas que analizan el comportamiento de los datos en sus valores centrales y son representativos en todas sus variantes. En este curso solo estudiaremos la media aritmética, la mediana y la moda.

Media Aritmética

La Media Aritmética es el promedio aritmético en una distribución de datos. Es el más usado de los promedios, siempre y cuando la serie no presente valores extremos, ya que esto distorsiona el valor de la media, en este caso sería aconsejable otra medida (la mediana por ejemplo). Es el valor típico representativo de un conjunto de datos y se caracteriza por depender de todas las medidas que forman la serie de datos.

- *Datos sin agrupar*

Viene representado por la sumatoria de todos los valores de la variable dividida entre el número de datos:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

donde x_i corresponde al valor de la variable en el individuo i, para $i = 1,2,...,n$ y n es el número de datos.

❖ *Ejemplo: Se desea calcular la media de los números 2, 5, 6, 8 y 4.*

$$\overline{X} = \frac{2+5+6+8+4}{5} = \frac{25}{5} = 5$$

La media de los números es 5.

- *Datos agrupados*

En el caso de que los datos esten agrupados, la media aritmética viene expresada por la fórmula

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

donde f_i es la frecuencia de la clase i y x_i es la marca de la clase i.

- *Ejemplo: La siguiente tabla se refiere a la estatura de 50 estudiantes (en metros) de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela.*

Estatura de 50 estudiantes de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela

$L_{ri} - L_{rs}$ (metros)	f_i	X_i	$f_i X_i$
1.45 – 1.48	2	1.465	2.930
1.48 – 1.51	7	1495	10.468
1.51 – 1.54	4	1.525	6.100
1.54 – 1.57	3	1.555	4.665
1.57 – 1.60	12	1.585	19.02
1.60 – 1.63	9	1.615	14.535
1.63 – 1.66	4	1.645	6.58
1.66 – 1.69	4	1.675	6.700
1.69 – 1.72	2	1.705	3.410
1.72 – 1.75	3	1.735	5.205
Total	50		79.61

Así:

$$\bar{x} = \frac{79.61}{50} = 1.59\, m$$

La estatura media de los 50 estudiantes de la Escuela de Nutrición y Dietética es 1.59 m.

Propiedades de la Media Aritmética

- Si a cada uno de los datos le sumamos una cantidad constante, la nueva media aritmética es igual a la anterior sumada por esa misma constante, es decir, sean los datos ($x_1 + k$, $x_2 + k$, $x_3 + k$,…, $x_n + k$) entonces

$$\overline{X}' = \overline{X} + k$$

- Si a cada uno de los datos se le multiplica por una misma constante, la nueva media aritmética es igual a la anterior multiplicada por la misma constante, es decir, sean los datos ($cx_1, cx_2, cx_3, \ldots, cx_n$) entonces

$$\overline{X}' = c\,\overline{X}$$

- Si se tienen varias muestras, entonces

$$\overline{X} = \frac{n_1\,\overline{x_1} + n_2\,\overline{x_2} + n_3\,\overline{x_3} + \cdots + n_k\,\overline{x_k}}{n_1 + n_2 + n_3 + \cdots + n_k}$$

Ventajas del uso de la Media Aritmética

- Fácil de entender y calcular.
- Hace uso de todos los datos de una distribución.
- Es el más conocido.
- Es usada en la inferencia estadística.
- Se presta a manipulación algebraica.

Desventajas del uso de la Media Aritmética

- Puede ser influenciada por los valores extremos que hagan perder su medida central.
- En el caso de variables discretas, el valor no es exactamente alguno de los datos, por lo que a veces se redondean.
- La media aritmética no puede ser calculada para una distribución con intervalos de clases abiertas, esto es, cuando los elementos están agrupados en intervalos de clase de tipo “por encima de” o “por debajo de”.

La Mediana

La Mediana es el valor de la variable que equidista de ambos extremos de la distribución cuando está ordenada de manera creciente, es decir, es el valor que deja por debajo de él el 50% de los datos, consecuentemente por encima de la mediana se halla el 50% de los datos. El valor de la mediana puede coincidir o no con un valor de la serie de datos.

- *Datos sin agrupar*

Tendremos dos casos dependiendo del valor de la población. Si dicho tamaño es:

- Impar: La medida coincidirá con la variante central, que se obtiene sumando 1 a la frecuencia total y dividiendo el resultado por dos.

$$Md = X_{\frac{n+1}{2}}$$

- Par: La mediana está representada por la media aritmética de las dos variantes centrales

$$Md = \frac{X_{n/2} + X_{(n+2)/2}}{2}$$

❖ *Ejemplo: Calcule la mediana de 2,8,5,3,4,6,2*

Primero se ordenan los datos en forma creciente → 2,2,3,4,5,6,8

En este caso n = 7, por lo que la mediana esta en la posición 4, así la mediana es 4

❖ *Ejemplo: Calcule la mediana de 1,8,10,3,4,2,3,5*

Primero se ordenan los datos en forma creciente→ 1,2,3,3,4,5,8,10

En este caso n = 8, por lo que la mediana está en la posición 4.5, así la mediana es 3.5

- *Datos agrupados*

En este caso, la mediana se obtendrá mediante la fórmula siguiente:

$$Md = L_i + \frac{\left(n/2\right) - F_{(ant)}}{f_i}.A$$

donde:

$n/2$ = Nos da la posición aproximada de la mediana en la distribución, de acuerdo al número de datos que se disponga.

L_i = límite inferior real de la clase medianal

$F_{(ant)}$ = frecuencia acumulada de la clase anterior a la clase medianal.

f_i = frecuencia absoluta de la clase medianal.

A = amplitud real del intervalo.

❖ *Ejemplo: Siguiendo con los datos de las estaturas de 50 estudiantes de la Escuela de Nutrición y Dietética tenemos:*

Estatura de 50 estudiantes de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela

L_{ri} – L_{rs} (metros)	**f_i**	**X_i**	**f_a**
1.45 – 1.48	2	1.465	2
1.48 – 1.51	7	1495	9
1.51 – 1.54	4	1.525	13
1.54 – 1.57	3	1.555	16
1.57 – 1.60	**12**	**1.585**	**28**
1.60 – 1.63	9	1.615	37
1.63 – 1.66	4	1.645	41
1.66 – 1.69	4	1.675	45
1.69 – 1.72	2	1.705	47
1.72 – 1.75	3	1.735	50
Total	50		

Hallamos la clase medianal. Tenemos que $n/2 = 50/2 = 25$.

La clase medianal estará en el intervalo donde se tenga por lo menos una frecuencia acumulada de 25 datos.

Así la clase medianal es (1.57, 1.60). Por lo que:

$$Md = 1.57 + \frac{25-16}{12} \times 0.03 = 1.59m$$

El valor de la mediana de las 50 estaturas de los estudiantes de la Escuela de Nutrición y Dietética es 1.59m.

Propiedades de la Mediana

- No es un estadígrafo suficiente, ya que no considera a todos los datos.

Ventajas del uso de la Mediana

- Los valores extremos no la afectan ya que está determinada por el número de observaciones y no por el valor de las mismas.
- Se puede calcular aunque los valores extremos sean abiertos.
- Es fácil de calcular.

Desventajas del uso de la Mediana

- No se presta a tratamientos algebraicos.
- Es necesario ordenar las variantes antes de que se pueda calcular la mediana.
- Es poco conocida.

La Moda

La Moda se define como el valor que tiene más frecuencia en una serie de datos. Puede que no exista o bien que existan varios valores candidatos a ser moda.

- *Datos sin agrupar.*

En una distribución de datos estadísticos, es el valor que más se repite.

❖ *Ejemplo: Encuentre el valor de la moda en el siguiente conjunto de datos 8,4,1,2,4,3,7,5,4,2,3,8* $\rightarrow$ La moda es 4.

- *Datos agrupados*

Si los datos están agrupados en distribuciones de frecuencias, la moda sería el valor de frecuencia más alta (clase modal). En este caso, la moda se calcula mediante la siguiente fórmula:

$$Mo = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}.A$$

L_i = límite inferior real de la clase modal

Δ_1 = Diferencia absoluta entre la frecuencia de la clase modal y la de la clase anterior.

Δ_2 = Diferencia absoluta entre la frecuencia de la clase modal y la de la clase posterior.

A = amplitud real del intervalo.

❖ *Ejemplo: Siguiendo con los datos de las estaturas de 50 estudiantes de la Escuela de Nutrición y Dietética tenemos:*

Estatura de 50 estudiantes de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela

$L_{ri} - L_{rs}$ *(metros)*	f_i	X_i	f_a
1.45 – 1.48	2	1.465	2
1.48 – 1.51	7	1495	9
1.51 – 1.54	4	1.525	13
1.54 – 1.57	3	1.555	16
1.57 – 1.60	**12**	**1.585**	**28**
1.60 – 1.63	9	1.615	37
1.63 – 1.66	4	1.645	41
1.66 – 1.69	4	1.675	45
1.69 – 1.72	2	1.705	47
1.72 – 1.75	3	1.735	50
Total	50		

La clase modal es (1.57-1.60). Por lo que:

$$Mo = 1.57 + \frac{|12-3|}{|12-3|+|12-9|} \times 0.03 = 1.59m$$

La estatura que más se repite es 1.59m.

Propiedades de la Moda

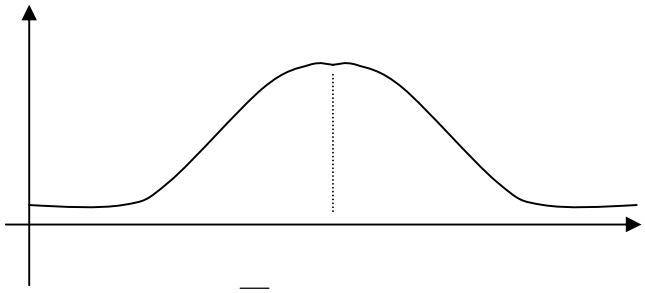
- Una distribución puede tener un solo valor modal, en este caso decimos que la distribución es *unimodal*. Si dos variantes se repiten con la misma frecuencia decimos que la distribución es *bimodal.* Si hay más de dos variantes con la misma frecuencia, la distribución es llamada *multimodal.*
- Si todos los datos tienen la misma frecuencia, no existe moda.
- La moda corresponde al valor donde el histograma alcanza la máxima altura.
- No es un estadígrafo suficiente, ya que no toma en cuenta todos los datos y si algunos datos se alteran, es posible que la moda siga igual.
- Carece de significación en distribuciones que contengas pocos datos y no ofrezcan una marcada tendencia central.
- No es afectada por los valores extremos.

Relación de las Medidas de Tendencia Central

Se cumple la siguiente relación empírica:

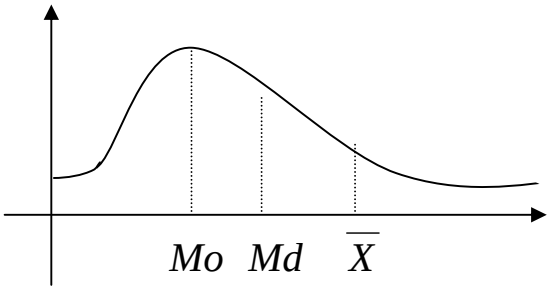
$$Mo = \overline{X} - 3\left(\overline{X} - Md\right)$$

Dependiendo de la forma de la distribución, puede ocurrir que si la distribución es simétrica, por ejemplo la distribución normal, la media, la mediana y la moda coinciden.

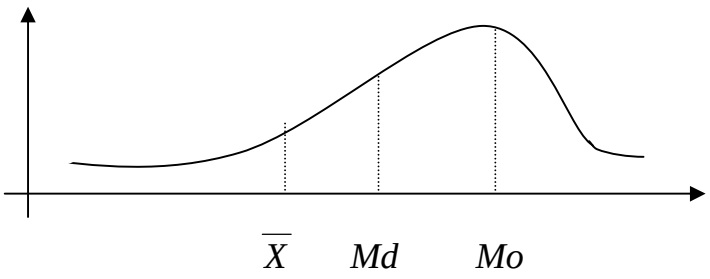


$$\overline{X} = Mo = Md$$

Si la distribución es asimétrica, puede ocurrir que $Mo < Md < \overline{X}$



o bien que $\overline{X} < Md < Mo$



¿Cómo cree Ud. que son los datos?

Medidas de Posición

Las medidas de posición son medidas estadísticas que dividen la distribución de los datos en partes iguales y describen la posición que tiene un dato dentro de una distribución, una vez que se ordena de forma creciente. Estudiaremos los Percentiles, Deciles y Cuartiles.

Percentiles

Son valores que dividen la distribución en 100 partes iguales y nos dan la situación de los datos según el lugar que ocupan en tanto por ciento. Son 99 y se denotan por $P_1, P_2, ..., P_{99}$. Así P_k corresponde al aquel valor que supera al $k\%$ de datos a lo más y a la vez es superado por el $(100-k)\%$ de los datos a lo máximo.

- *Datos sin agrupar*

Se ordenan los datos de forma creciente. Seguidamente calculamos

$$F_k = \frac{k.n}{100}$$

para determinar la posición del percentil k. Para hallar P_k buscamos en la columna de frecuencia acumulada, en qué elemento se ubican por lo menos F_k datos.

- *Datos agrupados.*

El percentil será hallado mediante la siguiente fórmula:

$$P_k = L_i + \frac{k.n/100 - F_{(ant)}}{f_i} \times A$$

donde:

$k.n/100$ = indica la posición donde está ubicado el percentil.

L_i = límite inferior real de la clase donde esta ubicado el percentil

$F_{(ant)}$ = frecuencia acumulada de la clase anterior de donde está ubicado el percentil.

f_i = frecuencia absoluta de la clase donde está ubicado el percentil.

A = amplitud real del intervalo.

Deciles

Son valores que dividen la distribución en 10 partes iguales, son 9 y se denotan por $D_1, D_2, ..., D_9$. Así D_2 por ejemplo, corresponde a aquel valor que supera al 20% de datos a lo más y a la vez es superado por el 80% de los datos a lo máximo.

- *Datos sin agrupar*

Se ordenan los datos de forma creciente. Seguidamente calculamos

$$F_k = \frac{k \,.n}{10}$$

para determinar la posición del decil k. Para hallar D_k buscamos en la columna de frecuencia acumulada, en qué elemento se ubican por lo menos F_k datos.

- *Datos agrupados*

El decil será hallado mediante la siguiente fórmula:

$$D_k = L_i + \frac{k.n/10 - F_{(ant)}}{f_i} \times A$$

donde:

$k.n/10$ = indica la posición donde está ubicado el decil.

L_i = límite inferior real de la clase donde esta ubicado el decil

$F_{(ant)}$ = frecuencia acumulada de la clase anterior de donde está ubicado el decil.

f_i = frecuencia absoluta de la clase donde está ubicado el decil.

A = amplitud real del intervalo.

Cuartiles

Son valores que dividen la distribución en 4 partes iguales, son 3 y se denotan por Q_1, Q_2, Q_3. Así Q_1 por ejemplo, corresponde a aquel valor

que deja por debajo de él, el 25% de datos y a la vez deja por encima el 75% de los datos.

- *Datos sin agrupar*

  Se ordenan los datos de forma creciente. Seguidamente calculamos

$$F_k = \frac{k.n}{4}$$

  para determinar la posición del cuartil k. Para hallar Q_k buscamos en la columna de frecuencia acumulada, en qué elemento se ubican por lo menos F_k datos.

- *Datos agrupados*

  El cuartil será hallado mediante la siguiente fórmula:

$$Q_k = L_i + \frac{k.n/4 - F_{(ant)}}{f_i} \times A$$

  donde:

  $k.n/4$ = indica la posición donde está ubicado el cuartil.

  L_i = límite inferior real de la clase donde esta ubicado el cuartil

  $F_{(ant)}$ = frecuencia acumulada de la clase anterior de donde está ubicado el cuartil.

  f_i = frecuencia absoluta de la clase donde está ubicado el cuartil.

  A = amplitud real del intervalo.

  <u>Propiedades.</u>

  Se cumple que:

$$Q_1 = P_{25} \qquad Q_3 = P_{75}$$
$$Q_2 = Me = P_{50} \qquad P_{10} = D_1$$
$$P_{20} = D_2 \qquad P_{90} = D_9$$

- *Ejemplo: Los siguientes datos corresponden a los sueldos semanales (en miles de bolívares) de 80 bioanalistas del Laboratorio X.*

Sueldo semanal (en miles de bolívares) de 80 bioanalistas del Laboratorio X

X_i (Bs.)	f_i	f_a
100	6	6
105	10	16
115	25	41
120	18	59
123	12	71
135	7	78
220	2	80

o *Calcule* P_{60}

La posición de este percentil es $F_{60} = \frac{60\,x\,80}{100} = 48$.

De esta forma $P_{60} = 120\ Bs.$

o *Calcule* Q_3.

La posición de este cuartil es $F_3 = \frac{3x80}{4} = 60$.

De esta forma $Q_3 = 123\ Bs.$

Calcule D_9.

La posición de este decil es $D_9 = \frac{9\,x\,80}{10} = 72$.

De esta forma $D_9 = 135\ Bs.$

- *Ejemplo: Siguiendo con los datos de las estaturas de 50 estudiantes de la Escuela de Nutrición y Dietética tenemos:*

Estatura de 50 estudiantes de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela

$L_{ri} - L_{rs}$ (metros)	f_i	f_a
1.45 – 1.48	2	2
1.48 – 1.51	7	9
1.51 – 1.54	4	13
1.54 – 1.57	3	16
1.57 – 1.60	12	28
1.60 – 1.63	9	37
1.63 – 1.66	4	41
1.66 – 1.69	4	45
1.69 – 1.72	2	47
1.72 – 1.75	3	50
Total	50	

o *Calcular* P_{66}

En principio calculamos la posición del percentil, $F_{66} = \frac{66\,x\,50}{100} = 33$

Así tenemos que

$$P_{66} = 1.60 + \frac{33-2}{9} x 0.03 = 1.62m$$

o *Calcular* Q_1

En principio calculamos la posición del cuartil, $F_1 = \frac{1x50}{4} = 12.5$

Así tenemos que

$$Q_1 = 1.51 + \frac{12.5 - 9}{4} x 0.03 = 1.54m$$

- *Calcular* D_3

En principio calculamos la posición del decil, $D_3 = \frac{3x50}{10} = 15$

Así tenemos que

$$D_3 = 1.54 + \frac{15 - 13}{3} x 0.03 = 1.56m$$

Medidas de Dispersión

Las Medidas de Tendencia Central o de Localización dan una visión del grupo, pero la misma es incompleta. Ellas dan información acerca del centro de los datos pero no qué tan dispersos son los mismos.

Para complementar las medidas de tendencia central se usan las medidas de variabilidad, ellas miden la dispersión de los datos alrededor de la medida de localización usada.

Las medidas de variabilidad indican qué tan diseminados son los datos del grupo al cual se le calcula la medida. Si un grupo tiene una baja variabilidad esto indica que está compuesto por individuos aproximadamente iguales, los datos están poco esparcidos, están bastante agrupados. La mayoría de los puntajes estarán alrededor de la medida de tendencia utilizada. En este caso se dice que los individuos poseen características homogéneas.

Pero si la variabilidad es alta, los puntajes estarán dispersos, los individuos u objetos que conforman el grupo serán disímiles. En este caso se dice que los individuos poseen características heterogéneas.

Desviación Típica

La Desviación Típica es una medida que da una mejor idea de cómo los datos se dispersan de la media. La Desviación Típica mide cómo los datos difieren de la Media Aritmética.

- *Datos sin agrupar*

Si los datos son **simples** (sin frecuencia) usaremos la fórmula:

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

Si los datos están **repetidos** (con frecuencias), usaremos la fórmula:

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2 \; f_i}{n}}$$

en donde x_i corresponde al valor de la característica, $\overline{X}$ la media de los datos y f_i la frecuencia de la característica i.

En caso tal que $n < 30$, usaremos la fórmula

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

Este estadístico se usa cuando se desea estimar la variabilidad de un conjunto de datos. Dicha corrección del denominador cuando n<30, utiliza también para datos repetidos o agrupados.

- *Datos agrupados*

La fórmula a utilizar es:

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

en donde x_i corresponde a la marca de clase del intervalo i, $\overline{X}$ la media de los datos y f_i la frecuencia del intervalo i.

<u>Características de la Desviación Típica</u>

- Proporciona la variación de datos respecto a la media aritmética.
- Su valor se encuentra en relación directa con la dispersión de los datos, a mayor dispersión de ellos, mayor desviación típica; a menor dispersión, menor desviación típica.
- Es la medida de dispersión adecuada cuando la medida de tendencia central es la media.
- Es susceptible de los valores extremos.
- La mayor utilidad de la desviación típica se presenta en una distribución normal, al encontrar que en los intervalos:
  $\overline{x} \pm \sigma$ se concentra aproximadamente el 68% de los datos,
  $\overline{x} \pm 2\sigma$ se concentra aproximadamente el 95% de los datos,
  $\overline{x} \pm 3\sigma$ se concentra aproximadamente todos los datos.

Varianza

Se define como el cuadrado de la desviación típica. Se interpreta como la desviación típica solo que difiere en la magnitud y unidad de medida.

- *Datos sin agrupar*

La fórmula es:

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

En el caso en el que los datos estén **repetidos**, usaremos la fórmula

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

en donde x_i corresponde al valor de la característica, $\overline{X}$ la media de los datos y f_i la frecuencia de la característica i.

En caso tal que $n < 30$, usaremos la fórmula

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

Dicha corrección del denominador, se utiliza también para datos repetidos o agrupados, con n<30.

- *Datos agrupados*

  La fórmula es:

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

en donde x_i corresponde a la marca de clase del intervalo i, $\overline{X}$ la media de los datos y f_i la frecuencia del intervalo i.

Desviación Media

Es la desviación que presenta los datos con respecto a la mediana o a la media aritmética. Se usa usualmente cuando las desviaciones extremas influyen en la desviación típica.

- *Datos sin agrupar*

  Si los datos son simples, usamos la fórmula:

$$D.M. = \frac{\sum_{i=1}^{n} \left| x_i - \overline{X} \right|}{n}$$

Si los datos están repetidos entonces, usamos la fórmula:

$$D.M. = \frac{\sum_{i=1}^{n} \left| x_i - \overline{X} \right| f_i}{n}$$

en donde x_i corresponde al valor de la característica, $\overline{X}$ la media de los datos y f_i la frecuencia de la característica i.

- *Datos agrupados*

  Se calcula mediante la fórmula:

$$D.M. = \frac{\sum_{i=1}^{n} \left| x_i - \overline{X} \right| f_i}{n}$$

en donde x_i corresponde a la marca de clase del intervalo i, $\overline{X}$ la media de los datos y f_i la frecuencia del intervalo i.

Rango Cuartílico

Esta medida se basa en el cuartil 1 y cuartil 3, por lo que excluye el 25% inferior de los datos y el 25% superior de los mismos. Esto indica que el rango cuartílico mide la concentración de los datos en el 50% central de los mismos. El Rango Cuartílico expresa la distancia entre Q_1 y Q_3,

$$RQ = Q_3 - Q_1$$

En la medida que esa distancia sea menor, mayor será la concentración del 50% central de los datos. Si la distancia entre Q_1 y Q_3 es mayor, entonces hay una mayor dispersión del 50% central de los datos.

Características del Rango Cuartílico

- No es una medida segura de dispersión, ya que su valor se encuentra afectado por el 50% de los datos, 25% inferior y 25% superior. Igualmente obvia la distribución de datos entre Q_1 y Q_3.
- Es posible que dos series de datos con diferentes distribuciones presenten igual rango cuartílico, por ser iguales en los valores de Q_1 y Q_3.
- Una medida de dispersión derivada del rango cuartílicos, es la desviación semicuartil, que es la semisuma de Q_1 y Q_3.
- Es la medida de variabilidad adecuada cuando la mediana es la medida de tendencia central.

Coeficiente de Variación

Las medidas de variabilidad en general se expresan en las mismas unidades de los datos. A menudo es deseable comparar la variabilidad cuando las unidades de medición son diferentes. Así el Coeficiente de Variación es un índice de variabilidad que permite comparar el grado de dispersión entre distribuciones con respecto a la media aritmética. Nos permite expresar el grado de homogeneidad del grupo de datos considerados en su conjunto. Su fórmula es:

$$CV = \frac{S}{\overline{X}} 100\%$$

El coeficiente de variación depende de la desviación típica y de la media aritmética, por lo que a mayor coeficiente de variación, significa la existencia de mayor variabilidad entre los datos, y un CV pequeño indica menor variabilidad o mayor homogeneidad en los datos.

- *Ejemplo: Siguiendo con el ejemplo de los sueldos semanales (en miles de bolívares) de 80 bioanalistas del Laboratorio X.*

Sueldo semanal (en miles de bolívares) de 80 bioanalistas del Laboratorio X

X_i (Bs.)	f_i	F_a	$(X_i - \overline{X})^2 f_i$	$\lvert X_i - \overline{X} \rvert$	$\lvert X_i - \overline{X} \rvert$ f_i
100	6	6	2166	19.33	115.98
105	10	16	1960	14.33	143.30
115	25	41	400	4.33	108.25
120	18	59	18	0.67	12.06
123	12	71	192	3.67	44.04
135	7	78	1792	15.67	109.69
220	2	80	20402	100.67	201.66
TOTALES	80		26930		734.66

Se quiere calcular la desviación típica, la varianza, la desviación media, el rango cuartílico y el coeficiente de variación.

- *Desviación típica*

Calculamos en principio la media aritmética

$$\overline{X} = \frac{9546}{80} = 119 Bs$$

Así la desviación típica es:

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2 f_i}{n}} \Rightarrow$$

$$S=\sqrt{\frac{26930}{80}}=\sqrt{336.62}=18\,Bs$$

- *Varianza*

  Directamente se tiene que $S^2=(18Bs)^2=324Bs^2$

- *Desviación media*

  Se tiene que

$$D.M.=\frac{\sum|x_i-\overline{X}|f_i}{n}=\frac{734.66}{80}=9\,Bs.$$

- *Rango Cuartílico.*

  Calculamos Q_1 y Q_3. Para ello debemos hallar primero F_1 y F_3.

  $F_1=\frac{1x80}{4}=20$, esto implica que $Q_1=115\,Bs.$

  $F_3=\frac{3x\,80}{4}=60$, esto implica que $Q_3=123\,Bs.$

  Por lo tanto:

$$RC=Q_3-Q_1=123-115=8\,Bs$$

- *Coeficiente de Variación*

Se obtiene que

$$C.V. = \frac{18\,Bs}{119\,Bs} x100\% = 15\%$$

- *Ejemplo: Siguiendo con los datos de las estaturas de 50 estudiantes de la Escuela de Nutrición y Dietética tenemos:*

Estatura de 50 estudiantes de la Escuela de Nutrición y Dietética de la Universidad Central de Venezuela

$L_{ri} - L_{rs}$ (metros)	f_i	f_a	X_i	$X_i f_i$	$(X_I - \overline{X})^2 f_i$	$\left\| X_i - \overline{X} \right\| f_i$
1.45 – 1.48	2	2	1.465	2.93	0,031	0.250
1.48 – 1.51	7	9	1.495	10.46	0,063	0.665
1.51 – 1.54	4	13	1.525	6.10	0,017	0.260
1.54 – 1.57	3	16	1.555	4.66	0,004	0.105
1.57 – 1.60	12	28	1.585	19.02	0,000	0.060
1.60 – 1.63	9	37	1.615	14.53	0,006	0.225
1.63 – 1.66	4	41	1.645	6.58	0,012	0.220
1.66 – 1.69	4	45	1.675	6.70	0,029	0.340
1.69 – 1.72	2	47	1.705	3.41	0,026	0.230
1.72 – 1.75	3	50	1.735	5.20	0,063	0.435
	50			79.59	7,876	2.790

Se quiere calcular la desviación típica, la varianza, la desviación media, el rango cuartílico y el coeficiente de variación.

- *Desviación típica*

Calculamos en principio la media aritmética $\overline{X} = \frac{79.59}{50} = 1.59m$

Así la desviación típica es:

$$S = \sqrt{\frac{7.876}{50}} = 0.40\ m$$

- *Varianza*

  De forma directa se tiene que $S^2 = (0.40)^2 = 0.16m^2$

- *Desviación media*

  Se tiene que $D.M. = \frac{2.79}{50} = 0.0558m$

- *Rango Cuartílico*

  Calculamos Q_1 y Q_3. Para ello debemos hallar primero F_1 y F_3.

  $F_1 = \frac{1x50}{4} = 12.5$, por lo que $Q_1 = 1.51 + \frac{12.5 - 9}{4} x0.03 = 1.54m$

  y

  $F_3 = \frac{3x50}{4} = 37.5$, por lo que $Q_3 = 1.63 + \frac{37.5 - 37}{4} x0.03 = 1.63m$

  Por lo tanto $RC = (1.63 - 1.54)m = 0.09m$

- *Coeficiente de Variación*

  Se encontró que el $C.V. = \frac{0.40m}{1.59m} x100\% = 25\%$

Medidas de Forma

Una distribución queda bien caracterizada mediante la tendencia central y la variabilidad, pero quedará mejor si éstas medidas son acompañadas con medidas que describan la asimetría y apuntamiento de la distribución.

Sesgo

Las curvas que representan las observaciones de datos pueden ser simétricas o asimétricas (sesgadas). El Sesgo es un indicador que mide el grado de asimetría o falta de simetría de una distribución.

Así, el sesgo viene dado por la fórmula: $Sesgo = \frac{\overline{X} - Mo}{S}$

Una distribución se considerará simétrica si cada uno de los lados de la distribución que quedan a partes de la mediana, son exactamente de igual área y forma.

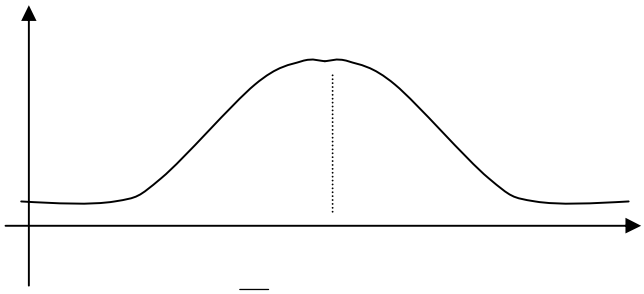


Figura 1

$$\overline{X} = Mo = Md$$

Si la acumulación de datos se encuentra hacia los valores bajos de la característica estudiada, se dice que la asimetría es positiva.

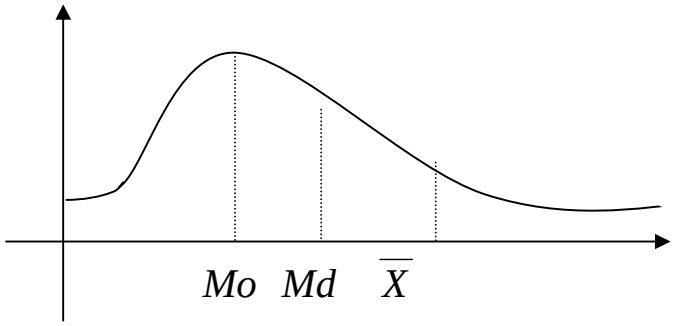


Figura 2

Si la acumulación de datos se encuentra hacia los valores altos de la característica estudiada, se dice que la asimetría es negativa.

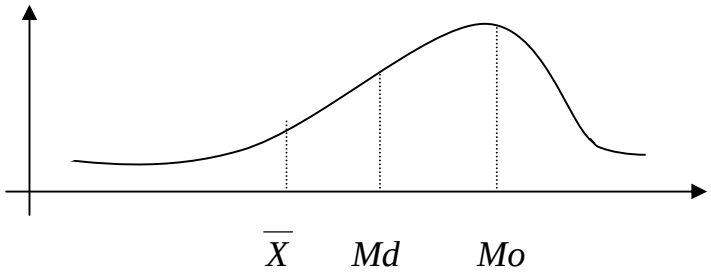


Figura 3

Características del Sesgo

- Si el sesgo es igual a 0, hay simetría. Figura 1.
- Si el sesgo es mayor a cero, la cola derecha es más larga que la izquierda respecto al valor central. Se dice que la asimetría es positiva. Figura 2.
- Si el sesgo es menor a cero, la cola derecha es más corta que la izquierda con respecto al valor central. La asimetría es negativa. Figura 3.
- Si una distribución tiene varias modas o carece de alguna, el sesgo se puede calcular mediante las siguientes fórmulas:

  - Sesgo de Pearson:

$$Sesgo = \frac{3\left(\overline{X} - Me\right)}{S}$$ para n mayor que 50

  - Sesgo Cuartílico (Bowley):

$$Sesgo = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

  - Sesgo Percentílico:

$$Sesgo = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

  - Sesgo por los momentos:

$$Sesgo = \frac{M_3}{S^3}$$

donde $M_k = \frac{\sum (x_i - \bar{X})^k}{n}$ es la fórmula de los momentos de orden *k*, en dado caso que los datos estén sin agrupar y

$M_k = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^k \, f_i}{n}$ es la fórmula de los momentos de orden *k*, en dado caso que los datos estén agrupados.

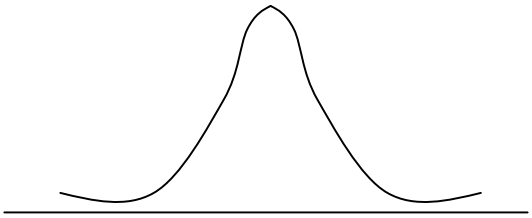
Curtosis

Es el grado de apuntamiento de una distribución con respecto a una curva modelo o curva normal de Laplace-Gauss. La fórmula es:

$K = \frac{M_4}{S^4}$ ó $K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$

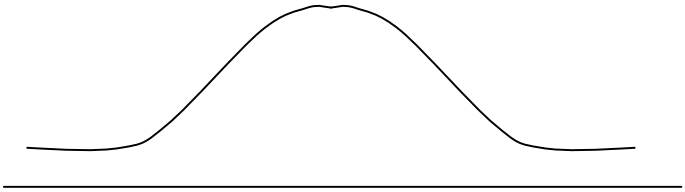
$K \to 3$ $K \to 0.263$

Si $K > 3$ ó $K > 0.263$ entonces la distribución es Leptocúrtica

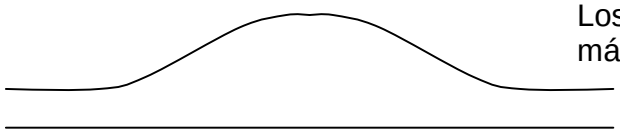


Los datos son menos dispersos

Si $K = 3$ ó $K = 0.263$ entonces la distribución es Mesocúrtica



Si $K < 3$ o $K < 0.263$ entonces la distribución es Platicúrtica.



- *Ejemplo: En la Escuela Bicentenaria, una muestra aleatoria de 80 niños en edad escolar, se observó su contenido de calcio(mg/dl) en la sangre que presentaban cada uno de ellos y se clasificaron según su consumo de leche diario:*
  Grupo A: 46 niños que toman menos de ½ litro de leche diario.
  Grupo B: 34 niños que tomas más de 1 litro de leche diario.
  Los datos obtenidos se muestran en la siguiente tabla:

$L_i - L_s$	f_i A	f_i B
7.1 – 7.6	*10*	*4*
7.7 – 8.2	*4*	*7*
8.3 – 8.8	*15*	*12*
8.9 – 9.4	*9*	*8*
9.5 – 10.0	*8*	*3*

Determine:

a) *El grado de variabilidad de cada muestra.*
b) *El grado de asimetría de cada muestra.*
c) *En función de los resultados anteriores, se podrá afirmar que la cantidad de ingesta de leche diaria influye en los contenidos de calcio en la sangre.*

Calculamos cada uno de los estadígrafos para ambos grupos:

Cantidad de calcio (mg/dl) en sangre de niños escogidos aleatoriamente del grupo A de la Escuela Bicentenaria

Nivel de calcio (mg/dL)		f_i	F_a	X_i	$(X_I - \overline{X})^2 f_i$
$L_i - L_s$	$L_i - L_s$				
7.1 7.6	7.05 7.65	10	10	7.35	15,625
7.7 8.2	7.65 8.25	4	14	7.95	1,69
8.3 8.8	8.25 8.85	15	29	8.55	0,0375
8.9 9.4	8.85 9.45	9	38	9.15	2,7225
9.5 10.0	9.45 10.05	8	46	9.75	10,58
Total		46			30,655

Datos suministrados por la unidad médica de la Escuela Bicentenaria.
Febrero, 1997.

El valor de la media es $\overline{X} = \frac{393.90}{46} = 8{,}6 mg/dl$

La desviación típica es $S = \sqrt{\frac{30.655}{46}} = 0.8 mg/dl$

De este modo, el grado de variabilidad obtenido es de $C.V. = \frac{0.8}{8.6} x100\% = 9\%$

Para calcular el grado de asimetría, necesitamos también obtener el valor de la moda: $Mo = 8.25 + \frac{11}{11+6} x0.6 = 8.6 mg/dl$. Así el sesgo será

$$Sesgo = \frac{8.56 - 8.63}{0.84} = -0.08$$

Cantidad de calcio (mg/dl) en sangre de niños escogidos aleatoriamente del grupo B de la Escuela Bicentenaria

Nivel de calcio (mg/dL)		f_i	F_a	X_i	$(X_I - \overline{X})^2 f_i$
$L_i - L_s$	$L_i - L_s$				
7.1 7.6	7.05 7.65	*4*	4	7.35	6,25
7.7 8.2	7.65 8.25	*7*	11	7.95	2,9575
8.3 8.8	8.25 8.85	*12*	23	8.55	0,03
8.9 9.4	8.85 9.45	*8*	31	9.15	2,42
9.5 10.0	9.45 10.05	*3*	34	9.75	3,9675
Total		34			15,625

Datos suministrados por la unidad médica de la Escuela Bicentenaria. Febrero, 1997.

.

El valor de la media es $\overline{X} = \frac{290.10}{34} = 8.5 mg/dl$

La desviación típica es $S = \sqrt{\frac{15.625}{34}} = 0.5 mg/dl$

De este modo, el grado de variabilidad obtenido es de

$$C.V. = \frac{0.5}{8.5} x100\% = 6\%$$

Para calcular el grado de asimetría, necesitamos también obtener el valor de la moda: $Mo = 8.25 + \frac{5}{5+4} x0.6 = 8.6 mg/dl$. Así el sesgo será

$$Sesgo = \frac{8.53 - 8.58}{0.70} = -0.07$$

Si observamos la muestra, no se podría decir que los niños que toman 1 litro de leche diaria (grupo B) posean más calcio en sangre que los que ingieren medio litro diario (grupo A), por el contrario se observa que en algunos casos, los niños que consumen menos de ½ litro poseen mayores niveles de calcio en sangre. Está inclinación no es muy fuerte, pues las muestras de ambos grupos A y B, son muy parecidas. Es así como el consumo de leche debe considerarse como una posible variable pero no de mucha importancia. Obsérvese los valores de la moda, media aritmética, desviación estándar, coeficiente de variación y sesgo.

Conclusión final: No hay evidencia de diferencia entre los niveles de calcio sérico entre niños que consumen un litro de leche diario y los que consumen medio litro.

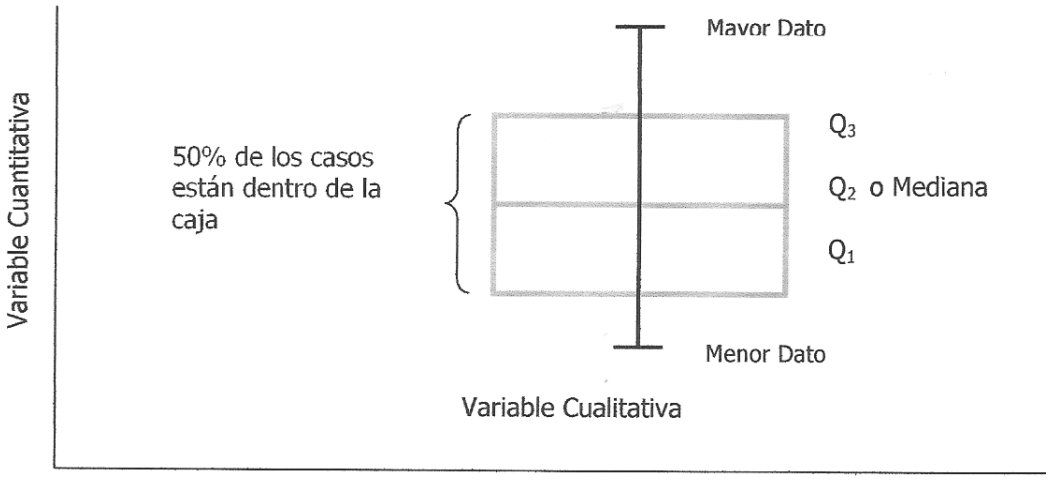
Aplicación: Diagrama de Caja

Dispositivo visual útil para comunicar la información contenida en un conjunto de datos. Para su construcción se usan los cuartiles o percentiles:

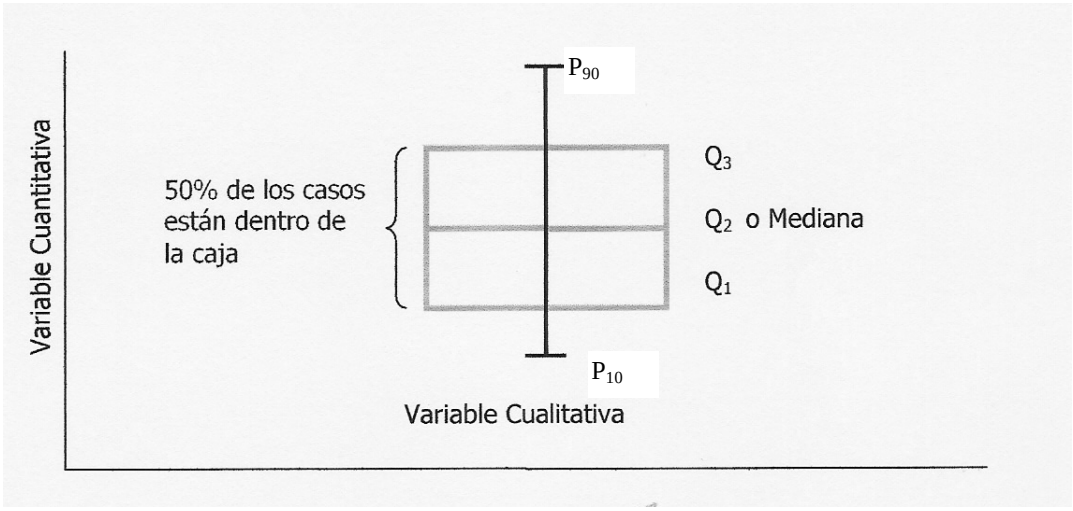
- Se representa la variable cuantitativa en el eje de las Y, la variable cualitativa en el eje de las X (una o más)
- Dibujar un eje vertical que se extienda desde la observación más pequeña hasta la más grande en los datos, cerrando cada observación con un apequeña línea horizontal.
- Dibujar sobre el eje vertical un cuadro que se extienda desde el cuartil Q_1 (extremo inferior) y cuartil Q_3 (extremo superior)
- Dividir el cuadro en dos partes con una línea que pase por el cuartil Q_2 o mediana.

Los diagramas de caja se caracterizan por requerir un solo eje; aquel en el cual se presentan sólo un resumen de los datos. La caja central representa valores de la mediana, los cuartiles 1 y 3. Dependiendo del usuario o paquete estadístico usado, los extremos de la línea vertical que divide la caja podrán significar diferentes medidas. Para ello observemos los siguientes ejemplos:

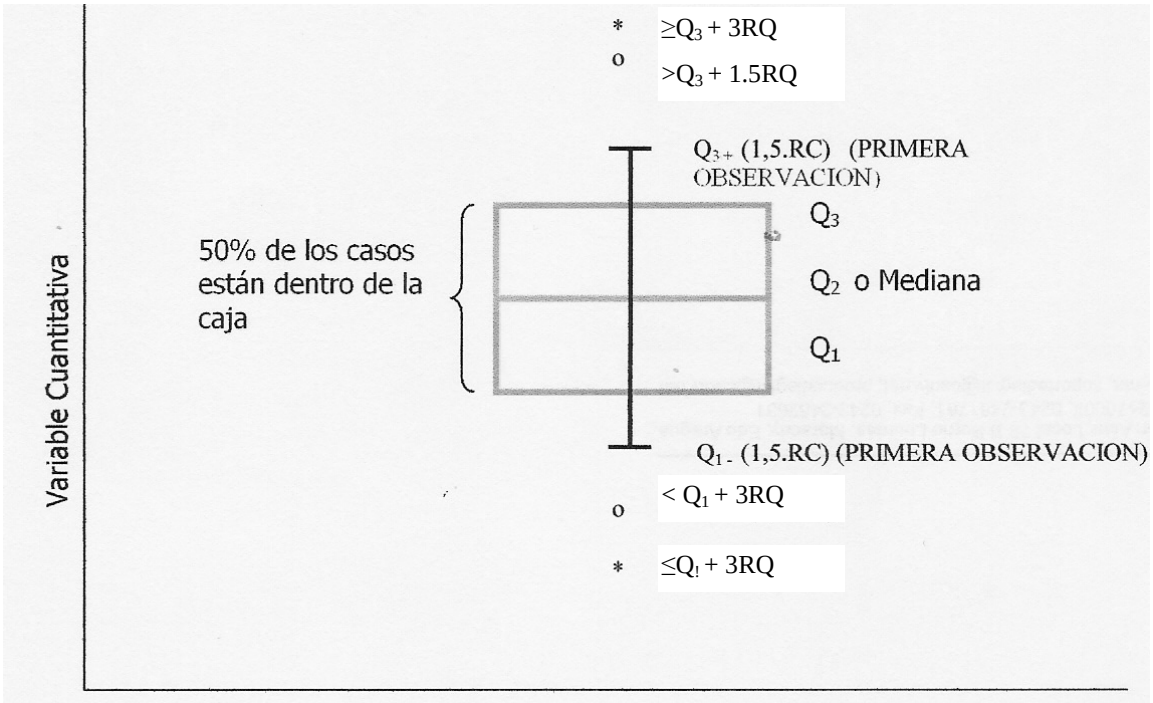
a) Gráfico de caja sencillo, representa el menor y mayor dato de la distribución.



b) Gráfico de caja que representa los percentiles 10 y 90 de la distribución.



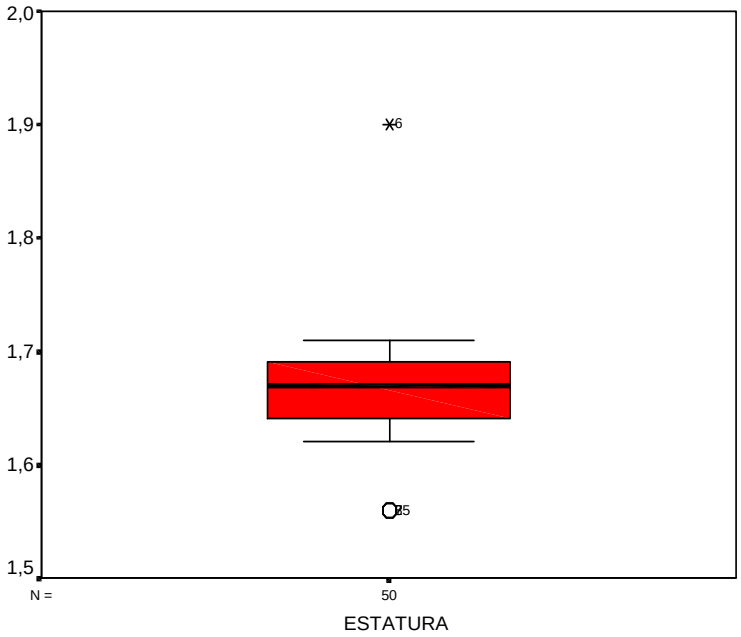
c) Gráfico de caja generado por el paquete SPSS, utiliza cuartiles y rango cuartílico. Además representa valores atípicos y aberrantes de la distribución.



Para entender mejor su construcción, mostremos en un diagrama de caja la distribución de 50 observaciones correspondientes a las estaturas de un grupo de estudiantes a la Escuela de Bioanálisis de la Universidad Central de Venezuela. Los datos los enumeramos a continuación:

1.56	1.56	1.56	1.62	1.62	1.63	1.63	1.64	1.64	1.64
1.64	1.64	1.64	1.65	1.65	1.65	1.65	1.65	1.65	1.65
1.65	1.65	1.65	1.65	1.67	1.67	1.67	1.68	1.68	1.68
1.68	1.68	1.68	1.68	1.68	1.68	1.69	1.69	1.69	1.69
1.69	1.69	1.69	1.69	1.70	1.70	1.70	1.70	1.71	1.90

Diagrama de Caja de la estatura de 50 estudiantes de la Escuela de Bioanálisis de Universidad Central de Venezuela



La caja central, que aparece de forma vertical en el ejemplo mostrado anteriormente, se extiende desde el cuartil 1 hasta el cuartil 3. Estos valores corresponden a $Q_1 = 1.64\,\text{m}$ y $Q_3 = 1.69\,\text{m}$. La línea que corre entre estos percentiles es la mediana, esto es, $Me = 1.67\,\text{m}$. Si la mediana se ubica aproximadamente a la mitad, entre los dos cuartiles, esto implica que las observaciones en el centro del conjunto de datos son aproximadamente simétricas.

Las líneas que se proyectan fuera de la caja a ambos lados se extienden a los valores adyacentes del diagrama. Los valores adyacentes son las

observaciones más extremas en el conjunto de datos, constituyendo estas aproximadamente el percentil 10 y el percentil 90. Para este ejemplo, el $P_{10} = 1.62\,\text{m}$ y $P_{90} = 1.70\,\text{m}$. Todos los puntos fuera de éste rango representados con círculos, constituyen las observaciones que se consideran valores atípicos, o puntos de datos que no son representativos del resto de los valores. Para este caso, se tienen datos atípicos hacia valores bajos, que corresponden a los tres individuos que miden 1.56m y un dato atípico por hacia valores altos, correspondiente al individuo que mide 1.90m.

En cuanto a las características generales de la distribución de los datos, se observa que existe una alta concentración de datos que se encuentran a valores bajos de la estatura.

CAPÍTULO III

Probabilidades

El problema central de la Estadística es el manejo del azar y la incertidumbre. Los eventos aleatorios siempre se han considerado como misteriosos. Los avances científicos de los siglos que siguieron al Renacimiento, enfatizando la observación y la experimentación cuidadosa, dieron lugar a la Teoría de Probabilidad para estudiar las leyes de la naturaleza y los problemas de la vida cotidiana.

Las estadísticas reemplazan las palabras imprecisas "pudo ser", "casi con seguridad", por un número que va de 0 a 1; esto indica una forma más precisa de qué tan probable o improbable es un evento.

En el campo médico los conceptos de probabilidad son útiles para comprender e interpretar datos presentados en cuadros y gráficas de informes publicados, además, permiten hacer enunciados acerca de cuánta es la confianza que se tiene en estimaciones de medias, proporciones y/o riesgos relativos.

Definición de Probabilidad

Clásica

La probabilidad que se dé un fenómeno determinado es igual al cociente entre el número de casos favorables al fenómeno y el número total de casos posibles.

Estadística

La probabilidad estimada de un suceso se toma como la frecuencia relativa de la aparición del suceso, cuando *n* es muy grande.

Conceptos Básicos

Experimentos aleatorios

Hay experimentos en los cuales los resultados no son esencialmente los mismos a pesar de que las condiciones sean aproximadamente idénticas; estos experimentos son denominados aleatorios. Para ello, es necesario que se satisfagan las siguientes condiciones:

a) Se puede repetir indefinidamente bajo idénticas condiciones.
b) Se conoce el conjunto de posibles resultados del experimento.

c) La aparición de cada resultado depende del azar.

Espacio Muestral o Universo (conjunto de puntos muestrales)

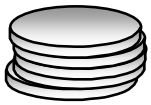
Es un conjunto que está formado por todos los resultados posibles de un experimento aleatorio; a cada uno de los resultados se denominan punto muestra. El espacio muestral usualmente se denota por *S.* El espacio muestral puede ser:

a) finito: tiene un número finito de puntos. (Discreto)
b) Infinito contable: tiene tantos puntos como números naturales. (Discreto)
c) Infinito no contable: tiene tantos puntos como hay en algún intervalo. (Continuo)

Sucesos o Eventos

Es un subconjunto del espacio muestral. Los eventos son denotados por (A, B, C,...). Si un suceso contiene un solo punto muestral, lo llamaremos suceso simple, en cambio que si contiene 2 o más puntos muestrales, lo llamaremos suceso compuesto.

❖ *Ejemplo 1: Si nos fijamos en el experimento de lanzar la moneda, el mismo será un experimento de una sola prueba y su espacio muestral tiene tan solo dos puntos muestrales(evento): S = {cara o sello}.*



❖ *Ejemplo 2: En una escuela rural, se va a seleccionar una muestra aleatoria de 2 niños cursantes del cuarto grado. Si se va a observar en cada niño la presencia o no de gingivitis ¿Cuál es el espacio muestral?*

Tenemos dos posibles casos:

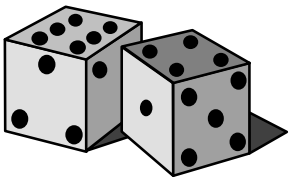
Gingivitis positiva → ☹

Gingivitis negativa → ☺

Luego:

S = { (☹☹) , (☹☺) , (☺☹) , (☺☺) }

❖ *Ejemplo 3: En el caso del lanzamiento de un dado (experimento), encontramos como espacio muestral a S = { 1,2,3,4,5,6 }*



El caso de un evento

Si en un espacio muestral tenemos un número finito de puntos muestrales, y cada uno tiene la misma probabilidad de darse; siendo s_i un punto muestral ($i = 1,2,...,m$) y m es el número total de puntos muestrales del espacio, la probabilidad de que se dé el punto muestral s_i es:

$$P(s_i) = \frac{1}{m}$$

El caso de dos o más eventos

Por otra parte, si m_a es el número de elementos del suceso A, A es un subconjunto de S y el número de elementos de S es m, entonces

$$P(A) = \frac{m_a}{m}$$

Es decir, la probabilidad de que se dé un determinado suceso A, es igual al cociente del número de casos favorables y el número total de casos posibles con la condición de que todos tengan la misma probabilidad de ocurrencia.

❖ *Ejemplo 4: ¿Cuál es la probabilidad de que al lanzar un dado, salga el número 5?*

La respuesta es P (salga un 5 en un solo lanzamiento) = 1 / 6

❖ *Ejemplo 5: ¿Cuál es la probabilidad de que al lanzar un dado, salga una cifra par? ¿y una cifra impar?*

Se tiene que

P(obtener un número par) = P(obtener un número impar) = 3 / 6 = 1 / 2

- *Ejemplo 6: ¿Cuál es la probabilidad de sacar una carta de corazón en un juego de cartas?*

  La respuesta es P(sacar un corazón en un juego de cartas) = 13 / 52.

Tipos de eventos

Eventos mutuamente excluyentes

Sean A y B dos subconjuntos de S. Decimos que A y B son mutuamente excluyentes si $A \cap B = \varnothing$, es decir, la aparición de uno de ellos impide la ocurrencia simultanea del otro. Para este caso, tenemos que:

$$P(A \cup B) = P(A) + P(B) \quad \text{(Regla aditiva)}$$

- Ejemplo 7: ¿Cuál es la probabilidad de extraer un tres de un juego de carta o de extraer un diez?

  Tenemos que

  P(A) =P(extraer un tres) = 4 / 52
  P(B) = P(extraer un diez) = 4 / 52

  Así $P(A \cup B) = 4 / 52 + 4 / 52 = 8 / 52$

- Ejemplo 8:¿Cuál es la probabilidad de extraer un siete de un juego de cartas o un dos?

  Tenemos que

  P(A) = P(extraer un siete) = 4 / 52
  P(B) = P(extraer un dos) = 4 / 52

  Así $P(A \cup B) = 4 / 52 + 4 / 52 = 8 / 52$

Eventos no mutuamente excluyentes

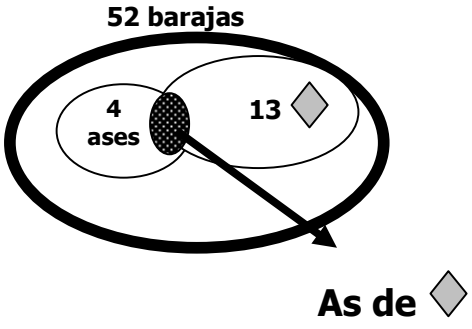
Para este caso, utilizaremos la fórmula:

$$P(A \cup B) = P(A) + P(B) – P(A \cap B)$$

En este caso los eventos parecen ser mutuamente excluyentes, pero existe una intersección en los eventos A y B, es decir, puede ocurrir que en el espacio muestral exista un evento que incluya a los eventos A y B, por lo tanto debemos restar dicha intersección para evitar contarla en las probabilidades de A y de B.

- *Ejemplo 9: ¿Cuál es la probabilidad de extraer un diamante de un juego de carta o un as?*

Tenemos que:



P(A) = P (extraer un diamante) = 13 / 52
P (B) = P (extraer un as) = 4 / 52
P(A ∩ B) =P (un as de diamante)= 1/ 52
P (AUB) = P (A) +P (B) - P (A ∩ B)

P (AUB) =13/52 + 4/52 - 1/52 = 16/52

- *Ejemplo 10: ¿Cuál es la probabilidad de extraer un trébol de un juego de carta, un diez o un dos?*

Tenemos que:

P(A) = P (extraer un trébol) = 13 / 52
P (B) = P (extraer un diez) = 4 / 52
P(C) = P (extraer un dos) = 4 / 52
P(A ∩ B) = P (un diez de trébol) = 1 / 52
P (B ∩ C) = 0
P(A ∩ C) =P (un dos de trébol) = 1 / 52

Así:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) – P(A \cap B) – P(A \cap C) – P(B \cap C)$$
$$= 13 / 52 + 4 / 52 + 4 / 52 – 1 / 52 – 0 – 1 / 52$$
$$= 19 / 52$$

Evento condicional

Sean A y B dos sucesos tales que P(A) > 0. Denotemos por P (B / A) la probabilidad de B dado que A ha ocurrido. Puesto que A ya ha ocurrido, el espacio muestral restante es reemplazado del original S. En este caso usamos la fórmula:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Otra forma de calcular la probabilidad de que tanto el suceso A como el suceso B ocurran, es usando la siguiente fórmula

$$P(A \cap B) = P(A)P(B/A) \quad \text{Regla multiplicativa}$$

- *Ejemplo 11: Se sabe que el 50% de la población fuma y que el 10% fuma y es hipertensa. ¿Cuál es la probabilidad de que si se escoge una persona fumadora, ésta sea hipertensa?*

  Tenemos que:

  P(F) =P(persona fumadora) = 0.50
  P(H) = P(persona hipertensa)
  P(F ∩H) = P(fumador e hipertensa) = 0.10
  Así:

$$P(H/F) = \frac{0.10}{0.50} = 0.20$$

- *Ejemplo 12: En una caja, hay 3 pelotas blancas y 5 negras. ¿Cuál es la probabilidad de sacar 1 blanca y 1 negra sin distinción de orden? Obtenga dicha probabilidad realizando un muestreo sin reposición.*

  Denotemos los eventos por A = extraer una bola blanca y B = extraer una bola negra.

  Como no nos interesa el orden de extracción, puede ocurrir.

  A1 = sacar la 1°blanca A2 = sacar la 1°negra
  B1 = sacar la 2°negra B2 = sacar la 2°blanca

  Así:

  P(extraer bola blanca y bola negra) = P(A1 ∩ B1) + P(A2 ∩ B2)

  donde

$P(A1 \cap B1) = P(A1) \,.\, P(B1 / A1) = 3 / 8.\ 5 / 7 = 15 / 56$
$P(A2 \cap B2) = P(A2).\ P(B2 / A2) = 5 / 8.\ 3 / 7 = 15 / 56$

Finalmente P(extraer bola blanca y bola negra) = 30 / 56

- Ejemplo 13: *En una caja, hay 3 pelotas blancas y 5 negras. ¿Cuál es la probabilidad de sacar 1 blanca y 1 negra sin distinción de orden? Obtenga dicha probabilidad realizando un muestreo con reposición.*

Denotemos los eventos por A = extraer una bola blanca y B = extraer una bola negra.

Como no nos interesa el orden de extracción, puede ocurrir.

A1 = sacar la 1°blanca	A2 = sacar la 1°negra
B1 = sacar la 2°negra	B2 = sacar la 2°blanca

Así:

P(extraer bola blanca y bola negra) = $P(A1 \cap B1) + P(A2 \cap B2)$

donde

$P(A1 \cap B1) = P(A1).\ P(B1 / A1) = 3 / 8.\ 5 / 8 = 15 / 64$
$P(A2 \cap B2) = P(A2).\ P(B2 / A2) = 5 / 8.\ 3 / 8 = 15 / 64$

Así:

P(extraer bola blanca y bola negra) = 30 / 64

Evento independiente

Dos sucesos A y B son independientes cuando la ocurrencia de uno no afecta la aparición del otro. En este caso tenemos que:

$$P(A \cap B) = P(A)P(B)$$

- *Ejemplo 14: Se lanzan 2 monedas. Hallar la probabilidad de que al lanzar la primera el resultado sea cara y al lanzar la segunda el resultado sea sello.*

Se observa que el espacio muestral es que S = { cc, cs, sc, ss }, donde

P(A) = P(primera moneda sea cara) = ½
P(B) = P(segunda moneda sea sello) = ½

De este modo la P(A ∩ B) = ½ . ½ = ¼

Axiomas de Probabilidad

Sean $A_1, A_2, ..., A_n$ subconjuntos del espacio muestral S. Se cumple que:

- Axioma 1

  Para todo A_i, se cumple que $0 \leq P(A_i) \leq 1$

- Axioma 2

  La suma de probabilidades de los ensayos en un conjunto mutuamente excluyentes es 1, es decir

$$\sum_{i=1}^{n} P(A_i) = 1$$

- Axioma 3

  Si $A_1, A_2, ..., A_n$ son mutuamente excluyentes, entonces:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

Teorema

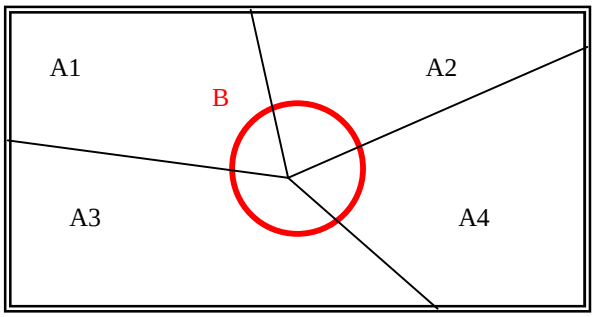
Sea A^c el suceso complemento de A, entonces

$$P(A^c) = 1 - P(A)$$

Particiones

Supongamos que tenemos los eventos A_1, A_2, A_3, A_4 los cuales son mutuamente excluyentes ($A_i \cap A_j = \varnothing$).

S



Entonces, obsérvese que $S = A_1 \cup A_2 \cup A_3 \cup A_4$

Además de que $B = S \cap B = (A_1 \cup A_2 \cup A_3 \cup A_4) \cap B$, y recordando las propiedades de conjunto tenemos

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup (A_4 \cap B)$$

De esta forma

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + P(A_4 \cap B) \\ &= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3) + P(A_4)P(B/A_4) \end{aligned}$$

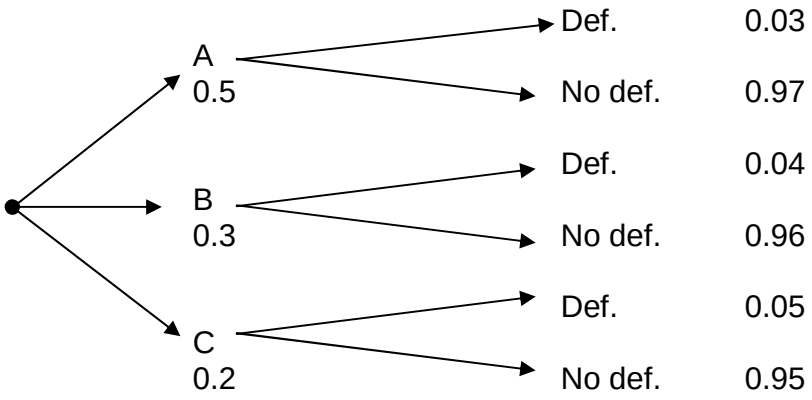
Esto se denota por la probabilidad total del evento B.

Teorema de Bayes

Sea S un espacio muestral que contiene *n* eventos $A_1, A_2, ..., A_n$ .Sea B un evento de S tal que $P(B) > 0$. La probabilidad de cualquier evento A_i, $i = 1,2,...,n$ dado el evento B es:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B/A_i)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \cdots + P(A_n)P(B/A_n)}$$

- *Ejemplo 15: Tres máquinas A, B y C producen respectivamente 50%, 30% y 20% del número total de artículos de una fábrica médica. Los porcentajes de defectos de producción de estas máquinas son 3%, 4% y 5% respectivamente. Si se selecciona al azar un artículo:*
  - *a) Hallar la probabilidad de que el artículo sea defectuoso.*
  - *b) Hallar la probabilidad de que el artículo sea no defectuoso.*
  - *c) Hallar la probabilidad de que si el artículo es defectuoso, provenga de la máquina A.*
  - *d) Hallar la probabilidad de que si el artículo no es defectuoso, provenga de la máquina B.*



a) El evento H buscado es encontrar un artículo defectuoso. Por lo que:

$$\begin{aligned} P(H) &= P(A \cap H) + P(B \cap H) + P(C \cap H) \\ &= P(A)P(H/A) + P(B)P(H/B) + P(C)P(H/C) \\ &= 0.5\,x\,0.03 + 0.3\,x\,0.04 + 0.2\,x\,0.05 \\ &= 0.037 \end{aligned}$$

b) El hecho de que un artículo sea no defectuoso, cae en el evento que es el complemento de H, por lo que

$$P(H^c) = 1 - P(H) = 1 - 0.037 = 0.963$$

c) En este caso tenemos

$$P(A/H) = \frac{P(A)P(H/A)}{P(A)P(H/A) + P(B)P(H/B) + P(C)P(H/C)}$$
$$= \frac{0.5\,x\,0.03}{0.5\,x\,0.03 + 0.3\,x\,0.04 + 0.2\,x\,0.05}$$
$$= 0.4054$$

d) En este caso tenemos

$$P(B\,No\,H) = \frac{P(B)P(No\,H/B)}{P(A)P(No\,H/A) + P(B)P(No\,H/B) + P(C)P(No\,H/C)}$$
$$= \frac{0.3\,x\,0.96}{0.5\,x\,0.97 + 0.3\,x\,0.96 + 0.2\,x\,0.95}$$
$$= 0.299$$

- *Ejemplo 16: Los datos recopilados en The Nacional Health Interview Survey de 1980-81. Los datos pertenecían a los daños al oído por lesiones sufridas por individuos mayores de 17 años. Las 163157 personas incluidas en el estudio se subdividieron en tres categorías mutuamente excluyentes:*

Condición Laboral	***Población***	***Presentaron daños***
Empleados	*98.917*	*552*
Desempleados	*7.462*	*27*
Fuera de la fuerza laboral	*56.778*	*368*
Total	*163.157*	*947*

a) Calcule la probabilidad de que al seleccionar una persona, el mismo esté en condición de empleado.
b) Calcule la probabilidad de que al seleccionar una persona, el mismo esté en condición de desempleados.
c) Calcule la probabilidad de que al seleccionar una persona, el mismo esté fuera de la fuerza laboral.
d) Calcule la probabilidad de que un individuo presente un daño en el oído sabiendo que se encuentra empleado.
e) Calcule la probabilidad de que un individuo presente un daño en el oído sabiendo que se encuentra desempleado.
f) Calcule la probabilidad de que un individuo presente un daño en el oído sabiendo que se encuentra fuera de la fuerza laboral.
g) Calcule la probabilidad de que un individuo seleccionado al azar presente lesión en el oído.
h) Calcule la probabilidad de que un individuo con lesión en el oído, se encuentre en condición laboral empleado.

a) Dicha probabilidad es:

$$P(individuo\,empleado) = \frac{98917}{163157} = 0.6063$$

b) Dicha probabilidad es:

$$P(individuo\ desempleado) = \frac{7462}{163157} = 0.0457$$

c) Dicha probabilidad es:

$$P(individuo\,fuera\ de\ la\ \ fuerza\ laboral) = \frac{56778}{163157} = 0.3480$$

d) Dicha probabilidad es:

$$P(daño\ en\ el\ oído / empleado) = \frac{552}{98917} = 0.0056$$

e) Dicha probabilidad es:

$$P(daño\ en\ el\ \ oído / desempleado) = \frac{27}{7462} = 0.0036$$

f) Dicha probabilidad es:

$$P(daño\ en\ el\,oído /\ \ fuera\,de\ la\ \ fuerza\ laboral) = \frac{368}{56778} = 0.0065$$

g) Dicha probabilidad es:

$$\begin{aligned} P(individuo\,con\ lesión) &= P(daño \cap empleado) + P(daño \cap desempleado) \\ &\quad + P(daño \cap \ fuera\ fuerza\ laboral) \\ &= 0.6063\,x\,0.0056 + 0.045\,x\,0.0036 + 0.3480\,x\,0.0065 \\ &= 0.034 + 0.0002 + 0.0023 \\ &= 0.0059 \end{aligned}$$

h) Dicha probabilidad es:

$$\begin{aligned} P(empleado / daño\ oído) &= \frac{P(empleado \cap daño)}{P(empleado \cap daño) + P(desempleado \cap daño) + P(\ fuera\ \ fuerz. \cap daño)} \\ &= \frac{0.0034}{0.0059} \\ &= 0.5762 \end{aligned}$$

Sensibilidad, Especificidad y Valores que Predicen Positividad y Negatividad

En el campo de las ciencias de la salud se aplican las leyes de la probabilidad y conceptos relacionados en la evaluación de pruebas de detección y criterios de diagnóstico. En nuestro campo, nos interesa tener mayor capacidad de predecir correctamente la presencia o ausencia de enfermedad a partir del conocimiento de los resultados positivos o negativos y el estado de los síntomas (presentes o ausentes).

En pruebas de detección pueden ocurrir los siguientes resultados:

- ☞ Valores falsos positivos: una prueba da positiva cuando debería dar negativa.
- ☞ Valores falsos negativos: una prueba da negativa cuando debería dar positiva.

Por lo tanto, las pruebas de detección no siempre son pruebas infalibles y se debe evaluar la utilidad de los resultados de la prueba y los síntomas del paciente para determinar si el individuo tiene o no alguna enfermedad.

Para ubicarnos, partiremos de la siguiente tabla de contingencia:

RESULTADO DE LA PRUEBA	ENFERMEDAD		TOTAL
	PRESENTE (E)	AUSENTE ($\overline{E}$)	
POSITIVO (+)	a	b	a+b
NEGATIVO (-)	c	d	c+d
TOTAL	a+c	b+d	N

Sensibilidad

La sensibilidad de una prueba o síntoma es la probabilidad de un resultado positivo de la prueba (presencia del síntoma) dada la presencia de la enfermedad. Sería calcular la estimación de la probabilidad condicional:

$$P(+/E) = \frac{a}{a+c}$$

Dado que un individuo tiene una enfermedad o síntoma el resultado de una prueba dé positivo.

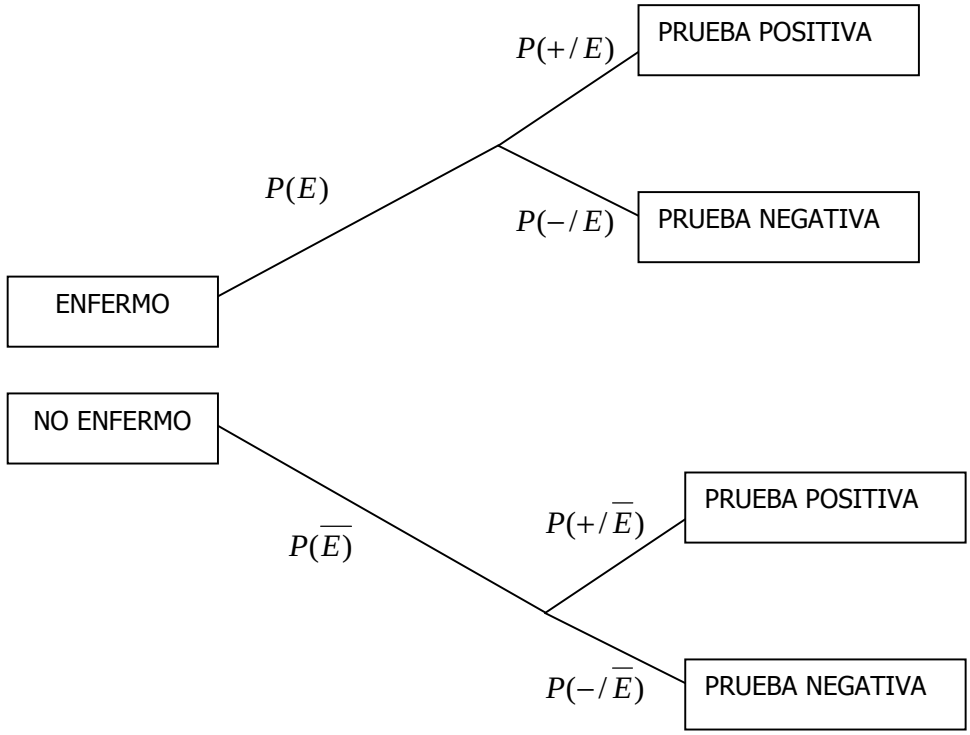
Especificidad

La especificidad de una prueba o síntoma es la probabilidad de un resultado negativo de la prueba (ausencia del síntoma) dada la ausencia de la enfermedad. Sería calcular la estimación de la probabilidad condicional:

$$P(-/\overline{E}) = \frac{d}{b+d}$$

Dado que un individuo No tiene una enfermedad o síntoma el resultado de una prueba dé negativo.

Para entender los Valores Predictivos de Positividad y Negatividad, será necesario plantearnos el siguiente árbol y aplicar el Teorema de Bayes:



Valor predictivo positivo. (VPP)

El valor que predice la positividad de una prueba de detección es la probabilidad de que un individuo tenga la enfermedad, dado que el individuo presenta un resultado positivo en la prueba de detección.

$$P(E/+) = \frac{P(+/E).P(E)}{P(+/E).P(E) \quad + \quad P(+/\overline{E}).P(\overline{E})}$$

Valor predictivo negativo. (VPN)

El valor que predice la negatividad de una prueba de detección es la probabilidad de que un individuo no tenga la enfermedad, dado que el individuo presenta un resultado negativo en la prueba de detección.

$$P(\overline{E}/-) = \frac{P(-/\overline{E}).P(\overline{E})}{P(-/\overline{E}).P(\overline{E}) \quad + \quad P(-/E).P(E)}$$

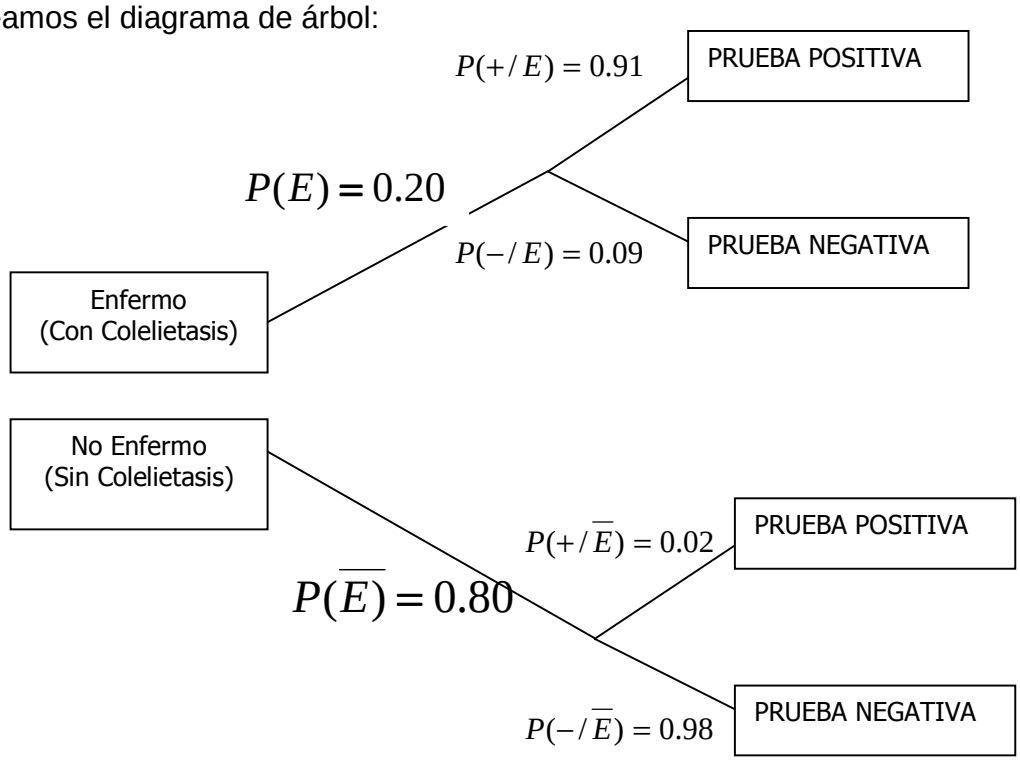
- *Ejemplo 17.- Un equipo de investigación pretende concer la sensibilidad y especificidad de una prueba de detección para VIH. La prueba se basa en una muestra aleatoria de 450 enfermos y portadores de la enfermedad y en otra aleatoria independiente de 500 pacientes que no presentan síntomas de enfermedad. Los resultados son los siguientes:*

RESULTADO DE LA PRUEBA	HIV		TOTAL
	PRESENTE (E)	AUSENTE ($\overline{E}$)	
POSITIVO (+)	436	5	441
NEGATIVO (-)	14	495	509
TOTAL	450	500	950

$$Sensibilidad = \frac{436}{450} = 0.97 \qquad Especificidad = \frac{495}{500} = 0.99$$

- *Ejemplo 18. Con el objeto de diagnosticar la colelietasis se usan los ultrasonidos. Tal técnica tiene una sensibilidad del 91% y una especificidad del 98%. En la población que nos ocupa, la probabilidad de padecer la enfermedad es de 0,2. ¿Cuál es el valor que predice la positividad de la prueba?*

Veamos el diagrama de árbol:



$$P(E/+) = \frac{0.20\,x\,0.91}{0.20\,x\,0.91 + 0.80\,x\,0.02} = 0.92$$

Valor que Predice la Positividad: 92%

Distribución de Probabilidades

Para referirse a las distribuciones probabilísticas existentes, es necesario y obligatorio hablar primero de lo que significa una variable aleatoria, debido a que en las distribuciones probabilísticas se trabajan generalmente con éste tipo de variables.

Variables Aleatorias (V.A.)

Una variable X es una variable aleatoria si es una magnitud susceptible de tomar diversos valores con determinadas probabilidades.

Existen dos tipos de variables aleatorias:

- ✓ *Discretas (V.A.D.):* es la que únicamente puede tomar un determinado número de valores en un intervalo.
- ✓ *Continuas (V.A.C.):* es la que puede tomar cualquier valor en un intervalo.

Definición de Distribución de Probabilidad y Función de Probabilidad

Denotamos por $P(X = a)$ la probabilidad del suceso correspondiente de que X tome el valor a y por $P(a \leq X \leq b)$ la probabilidad de que X tome valores desde a hasta b; entonces tenemos que dicho conjunto constituye la Distribución de Probabilidad.

Supongamos que tenemos una Variable Aleatoria X que puede tomar los valores $x_1, x_2, ..., x_n$ que pueden ser discretos o continuos, entonces cada uno de los valores tiene cierta probabilidad que en la práctica se desconoce, sin embargo a través de planteamientos teóricos podemos obtener dichas probabilidades, a las cuales designamos por $f(X)$, y al desarrollo que toman estos valores de $f(X)$, es lo que se llama Distribución de Probabilidad de la Variable Aleatoria.

Por otra parte, la Función de Probabilidad es aquella función $f(X)$ que mide la probabilidad que la Variable Aleatoria X tome determinados valores.

La Función de Probabilidad de una Variable Aleatoria Discreta X satisface que para todo $i,\ i = 1,2,...,n$

$$f(X = x_i) \geq 0$$

$$\sum_{i=1}^{n} f(X = x_i) = 1$$

La Función de Probabilidad de una Variable Aleatoria Continua X satisface que para todo $i,\ i = 1,2,...,n$

$$f(X = x_i) \geq 0$$

$$\int f(X = x_i) = 1$$

En este curso, solo estudiaremos las Distribuciones Discretas llamadas Binomial y Poisson, y entre las Distribuciones Continuas llamadas Normal, T de Student y Chi-Cuadrado.

Distribuciones de Probabilidad Discretas

Distribución Binomial

Los principios básicos de la Distribución Teórica Binomial los desarrolló el matemático suizo Jacob Bernoulli, en el siglo XVII. La Distribución Binomial proporciona la probabilidad de que un resultado específico ocurra de un número determinado de pruebas independientes. Bajo el supuesto de que durante n pruebas, la probabilidad de éxito en una sola prueba se mantenga fija; la determinación de la probabilidad de obtener un número dado de éxitos r, en las n pruebas, se simplifica utilizando la Distribución Binomial. Un experimento Binomial es aquel cuyo experimento consta de n pruebas idénticas en donde cada respuesta tiene dos posibles resultados: éxito o fracaso. Definamos la variable Aleatoria X como el número de éxitos al realizar n veces el experimento. En general, cualquier forma de obtener r éxitos en n veces tendrá probabilidad

$$p^r\, q^{n-r}$$

y la forma de obtener r éxitos y $n-r$ fracasos es igual a $\binom{n}{r}$.

Así:

$$P(X = r) = \binom{n}{r} p^r q^{n-r}$$

donde p es la probabilidad de éxito y $q = 1 - p$ es la probabilidad de fracaso.

Algunos ejemplos que podríamos citar, son determinar el número de llamadas en una central, número de colonias bacterianas por cajas de Pietro, entre otros.

Propiedades

- Media: $\mu = n.p$
- Varianza: $\sigma^2 = n.p.q$
- Desviación Típica: $\sigma = \sqrt{n.p.q}$
- Sesgo: $Sesgo = \frac{q-p}{\sigma}$
- Curtosis: $Curtosis = 3 + \frac{1-6.p.q}{n.p.q}$

Cuando n tiende al infinito, manteniendo p constante, la Distribución Binomial tipificada tiende a la Distribución Normal como límite.

❖ *Ejemplo 17: Un dado corriente se lanza 7 veces; llamamos a un lanzamiento un éxito si sale un 5 o un 6.*

a) Hallar la probabilidad de que salga un 5 ó un 6.

b) Hallar la probabilidad de que un 5 ó un 6 salga por lo menos una vez.

Definimos como X = salga un 5 o un 6 al lanzar el dado.

a) Se tiene que p = P (extraer un 5 ó un 6) = 1 / 6 + 1 / 6 = 2 / 6 = 1 / 3

por lo que q = 2 / 3.

b) Se tiene que

$$P(X) = 1 - P(no\ salga\ alguno)$$
$$= 1 - \binom{7}{0}\left(1/3\right)^0\left(2/3\right)^7 = 1 - 0.054$$
$$= 0.946$$

❖ *Ejemplo 18: Supongamos que en un cierto Centro de Rehabilitación cada persona tiene una probabilidad de 0.65 de recuperarse de su enfermedad en una semana. Encuentre la probabilidad de que si se seleccionan al azar 10 pacientes:*

a) A lo sumo 4 se recuperen en una semana.

b) *Se recuperen por lo menos 2 personas en una semana.*

Definamos por X = número de personas que se recuperan en una semana. El valor de $p = 0.65$ y $n = 10$.

a) Se tiene que:

$$\begin{aligned} P(X \le 4) &= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \\ &= \binom{10}{0}(0.65)^0(0.35)^{10} + \binom{10}{1}(0.65)^1(0.35)^9 \\ &\quad + \binom{10}{2}(0.65)^2(0.35)^8 + \binom{10}{3}(0.65)^3(0.35)^7 \\ &\quad + \binom{10}{4}(0.65)^4(0.35)^6 \\ &= 0.0949017 \end{aligned}$$

b) Usando el teorema del complemento

$$\begin{aligned} P(X \ge 2) &= 1 - [P(X=0) + P(X=1)] \\ &= 1 - \left[\binom{10}{0}(0.65)^0(0.35)^{10} + \binom{10}{1}(0.65)^1(0.35)^9\right] \\ &= 0.9994603 \end{aligned}$$

Distribución Poisson

La Distribución Teórica de Poisson se debe al matemático francés Simeón Poisson, es aplicable a fenómenos aleatorios que se caracterizan por el número de sucesos que ocurren en un determinado período de tiempo o en un determinado espacio. La hipótesis básica en este tipo de fenómenos es que los sucesos son independientes. Puede usarse para determinar la probabilidad de eventos poco frecuente, es decir, proporciona la probabilidad de que un resultado suceda un número específico de veces cuando la

cantidad de pruebas es grande y la probabilidad de ocurrencia es pequeña.

Sea X una variable aleatoria que representa el número de veces que ocurre el suceso, entonces

$$P(X = r) = f(r) = \frac{e^{-\mu}\mu^{r}}{r!}$$

para $r = 0,1,2,3,...$ y donde μ es un parámetro mayor que cero que representa el valor promedio que describe el evento.

Por ejemplo, esta distribución sirve para planear el número de camas que un hospital necesita en su unidad de cuidados intensivos, el número de células en un volumen determinado de líquido, el número de partículas que emite una cantidad específica de material radioactivo, etc.

Propiedades

- Media: $\mu = \lambda = n.p$
- Varianza: $\sigma^2 = \lambda$
- Desviación Típica: $\sigma = \sqrt{\lambda}$
- Sesgo: $Sesgo = 1/\sqrt{\lambda}$
- Curtosis: $Curtosis = 3 + 1/\lambda$

Esta distribución tiene aplicación cuando estamos en presencia de "eventos raros", los cuales se caracterizan por tener una probabilidad de ocurrencia muy pequeña en una población muy grande, por lo que generalmente $\mu = n.p < 5$, de esta forma, la distribución Binomial tiende a la distribución de Poisson.

❖ *Ejemplo 19: Supóngase que 300 erratas están distribuidas al azar a lo largo de un libro de 500 páginas. Hallar la probabilidad p de que una página dada contenga*

a) 2 erratas.

b) *2 o más erratas*.

Definimos a X como el número de erratas por página. Se tiene que $n = 300$, y la probabilidad de encontrar una errata en las 500 páginas es $1/500$. De este modo $\mu = n.p = 300x\,1/500 = 0.6$

a) Para este primer caso tenemos,

$$P(X = 2) = \frac{e^{-0.6}\ (0.6)^2}{2!} = 0.0988$$

b) Ahora queremos calcular la probabilidad de obtener 2 o más erratas:

$$\begin{aligned} P(X \geq 2) &= 1 - \left[P(X = 0) + P(X = 1)\right] \\ &= 1 - \left[\frac{e^{-0.6}\ (0.6)^0}{0!} + \frac{e^{-0.6}\ (0.6)^1}{1!}\right] \\ &= 0.122 \end{aligned}$$

❖ *Ejemplo 20: En un gran hospital, la probabilidad de recibir pacientes que presentan hemofilia es de 0.01. Si en una semana se atienden, aproximadamente 400 personas, cuál es la probabilidad de encontrar en este grupo un máximo de 3 hemofílicos.*

En este caso tenemos X = número de personas que presentan hemofilia. Se tiene que $p = 0.01$ y $n = 400$. Así el número promedio de hemofílicos por semana es $\mu = n.p = 400\ x\ 0.01 = 4$. Finalmente

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{e^{-4}(4)^0}{0!} + \frac{e^{-4}(4)^1}{1!} + \frac{e^{-4}(4)^2}{2!} + \frac{e^{-4}(4)^3}{3!} \\ &= 0.0183 + 0.0732 + 0.1465 + 0.1953 = 0.4333 \end{aligned}$$

Distribución de Probabilidades Continuas

Distribución Normal

Gauss y Laplace estudiaron la distribución de errores de las observaciones, concluyendo que todas las distribuciones estadísticas se aproximan a una curva que llamaron Normal, cuando el número de observaciones es grande.

La Distribución Normal es la distribución de probabilidad más famosa. Fue descubierta por primera vez por el matemático francés Abraham DeMoivre, quien publicó sus trabajos en 1733. Sin embargo, dos astrónomos matemáticos, Pierre-Simon Laplace de Francia y Carl Friedrich Gauss de Alemania, se ocuparon de establecer los principios científicos de la distribución normal.

La media μ y la desviación estándar σ son los parámetros de la distribución normal, esto es, μ y σ determinan completamente la ubicación de las cantidades y la forma de la curva.

La distribución Normal, curva normal o distribución de Gauss es una función entre dos variables continuas, x e y, dada por la ecuación:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \quad e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde μ es la media aritmética y σ es la desviación típica de los datos. Haciendo un cambio de origen y escala a través de la ecuación

$$z = \frac{x - \mu}{\sigma}$$

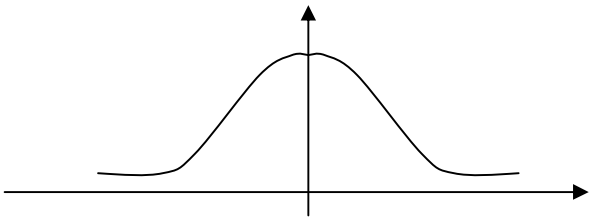
encontramos la nueva ecuación

$$y = \frac{1}{\sqrt{2\pi}} \quad e^{-\frac{1}{2}z^2}$$

la cual es la forma tipificada de la distribución normal. Si X se distribuye normalmente, esto lo denotaremos por $X \sim N(\mu, \sigma^2)$. Usando el cambio de escala para trabajar con la forma tipificada, nos queda $Z \sim N(0,1)$, lo cual nos permitirá encontrar probabilidades a partir de la tabla de la normal.

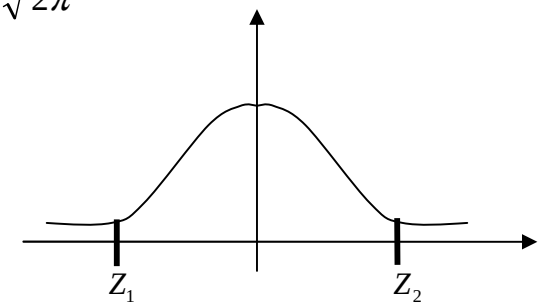
Propiedades

- Para la curva normal tipificada, se tiene que la media es igual a 0 y la desviación típica es 1.
- Esta curva es simétrica con respecto a la media, de modo que la media, la mediana y moda coinciden.
- El valor del sesgo es 0.
- Es una curva mesocúrtica.
- El área bajo la curva es igual a 1.



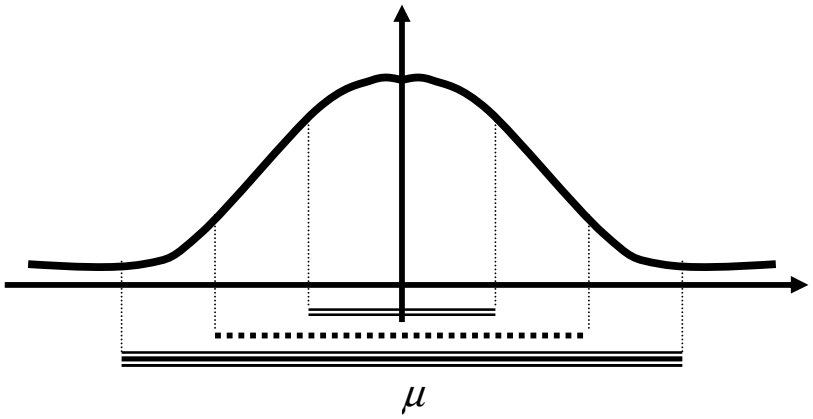
- El área bajo la curva comprendida entre las ordenadas correspondientes a los valores X_1 y X_2 ó entre sus tipificaciones Z_1 y Z_2, es la probabilidad de que la variable X o Z tome valores comprendidos entre esos puntos

$$A = \int_{Z_1}^{Z_2} \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2} z^2} \; dz = P(Z_1 \leq Z \leq Z_2)$$



Estos valores los buscaremos en una tabla.

Regla Empírica



A partir del eje de simetría μ tenemos que

═══ $\mu \pm \sigma$ contiene aproximadamente el 68% de los datos.

▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪ $\mu \pm 2\sigma$ contiene aproximadamente el 95% de los datos.

═══════════════════ $\mu \pm 3\sigma$ contiene aproximadamente todos los datos.

Observación:

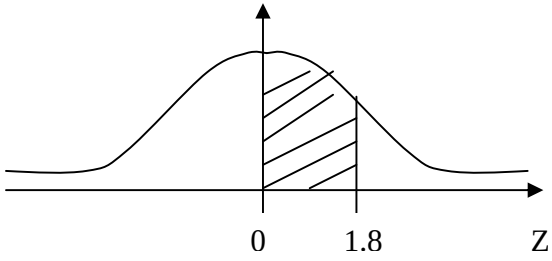
Trabajaremos con las siguientes tablas:

1.- Áreas bajo la curva normal tipificada de 0 a z.
2.- Distribución t de Student con ν grados de libertad.
3.- Distribución Chi-cuadrado con ν grados de libertad.

❖ *Ejemplo 21: En la tabla de la curva normal tipificada de 0 a z, halle las siguientes áreas:*

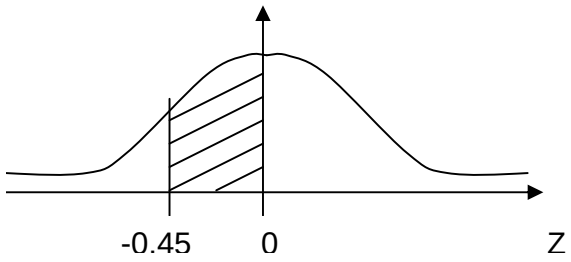
a) Area entre Z=0 y Z=1,80

P(0< Z < 1,80) = 0,4641



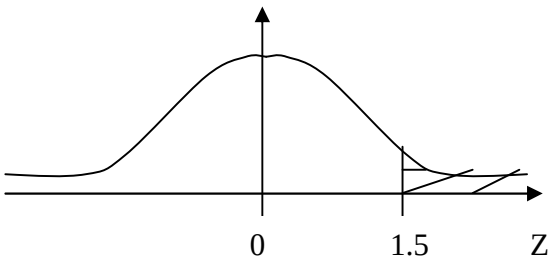
b) Area entre Z = -0,45 y Z=0

P(-0,45 < Z < 0) = P(0 < Z < 0,45)
= 0,1736



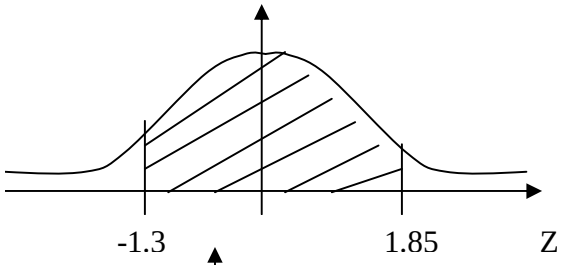
c) Área a la derecha de 1,50

P(Z > 1,50) = 0,5 –P(Z < 1,50)
= 0,5 – 0,4332
= 0,0668



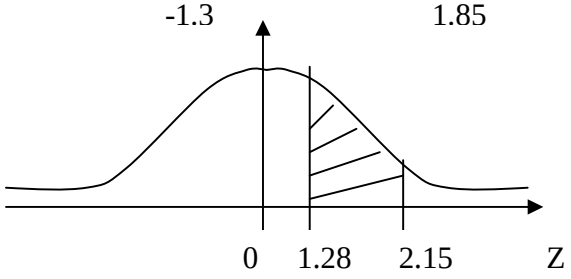
d) Area entre Z = -1,3 y Z = 1,85

P(-1,3 < Z < 1,85) =
= P(Z < 1,85) + P(Z < 1,3)
= 0,4678 + 0,4032
= 0,871



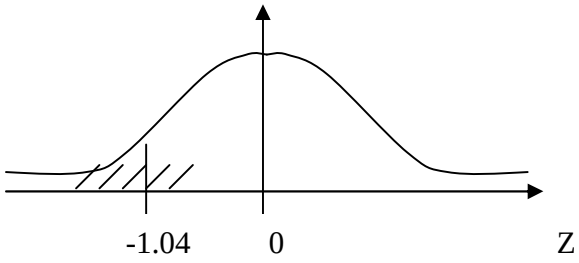
e) Área entre Z = 1,28 y Z = 2,15

P(1,28 < Z < 2,15) =
P(Z < 2,15) – P(Z < 1,28)
= 0,4842 – 0,3997
= 0,0845



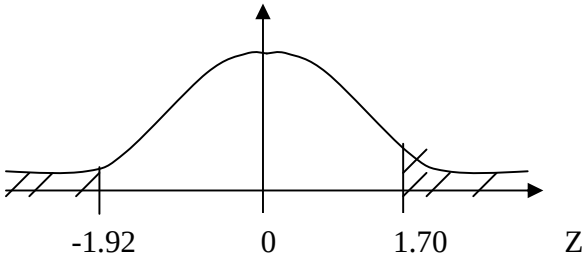
f) Área a la izquierda de –1,04

P(Z < - 1,04) = 0,5 – P(Z < 1,04)
= 0,5 – 0,3508
= 0,1492



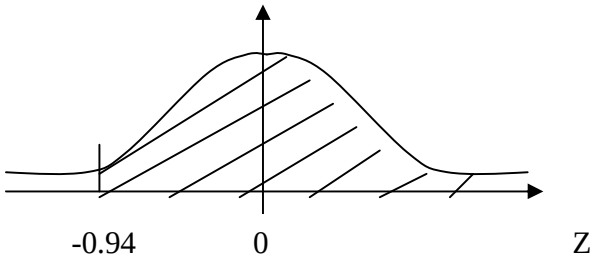
g) Área a la izquierda de –1,92 y a la derecha de 1,70

= P(Z < - 1,92) + P(Z < 1,7)
= 1 – P(- 1,92 < Z < 1,7)
= 1 – (0,4554 + 0,4726) = 0,072



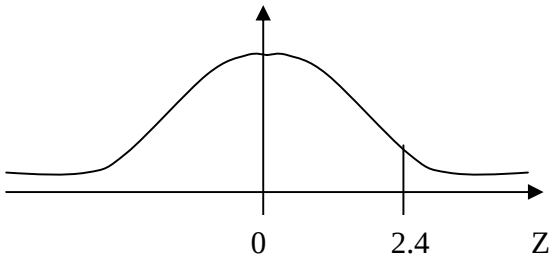
h) Área a la derecha de –0,94

A = 0,5 + P(Z < 0,94)
= 0,5 + 0,3264 = 0,8264



❖ *Ejemplo 22: Las calificaciones de un examen se distribuyen normalmente con una media aritmética de 50 puntos y una desviación típica de 5 puntos. Si el total de alumnos es de 80, hallar la probabilidad de que una nota elegida al azar sea mayor a 62 puntos y el número esperado de personas con más de 62 puntos.*

Tenemos que X = calificación de un examen. Se tiene que la media aritmética $\mu = 50\,ptos$, $\sigma = 5\,ptos$ y $n = 80$



Hallamos la distribución normal tipificada

$$z = \frac{x - \mu}{\sigma} = \frac{62 - 50}{5} = 2.4$$

Luego

$$\begin{aligned} P(X > 62) &= P(Z > 2.4) \\ &= 0.5 - P(0 \le Z \le 2.4) \\ &= 0.5 - 0.4918 \\ &= 0.0082 \end{aligned}$$

Por otro lado, el número esperado de personas con calificación superior a 62 puntos es

$$\#esperado = n.p = 80\,x\,0.0082 = 0.656 \to 1$$

CAPÍTULO IV

Inferencia Estadística

"Uno de los propósitos de la investigación es el realizar inferencias o generalizar de una muestra a una población más grande. Para poder comprender los fundamentos del muestreo, resulta imprescindible el conocer las definiciones de Universo, Población y Muestra.

Universo / Población : *El Universo en definitiva constituye una población teórica sobre la cual los estadísticos han creado toda la teoría del muestreo; se suele asimilar a la población más amplia que se quiere conocer con un estudio pero que por obvias razones es imposible de alcanzar. La* ***población,*** *es un conjunto o colección grande de artículos que poseen algo en común. Esta definición traducida al uso en medicina podría ser: el conjunto de sujetos u organismos que poseen una característica en común, susceptible de estudio, medición u observación. Para hacer más digerible esta definición, supongamos que un investigador desea establecer la prevalencia de uso de aretes en el ombligo, en mujeres rubias menores de 20 años de edad, de ojos verdes, cuya estatura sea mayor de 1.75m y residan en la ciudad de Quito. Aparentemente el investigador busca firmemente una candidata a Miss, pero bueno en este caso están claramente definidas las características de la población a investigar:* ***"rubias", "ojos verdes", "estatura 1.75m", "residentes en Quito"*** *en definitiva se trata de un grupo de sujetos (en este caso mujeres) que deben tener en común necesariamente las características citadas".*

La Inferencia Estadística constituye una parte de la Estadística en la que se hacen estimaciones e inferencias para la toma de decisiones. Aquí utilizaremos técnicas de muestreo apropiadas a fin de estudiar determinadas características de la población que nos interesa analizar, tomando en cuenta que la muestra a estudiar, debe ser representativa de la población.

Muestreo Estadístico

El conjunto de técnicas que nos permiten diseñar la muestra más apropiada para un experimento, garantizando que esta sea representativa de la población de origen y controlar los errores cometidos, es lo que se conoce como Muestreo Estadístico. Dependiendo de la investigación a realizar, utilizaremos el método que mejor se adecua al mismo (muestreo aleatorio simple, muestreo estratificado, entre otros).

Se aconseja la utilización del muestreo cuando la población es infinita, cuando las muestras son homogéneas, cuando el proceso de investigación de la característica de un elemento resulte destructivo.

Ventajas del Muestreo

- ✓ Economía y rapidez en su realización.
- ✓ Más alcance en la investigación.
- ✓ Más entrenamiento, formación y control del personal.
- ✓ Mayor rapidez de procesamiento y presentación de resultados.
- ✓ Fácil verificación posterior.
- ✓ Mayor confiabilidad de los datos obtenidos.

Limitaciones del Muestreo

- ✓ No permite hacer cálculos, tabulaciones o proyecciones con respecto a área o grupo pequeños.
- ✓ Presenta el error de muestreo.
- ✓ Se requiere de una preparación estadístico-matemática.

A la hora de escoger una muestra, se supone que las muestras son obtenidas a través del muestreo simple aleatorio, según el cual cada elemento de la población tiene idéntica probabilidad de ser escogido en una muestra. Los elementos de la muestra son variables aleatorias independientes y la muestra recibe el nombre de Muestra Aleatoria. Dicha muestra aleatoria tiene asociada una función de densidad $f(x)$.

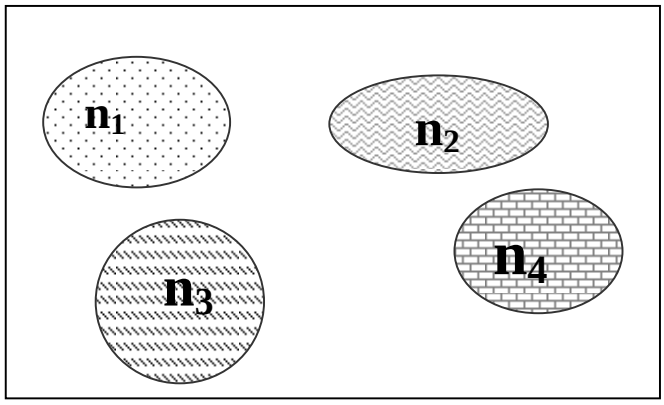
Tenemos dos formas de escoger los elementos de una muestra, una cuando cada elemento que se selecciona puede ser seleccionado nuevamente para constituir la muestra, el cual constituye un muestreo con reemplazamiento, y otra, por el contrario, cuando cada elemento no puede ser seleccionado más de una vez para formar la muestra, el cual constituye el muestreo sin reemplazamiento.

Para el caso del muestreo con reemplazamiento permanece constante la probabilidad de selección de cada elemento que va a integra la muestra y en el caso del muestreo sin reemplazamiento, la probabilidad cambia cada vez, puesto que vamos a seleccionar entre un número menor de elementos.

Distribuciones muestrales

Si tenemos una población de tamaño N, de ella es posible extraer con o sin reemplazamiento una serie de muestras (n).

Población(N)



Para cada muestra se puede calcular cualquier estadístico: media, varianza, desviación típica, otros, los cuales serán distintos para cada muestra.

La agrupación de los valores así obtenidos para un estadístico es la distribución muestral de ese estadístico. Si ese estadístico es la media, entonces tendremos la distribución muestral de la media de los datos, si el estadístico es la varianza, entonces tendremos la distribución muestral de la varianza de los datos, y así sucesivamente.

Para cada distribución muestral, es posible calcular la media por ejemplo, obteniendo así la media de la distribución muestral.

Obsérvese que los estadísticos muestrales son variables aleatorias, por ello tendrán asociado una distribución de probabilidad.

Observación: Existe el llamado Teorema del Límite Central el cual en esencia dice que cuando n aumenta, la distribución de las medias tiende a una ley normal con media y varianza específica. Dicho teorema se verifica cuando el tamaño de la muestra es igual o superior a 30. Enunciemos dicho teorema.

Teorema del Límite Central

Sea una población en la que una variable aleatoria X sigue una ley de probabilidad cualquiera, de media μ y varianza σ^2. Si extraemos de dicha población muestras al azar formadas cada una de ellas por un

conjunto de n observaciones independientes $x_1, x_2, ..., x_n$. La distribución del conjunto de la media $\overline{X}$ observadas en dichas muestras tiene por media a $\mu_{\overline{X}}$ y por varianza a $\sigma^2_{\overline{X}} = \sigma^2 / n$.

En lo sucesivo trabajaremos con muestras que son independientes.

Distribución de la Media Muestral

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de tamaño n de una población. La Media Muestral es

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

La población tiene distribución

$N(\mu, \sigma^2)$

Sea el estadístico $\overline{X}$ de una muestra aleatorio extraída de una población de media μ y varianza σ^2. La media muestral de una muestra aleatoria de tamaño n tomada de una población normal, es una variable normalmente distribuida con media

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

y varianza

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$$

Desviación Típica

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

Si n es mayor o igual a 30 entonces

$$\bar{x} \approx N(\mu, \sigma^2/n)$$

La distribución de la población tiene media μ pero no se conoce la varianza

En este caso utilizaremos la distribución t de Student, usando una función de $\bar{X}$ que contiene la varianza muestral en lugar de σ^2.

Tendremos que

$$\frac{\bar{x}-\mu}{S/\sqrt{n}} \approx t_{n-1}$$

donde t_{n-1} es la distribución t de Student con $n-1$ grados de libertad y

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

Distribución muestral de proporciones. (población finita)

Se considera todas las posibles muestras de tamaño n extraída de una población y para cada muestra se determina la proporción p de éxito. Entonces se obtiene una distribución muestral de proporciones cuya media es μ_p y desviación típica σ_p y viene dada por una media de

$$\mu_{\hat{p}} = p$$

y desviación típica de

$$\sigma_{\hat{p}} = \sqrt{\frac{p.q}{n}}$$

Para n ≥ 30, la distribución muestral se aproxima mucho a una distribución normal, la población se distribuye binomialmente. La proporción de éxitos

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

en una muestra aleatoria tomada de una población punto-binomial, se distribuye

$$\hat{p} \approx N(p, \frac{p.q}{n})$$

Distribución muestral de las diferencias

Si se tienen dos poblaciones, de las cuales se extraen muestras de tamaño n_1 de la población 1(llamémosla X_1) y n_2 de la población 2 (llamémosla X_2). De la primera se calcula un estadístico S_1, donde la distribución muestral tiene media μ_{S_1} y desviación típica σ_{S_1}. De igual forma, para la segunda población se calcula el estadístico S_2, cuya media es μ_{S_2} y desviación típica σ_{S_2}. De acuerdo a ello, podemos obtener una distribución muestral de diferencias $S_1 - S_2$ que se conoce como distribución muestral de las diferencias del estadístico, con media $\mu_{S_1-S_2}$ y desviación típica $\sigma_{S_1-S_2}$. Tomando en cuenta que las muestras sean independientes y las varianzas son conocidas, tenemos:

Distribución muestral de la diferencia de medias:

La media muestral es:

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

y la desviación típica es

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$$

Para n_1 y n_2 ambos mayores o iguales a 30, la distribución se comporta como una normal.

Por otra parte si las varianzas poblacionales no son conocidas, como

generalmente sucede, éstas se aproximan por las cuasivarianzas muestrales, quedando:

$$\sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}$$

De modo que en este caso

$$\overline{X}_1 - \overline{X}_2 \approx N\left(\mu_1-\mu_2, \frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}\right)$$

Si $\sigma^2 = \sigma_1^2 = \sigma_2^2$ entonces

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1-\mu_2\right)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2}\right)}} \approx t_{n_1+n_2-2}$$

Distribución muestral de la diferencia de proporciones:

La media muestral es:

$$\mu_{\hat{p}_1-\hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} + \frac{x_2}{n_2}$$

y la desviación típica es

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \qquad \text{donde } p = \frac{x_1+x_2}{n_1+n_2}$$

Intervalos de Confianza

Si queremos estimar un parámetro de una población mediante una muestra de tamaño n, podemos obtener muchos valores distintos de ese parámetro muestral. Lo más recomendable es encontrar un intervalo alrededor del valor del estimador, acompañado de alguna medida que nos diga la confianza que se puede tener de que ese intervalo contenga el verdadero valor del parámetro.

Dada una muestra aleatoria de tamaño n de una población con función de densidad $f(x)$, un intervalo de confianza del $100(1-\alpha)\%$ para un parámetro desconocido θ, es un intervalo determinado por dos números: $(\theta-\delta_1, \theta-\delta_2)$ calculados con base en los datos de la muestra tales que

$$P(\hat{\theta}-\delta_1 \leq \theta \leq \hat{\theta}+\delta_2) = 1-\alpha$$

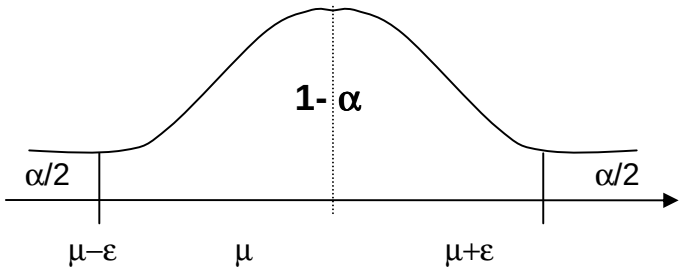
El valor $100(1-\alpha)\%$ es conocido como coeficiente de confianza.

Los límites superiores e inferiores del intervalo de confianza que contiene una proporción $1-\alpha$ de las medidas, los cuales constituyen variables aleatorias, se hallan por medio de la fórmula:

$$x = \mu \pm Z_{\alpha}\,\sigma$$

El intervalo de probabilidad $1-\alpha$ de una media por ejemplo, se halla como sigue:

$$\mu \pm e = \mu \pm Z_{\alpha}\frac{\sigma}{\sqrt{n}}$$



Teoría de Estimación Estadística

Sea μ_S y σ_S la media y desviación típica de la distribución muestral del estadístico S. Si esta distribución muestral se aproxima a una normal, cabe esperar en muestras extraídas el estadístico S se encuentre en los intervalos $\mu_S \pm \sigma_S$; $\mu_S \pm 2\sigma_S$; $\mu_S \pm 3\sigma_S$. Por ello se puede llamar intervalo de confianza para la estima μ_S. Los números extremos de estos intervalos se llaman límites de confianza.

El porcentaje de confianza es llamado nivel de confianza y los valores Z_c correspondientes a los límites de confianza se llaman coeficientes de confianza o valores críticos.

Intervalo de confianza para la media

Si el estadístico S es la media muestral $\bar{x}$, entonces los límites de confianza son dados por

$$\bar{x} \pm Z_c \sigma_{\bar{x}}$$

donde Z_c depende del nivel de confianza y $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

Intervalo de confianza para proporciones

Si el estadístico S es la proporción de éxitos en una muestra de tamaño n extraída de una población binomial en la que p es la proporción de éxitos, los límites de confianza para p vienen dados por

$$\hat{p} \pm Z_c \sigma_{\hat{p}}$$

donde p es la proporción de éxitos en la muestra de tamaño n y $\sigma_{\hat{p}} = \sqrt{p.q / n}$

Intervalo de confianza para la diferencia de medias

Para este caso, tendremos

$$\overline{X_1} - \overline{X_2} \pm Z_c \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Intervalo de confianza para la diferencia de proporciones

Para este caso, tendremos

$$\hat{p}_1 - \hat{p}_2 \pm Z_c \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} \qquad \text{donde} \qquad p = \frac{x_1 + x_2}{n_1 + n_2}$$

Teoría de la Decisión Estadística, Ensayos de Hipótesis y Significación

Decisión estadística

Son decisiones sobre poblaciones, tomadas a partir de la información muestral de las mismas.

Hipótesis Estadística

Una hipótesis estadística es una conjetura que se realiza respecto a una población, más concretamente, respecto a un parámetro de la población el cual cuantifica una característica de ella. Se formulan hipótesis con el solo propósito de rechazarla o aceptarla. Trabajaremos con las llamadas hipótesis nula y la hipótesis alternativa.

Hipótesis nula

Se denota por H_0 y es la hipótesis que se establece con el propósito de ver su posible rechazo. Por ejemplo, se comienza por afirmar que la media de la población es igual a un valor dado μ_0, y se denota $H_o : \mu = \mu_0$

Hipótesis alternativa

Se denota por H_1 y es cualquier hipótesis que difiere de la hipótesis nula, referida la misma medida estadística, por lo tanto contradice a H_0. En una prueba hay generalmente una hipótesis nula, pero puede haber muchas hipótesis alternativas, a saber, $H_1 : \mu \neq \mu_0; H_1 : \mu > \mu_0; \; H_1 : \mu < \mu_0$

Siguiendo con el ejemplo de la media, comparamos la media muestral $\bar{x}$ con la media propuesta μ_0 específicamente, deseamos saber si la diferencia entre la media de muestreo y la media hipotética es demasiado grande para atribuirla a la pura casualidad.

Tipos de Error: Error tipo I y tipo II.

Al tomar una decisión respecto a una hipótesis nula considerada, partiendo del estadístico obtenido a partir de la muestra, se puede incurrir en errores, el cual lo veremos en el siguiente cuadro:

	Aceptar H_0	Rechazar H_0
H_0: verdadera	Correcto	Error tipo I
H_1: verdadera	Error tipo II	Correcto

Es decir, que el error tipo I se presenta al rechazar una hipótesis cuando ésta es verdadera y el error tipo II se presenta al aceptar una hipótesis siendo falsa.

Siguiendo con el ejemplo de la media, si existe evidencia que la muestra no puede provenir de una población con media μ_0, rechazamos la hipótesis nula. Esto ocurre cuando, en el supuesto que H_0 sea verdadera, la probabilidad de obtener una media de muestreo tan extrema o más que el valor observado $\bar{x}$ es suficientemente pequeña. Por tanto, concluimos que la media de la población no puede ser μ_0; se dice que dicho resultado de la prueba es estadísticamente significativa. Si no existe suficiente evidencia para dudar de la validez de la hipótesis nula, no podemos rechazar esta afirmación. Sin embargo, no decimos que aceptamos H_0, la prueba no demuestra la hipótesis nula (podría ser que la muestra elegida es demasiado pequeña, o hay error de diseño, entre otras justificaciones).

Cuando se realiza el experimento para probar una hipótesis estadística, se debe tratar de minimizar las probabilidades de cometer error tipo I y error tipo II, para así obtener mayor validez en las conclusiones. Llamaremos α a la probabilidad de error tipo I y a β a la probabilidad de error tipo II.

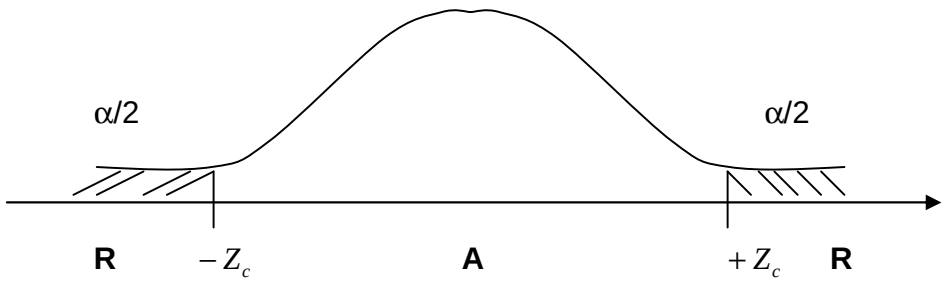
Al obtener el estadístico, éste se compara con el parámetro (valor H_0) para ver si es un valor cercano o no, es decir, si existe diferencia significativa o no con una cierta probabilidad dada de error.

Nivel de significación

La probabilidad máxima con la que en el ensayo de una hipótesis se puede cometer un error del tipo I, se llama nivel de significación del ensayo. El mismo es denotado por α. Comúnmente se usan los niveles 5% y 1%. Por ejemplo si es del 5%, se tiene que se está con un 95% de confianza de que se toma la decisión adecuada. En tal caso, se dice que la hipótesis ha sido rechazada al nivel de significación del 0.05, lo que significa que se puede cometer error con una probabilidad de 0.05.

Ensayos referentes a la distribución normal

Con una hipótesis dada, la distribución muestral de un estadístico S es una distribución normal con media y varianza específicos. La distribución de la variable tipificada es una normal tipificada. Un contraste de hipótesis es una función de decisión que lleva a aceptar o rechazar H_0.



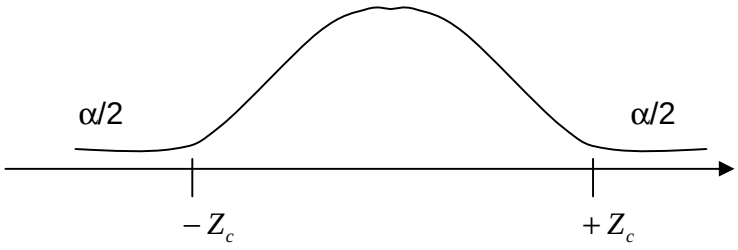
R = región crítico o de rechazo.
A = región de aceptación o no significación.
α = nivel de significación.

La regla de decisión:

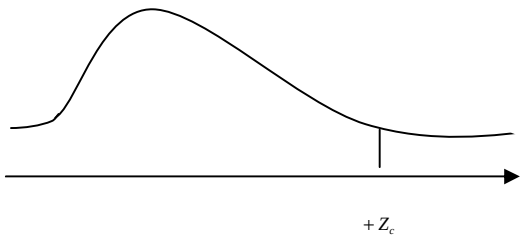
a) Se rechaza la hipótesis al nivel de significación de α sí el valor z_i obtenida para el estadístico S se encuentra fuera del rango $-Z_c$ a $+Z_c$.
b) Se acepta la hipótesis en caso contrario.

Ensayos de una cola y dos colas

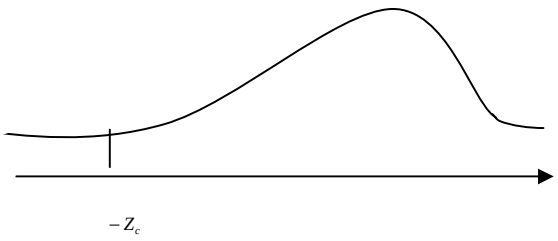
Si $H_i : \mu = \mu_0$, estas pruebas conducen a una prueba de 2 colas. Una prueba de dos colas es apropiada cuando el investigador no espera algo a priori respecto al valor a observar en la prueba. Solo desean saber si la muestra es diferente de la media de la población.



Si $H_1 : \mu > \mu_0$ o si $H_1 : \mu < \mu_0$, esta prueba conduce a una prueba de una cola, y es apropiada cuando el investigador tiene una idea a priori respecto al tamaño de la media.



$H_1 : \mu > \mu_0$



$H_1 : \mu < \mu_0$

Teoría de Muestras Grandes

Prueba de Hipótesis para la Media

	Región de rechazo
$H_0 : \overline{X} = \mu$	
$H_1 : \overline{X} \neq \mu$	$R : \lvert Z \rvert > Z_{\alpha/2}$
$H_1 : \overline{X} > \mu$	$R : Z > Z_{\alpha}$
$H_1 : \overline{X} < \mu$	$R : Z < -Z_{\alpha}$

donde el estadístico de la prueba para σ conocido es

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

donde μ es la media poblacional, $\bar{x}$ es la media muestral de la población y σ es la desviación típica de la población y n es el tamaño de la muestra.

Prueba de Hipótesis para la Proporción

	Región de rechazo
$H_0 : \hat{p} = p$	
$H_1^0 : \hat{p} \neq P$	$R : \lvert Z \rvert > Z_{\alpha/2}$
$H_1 : \hat{p} > p$	$R : Z > Z_{\alpha}$
$H_1 : \hat{p} < p$	$R : Z < -Z_{\alpha}$

y el estadístico de la prueba es:

$$Z = \frac{\hat{p} - P}{\sqrt{p.q/n}}$$

P es la proporción de éxitos de la población, $\hat{p}$ es la proporción de éxitos de la muestra y n el tamaño de la muestra.

Prueba de Hipótesis para la diferencia de las medias

Región de rechazo.

$$H_0 : \overline{X}_1 = \overline{X}_2$$

$$H_1 : \overline{X}_1 \neq \overline{X}_2 \qquad R : |Z| > Z_{\alpha/2}$$

$$H_1 : \overline{X}_1 > \overline{X}_2 \qquad R : Z > Z_{\alpha}$$

$$H_1 : \overline{X}_1 < \overline{X}_2 \qquad R : Z < -Z_{\alpha}$$

y donde el estadístico es para las varianzas conocidas

$$Z = \frac{\overline{X}1 - \overline{X}2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

donde σ_1, σ_2 desviaciones típicas de las muestras; $\overline{x}_1$, $\overline{x}_2$ son las medias muestrales de las poblaciones, n_1 y n_2 son los tamaños de las muestras 1 y 2 respectivamente.

Prueba de Hipótesis para la diferencia de las proporciones

Región de rechazo

$$H_0 : \hat{p}_1 = \hat{p}_2$$

$$R : |Z| \quad Z_{\alpha/2}$$

$$R : Z > Z_{\alpha}$$

$$R : Z < -Z_{\alpha}$$

$$H_1 : \hat{p}_1 \neq \hat{p}_2$$
$$H_1 : \hat{p}_1 > \hat{p}_2$$
$$H_1 : \hat{p}_1 < \hat{p}_2$$

donde el estadístico de la prueba es:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \qquad donde \; \bar{P} \; \frac{X_1 + X_2}{n_1 + n_2}$$

$\hat{p}_1$, $\hat{p}_2$ son las proporciones muestrales de éxito y $Q = 1 - P$, n_1 y n_2 son los tamaños de las muestras 1 y 2 respectivamente.

Etapas de las pruebas de hipótesis estadística

1. Determinación de H_0 y H_1.
2. Decisión sobre la prueba estadística apropiada.
3. Selección del nivel de significación para la prueba.
4. Determinación del valor que la prueba debe alcanzar para declararse significativa.
5. Cálculos.
6. Obtención de la conclusión.

❖ *Ejemplo 1 Se supone que el peso medio de todas las personas con sobrepeso, debido a problemas hormonales es de 90 Kg. Si la distribución del peso es normal con varianza igual a 100Kg^2. ¿Se podrá asegurar a un nivel de significación del 5% que el peso medio muestral de 89 Kg obtenido de una muestra aleatoria de 120 pacientes difiere significativamente del peso medio hipotético?*

Definimos primera la variable X = Peso de una persona que sufre de sobrepeso por proble mas hormonales.

Definimos las hipótesis.

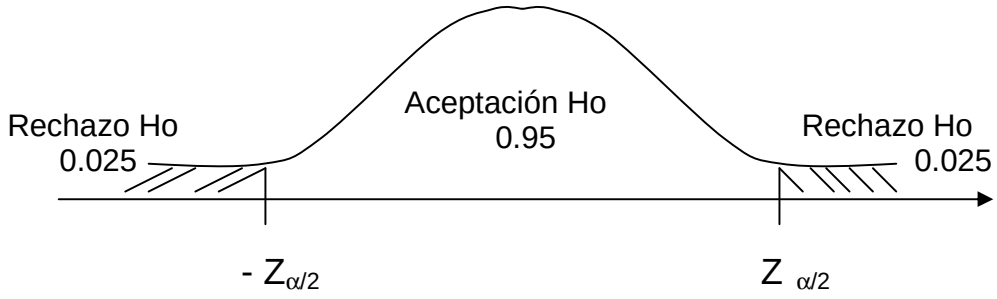
$H_0 : \mu = 90 Kgs$ El peso medio es de 90 Kgs.

$H_1 : \mu \neq 90 Kgs$ El peso medio difiere de los 90 Kgs.

El nivel de significación α = 5%. Esta es una prueba de dos colas, así $Z_\alpha = 1.96$

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{89 - 90}{10 / \sqrt{120}} = -1.09$$

Luego $-1.09 \in (-1.96, 1.96)$ por lo tanto decidimos aceptar con un nivel de significación del 5% que μ = 90 Kg



❖ *Ejemplo 2: Por lo general, una persona en nerviosa, tiene pulso promedio de 105 puls/min. Con una varianza de 25 puls/min. Un fabricante de medicamentos elabora un nuevo tranquilizante con la intención de reducir las puls/min. Para probar esta aseveración selecciona una muestra de 56 personas nerviosas; le suministra el nuevo medicamento y obtiene como resultado promedio 98 puls/min. ¿Es cierta la aseveración del fabricante? Pruébelo con un nivel del 1%*

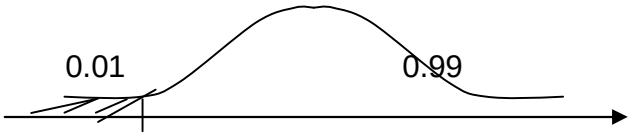
Datos:

$\mu = 105 \ puls/\min$

$\sigma^2 = 25 \ puls/\min$

$n = 56 \longrightarrow$ Normal

$\bar{x} = 98 \ puls/\min$



$\alpha = 0.01$ -z

La variable es X = Número de pulsaciones promedio en personas nerviosas. Las hipótesis asociadas son:

$H_0 : \bar{x} = \mu$ El nuevo tranquilizante no influye.
$H_1 : \bar{x} \neq \mu$ El nuevo tranquilizante si influye.

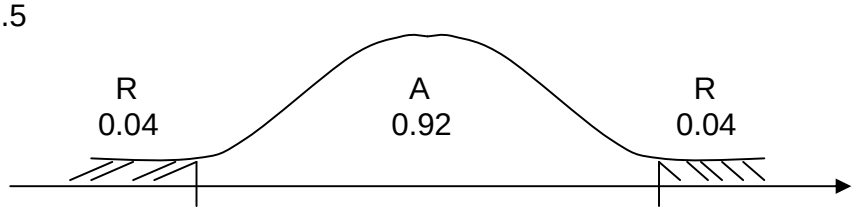
Hallamos el valor del estadístico:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{98\,puls/min - 105\,puls/min}{5/\sqrt{56}} = -2.09$$

El valor $Z_\alpha = -2.33$. El valor de Z entra en la región de aceptación, por lo que aceptamos Ho, es decir, no es cierta la aseveración del fabricante, porque luego de suministrar el nuevo tranquilizante, las personas no redujeron las pulsaciones por minuto, con un nivel de significación de 0.01.

- *Ejemplo 3: Se desea saber si el ayuno afecta los resultados de las hematologías completas. Para ello, se escogen dos muestras de personas normales: la primera muestra consta de personas que respetaron las 14 horas de ayuno antes de practicarse el examen; y la segunda muestra, por personas que hicieron sus comidas cotidianas. Después de realizar las hematologías, se obtuvieron los siguientes resultados:*
- en la primera muestra, de 80 personas, 50 de ellas tienen los valores dentro de los intervalos de referencia.
- en la segunda muestra, de 90 personas, 45 de ellas tienen los valores dentro de los intervalos de referencia.
¿Existe diferencia significativa entre las dos muestras? ¿ Influirá el ayuno en los resultados de las hematologías?. Pruébelo con un nivel del 8%

Datos:
$n_1 = 80$ ⟶ Normal
$p_1 = 50/80 = 0.625$
$n_2 = 90$ ⟶ Normal
$p_2 = 45/90 = 0.5$
$\alpha = 0.08$



-z z

Las hipótesis asociadas son

$H_0 : p_1 = p_2$ El ayuno no influye
$H_1 : p_1 \neq p_2$ El ayuno si influye

Hallamos el valor del estadístico:

$$Z = \frac{0.625 - 0.5}{\sqrt{\frac{0.55\ 0.45}{80} + \frac{0.55\ 0.45}{90}}} = 1.63$$

El valor $Z_{\alpha/2} = 1.75$. El valor de Z está en la región de aceptación, por lo que aceptamos con un nivel del 8% que el ayuno no influye en el resultado de las hematologías completas.

Teoría de pequeñas muestras

Distribución t de Student

Es una distribución de probabilidad continua y simétrica, pero más extendida que la normal y su amplitud depende del tamaño de la muestra; cuando ésta es muy grande coincide con la normal.

La función de esta distribución es:

$$Y = \frac{Y_0}{\left(1+\frac{t^2}{\upsilon}\right)^{\upsilon+1/2}}$$

Y_0 es una constante que depende de n y $\upsilon = n-1$ es el número de grados de libertad.

El número de grados de libertad de un estadístico se define como el número de observaciones independientes de la muestra.

Los valores del estadístico “t” vienen expresados en función del nivel de confianza y grados de libertad de la prueba.

De este modo, el intervalo de confianza para medias poblacionales viene dado por

$$\bar{x} \pm t_c \frac{S}{\sqrt{n-1}}$$

Prueba de Hipótesis para la media

El estadístico es

$$t = \frac{\bar{x} - \mu}{S}\sqrt{n-1}$$

con $\upsilon = n-1$ grados de libertad.

$$\bar{x} \pm t_c \frac{S}{\sqrt{n-1}}$$

Prueba de Hipótesis para la diferencia de medias

Dos muestras al azar de tamaños n_1 y n_2 con desviaciones típicas poblacionales iguales, donde las media $\bar{x}_1$ y $\bar{x}_2$ con desviaciones típicas S_1 y S_2, se tienen que

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde $\upsilon = n_1 + n_2 - 2$ y

$$\sigma = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

- *Ejemplo 4: Actualmente en el mercado, existen varios medicamentos que logran disminuir la temperatura en casos de fiebres muy altas en un tiempo promedio de 2 hr. Se quiere probar la eficacia de una nueva droga que produzca el mismo efecto en menor tiempo. Se escogió una muestra de 25 personas con fiebre alta y se les suministró el nuevo medicamento y se observó que en un tiempo promedio de 1 hr y 15 min. con una desviación de 7 minutos, se reduce la temperatura. ¿La nueva droga es más eficaz que los medicamentos actuales del mercado? Pruebe con un nivel de significación de 10%*

Datos:

$\mu = 2hr = 120\min$

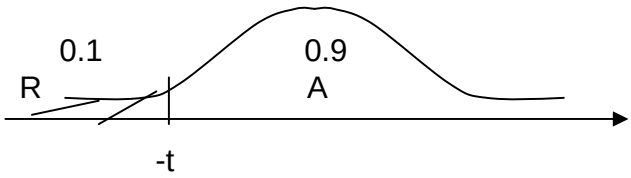
$S = 7$

$n = 25$ ⟶ T de Student

$\bar{x} = 75\min$

$\alpha = 0.1$

$\upsilon = 25 - 1 = 24$



Se define la variable X = tiempo promedio de disminución de la fiebre. Las hipótesis asociadas son:

$H_0 : \bar{x} = \mu$ El tiempo no disminuye y la eficacia es igual.

$H_1 : \bar{x} < \mu$ El tiempo disminuye y la eficacia de la droga es superior a los medicamentos actuales.

Hallamos el valor del estadístico:

$$t = \frac{\overline{X} - \mu}{S}\sqrt{n-1} = \frac{75\min - 120\min}{7\min}\sqrt{24} = -31.49$$

El valor $t_{\alpha/2,24} = -1.32$. En este caso el valor del estadístico t cae en la región de rechazo, por lo tanto con un nivel de significación del 10% rechazamos la hipótesis nula y aceptamos que el tiempo disminuye y la eficacia de la droga es mayor que los medicamentos actuales.

❖ *Ejemplo 5: Dos tipos de soluciones químicas A y B, fueron ensayadas para determinar su pH(grado de acidez de la solución). Análisis de 6 muestras de A dieron un pH medio de 7.56 con desviación típica de 0.24 y análisis de 5 muestras de B dieron un pH medio de 7.02 con desviación típica de 0.32. Mediante el valor α = 0.05 determinar si existe diferencia significativa entre las dos muestras*

Datos:

$n_A = 6$ ⟶ T de Student

$\bar{x}_A = 7.56$

$n_B = 5$ ⟶ T de Student

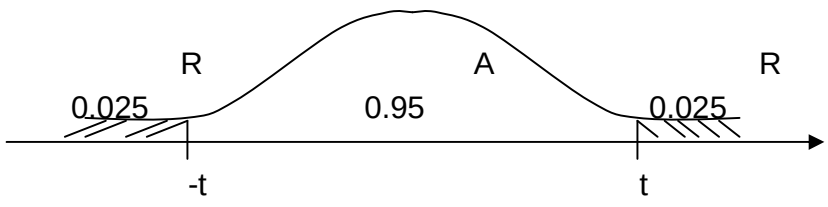
$\bar{x}_B = 7.02$

$X_B = 7.02$

$\alpha = 0.05$

$S_A = 0.24$

$S_B = 0.32$



Las hipótesis asociadas son :

$H_0 : \bar{x}_A = \bar{x}_B$ No existe diferencia significativa en la media de ambos grupos
$H_i : \bar{x}_A \neq \bar{x}_B$ Si existe diferencia significativa en la media de ambos grupos

Hallamos el valor del estadístico:

$$t = \frac{7.56 - 7.02}{\sqrt{\frac{6(0.24)^2 + 5(0.32)^2}{6+5-2}} \sqrt{\frac{1}{6} + \frac{1}{5}}} = 2.67$$

El valor $t_{6+5-2,1-\alpha/2} = t_{9,0.975} = 2.26$. El valor de t cae en la región de rechazo, por lo que aceptamos el hecho de que hay evidencia de que exista diferencia significativa entre las dos muestras.

Distribución Chi-Cuadrado

Es una distribución probabilística de tipo continua, con asimetría positiva y su función viene dada por la expresión

$$Y = Y_o \chi^{\upsilon - 2} e^{-\frac{1}{2}\chi^2}$$

siendo Y_0 un valor constante tomado en función de los grados de libertad y $\upsilon = n - 1$ es el número de grados de libertad.

Se puede obtener el intervalo de confianza para el estadístico χ^2, establecidos para los diferentes niveles de confianza, intervalos para estimar la desviación típica poblacional σ a partir del valor muestral S. Entonces debemos calcular

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

Prueba de Hipótesis para la varianza

Se utiliza couando se desea comparar la varianza poblacional σ^2 con la varianza muestral S^2. El estadístico es

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

el cual tiene asociados $\upsilon = n-1$ grados de libertad.

Los valores del estadístico χ^2 vienen expresados en función del nivel de confianza y de los grados de libertad.

❖ *Ejemplo 6: En el pasado, la desviación típica de los pesos de ciertos paquetes de 40 onzas llenados por una máquina era de 0.25 onzas. Una muestra al azar de 20 paquetes dio una desviación típica de 0.32 onzas. ¿Es significativo el incremento de variabilidad? Use un nivel del 5%*

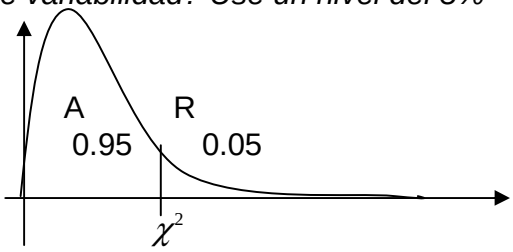
Datos:

$S = 0.32$ *onzas*

$n = 20$

$\sigma = 0.25$ *onzas*

$\alpha = 0.05$



La hipótesis asociada es:

$H_0 : \sigma^2 = S^2$ No hay variabilidad de los datos.

$H_1 : \sigma^2 < S^2$ Es significativo el incremento de la varianza.

Hallamos el valor del estadístico:

$$\chi^2 = \frac{20 \quad (0.32)^2}{(0.25)^2} = 32.768$$

En la tabla encontramos el siguiente valor:

$$\chi^2_{0.95,19} = 30.14$$

Así, rechazamos el hecho de que las varianzas sean iguales con un nivel del 5%, es decir, se acepta que fue significativo el incremento.

CAPÍTULO V

Regresión y Correlación

En esta parte del curso, se analiza el concepto de relación entre dos variables y extiende esta idea para predecir el valor de una variable a partir de la otra. Se describen pruebas estadísticas para determinar si una relación entre dos variable es significativa o no.

Anteriormente, las muestras consistían en mediciones de una sola variable aleatoria Y. Ahora si queremos estudiar dos o más variables de una misma población, entonces se deben aplicar las técnicas de regresión y correlación.

De una población de tamaño N se pueden estudiar dos variables X e Y, los cuales los tendremos por pares de observaciones $(x_i y_i)$ los cuales los disponemos en forma de tablas. Dichas tablas expresan cómo se distribuyen las observaciones en función de los pares $(x_i y_i)$ por lo que reciben el nombre de distribución bidimensional de frecuencias. En estos casos se pretende estudiar de este conjunto, la relación existente entre las variables.

Cada variable, separadamente, se estudia básicamente a través de su media aritmética y varianza. Por lo que encontramos $\overline{X}, \overline{Y}, S_{\overline{X}}, S_{\overline{Y}}$ respectivamente. Este análisis se realiza mediante el cálculo de una medida llamada covarianza cuya fórmula es:

$$S_{xy} = \frac{\sum_{i=1}^{N} x_i\, y_i}{N} - \frac{\sum_{i=1}^{N} x_i}{N}\,\frac{\sum_{i=1}^{N} y_i}{N}$$

Si $S_{XY} > 0$ entonces decimos que X e Y siguen el mismo comportamiento.
Si $S_{xy} < 0$ entonces decimos que X e Y se mueven en sentido contrario.
Si $S_{XY} = 0$ entonces no existe relación entre las variables.

Más que la covarianza, existe otra magnitud que indica el grado de relación entre las dos variable, es llamada *Coeficiente de correlación* r que viene dada por la fórmula:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}}$$

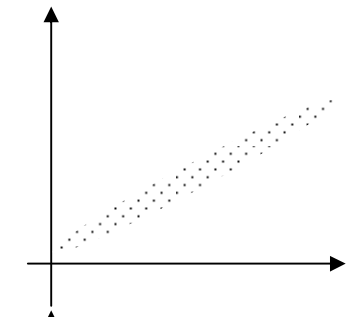
Dicho valor oscila entre –1 y 1 y se interpreta como sigue:

- Si r $\rightarrow$ 1 las variables tienen igual sentido.
- Si r $\rightarrow$ -1 las variables tienen relación inversa.
- Si r $\rightarrow$ 0 no hay relación lineal entre las variables

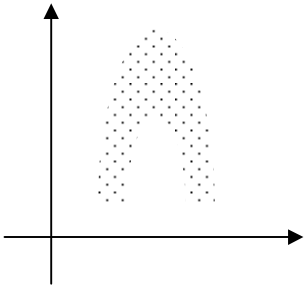
De esta forma la *Correlación* sirve para obtener una medida del grado de fuerza o relación que existe entre dos variables.

Diagrama de dispersión

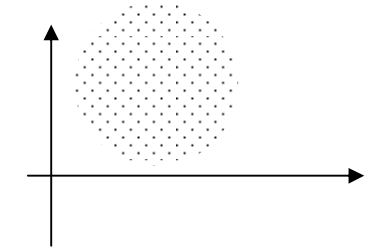
El diagrama de dispersión da una visualización y un medio más simple para estudiar la relación entre dos variables. En este diagrama, cada uno de los n pares de observaciones $(x_i y_i)$ se marca con un solo punto en la gráfica. Con la disposición de los puntos en la gráfica se detecta el patrón indicativo de la naturaleza de la forma funcional básica de los datos.



Los puntos sugieren una relación lineal



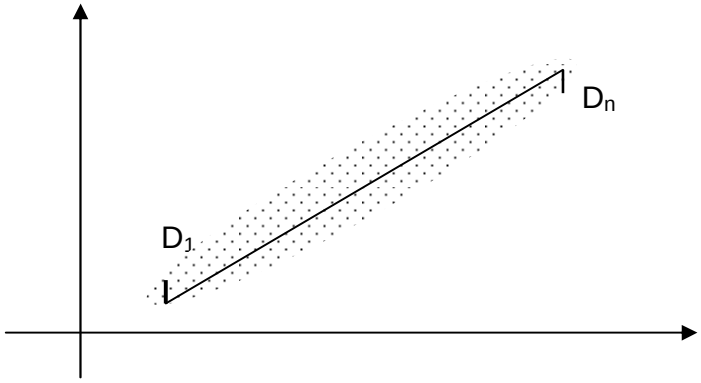
Los puntos sugieren una relación curvilínea.



Los puntos no sugieren relación alguna.

Modelo de Mínimos Cuadrados

Supongamos que un diagrama de dispersión consta de los puntos $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Para evitar la subjetividad del experimento en la elección de la mejor curva que se ajusta al diagrama de dispersión, es necesario tomar un modelo que impida este error. Para un valor dado de x por ejemplo, existirá una diferencia x_1 y el correspondiente valor de la curva c; esta diferencia se indica en la gráfica por D_1 que se conoce como desviación, error o residuo, el cual puede ser positivo, negativo o cero. Para los otros puntos también encontramos distancias $D_1, D_2, ..., D_n$. Una medida de la bondad de ajuste de la curva *c* a los datos viene suministrada por la cantidad $D_1^2 + D_2^2 + \cdots + D_n^2$. Si este valor es pequeño, entonces el ajuste es bueno, si es grande, el ajuste es malo.



De todas las curvas de aproximación a una serie de datos puntuales, la curva que tiene la propiedad de que $D_1^2 + D_2^2 + \cdots + D_n^2$ es mínima, se conoce como la mejor curva de ajuste.

Una curva que presente esta propiedad se dice que se ajusta a los datos por mínimos cuadrados y se llama curva de mínimos cuadrados.

Regresión

Consiste en obtener una ecuación que se pueda usar para predecir o calcular el valor de una variable correspondiente a un valor dado de la otra variable. Existen muchos modelos de regresión, a saber, lineal, cuadrático, logarítmico, logístico, exponencial, entre otros. En este curso, solo estudiaremos el modelo de regresión lineal

Regresión Lineal

En este caso, el diagrama de dispersión sugiere la idea de intentar expresar la relación entre las dos variables mediante una línea de regresión que sea recta. Si tenemos dos variables X e Y, decimos que están relacionadas según una línea recta cuando sus valores satisfacen la ecuación

$$Y = aX + b$$

donde a y b son constantes. La constante a se refiere a la inclinación de la recta y b es el valor por donde la recta corta al eje vertical y vienen expresados por las fórmulas:

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

y

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

Para determinar matemáticamente la ecuación de esta recta de regresión aplicando el proceso de mínimos cuadrados donde hay que lograr que

$$\sum D_i^2 = \sum \left(y_i - \left(y_{est}\right)_i\right)^2 \ sea \ \text{mín}\, ima.$$

En esta fórmula $(y_{est})_i$ es el valor de la ordenada de la recta de regresión para un valor $x = x_i$.

A su vez, podemos hallar el error de estimación al momento indicar cual es la recta de regresión lineal, el cual viene dado por la fórmula:

$$S = \sqrt{\frac{\sum\left(y - y_{est}\right)}{n}}$$

- *Ejemplo 1: Clínicamente se ha determinado que existe relación entre el peso y los niveles de glucosa en la sangre en personas que sufren de diabetes. Se quiere analizar el tipo de relación entre estas dos variables y para ello se selecciona un grupo de 14 diabéticos y se registraron los siguientes datos:*

Peso(Kg)	*58*	*69*	*75*	*67*	*71*	*59*	*72*	*78*	*77*	*70*	*68*	*65*	*80*	*76*
Glucosa(mg%)	*168*	*192*	*199*	*178*	*197*	*165*	*198*	*198*	*199*	*198*	*190*	*175*	*210*	*197*

Se desea saber:

a) ¿Qué tipo de relación existe entre el peso y los niveles de glucosa?Justifique su respuesta.
b) Estime el nivel de glucosa que puede tener un diabético que peso 60 Kg.

a) Construimos la tabla:

	Peso(X)	**Glucosa(Y)**	X^2	XY	Y^2
	58	168	3364	9744	28224
	69	192	4761	13248	36864
	75	199	5625	14925	39601
	67	178	4489	11926	31684
	71	197	5041	13987	38809
	59	165	3481	9735	27225
	72	198	5184	14256	39204
	78	198	6084	15444	39204
	77	199	5929	15323	39601
	70	198	4900	13860	39204
	68	190	4624	12920	36100
	62	175	4225	11375	30625
	80	210	6400	16800	44100
	76	197	5776	14972	38809
Total	985	2664	69883	188515	509254

Buscamos el coeficiente de correlación lineal:

$$r = \frac{14\,(188515) - 985\;(2664)}{\sqrt{\left[\left[14(69883) - (985)^2\right]\left[14(509254) - (2664)^2\right]\right]}} = 0.93$$

Existe correlación lineal positiva entre los datos ya que la asociación entre las variables es alta. Por lo tanto, si el peso de la persona aumenta, entonces el nivel de glucosa también aumenta y viceversa.

b) Buscamos los valores de a y b:

$$a = \frac{14(188515) - 985(2664)}{14(69883) - (985)^2} = 1.86$$

$$b = \frac{2664(69883) - 188515(985)}{14(69883) - (985)^2} = 59.11$$

De este modo la recta de regresión lineal que se ajusta a los datos es

$$Y = 1.86X + 59.11$$

Así, para un individuo que pesa 60Kgs, se estima que su nivel de glucosa será

$$Y = 1.86(60) + 59.11 = 170{,}7 \Rightarrow 171mg\%$$

CAPITULO VI

Ejercicios de estadística descriptiva

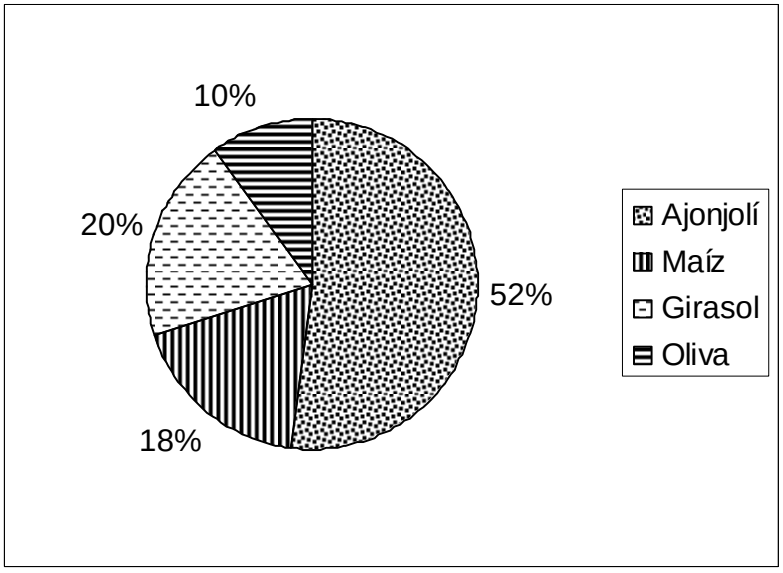
1.- Se realiza un estudio en 100 hogares de clase media en la Ciudad de Mérida para conocer el tipo de aceite usado en la cocina. Los resultados son:

Tipo de Aceite	Nro. de hogares
Oliva	7
Ajonjolí	21
Maíz	58
Girasol	14

Responda las siguientes preguntas:

a) ¿Cuál es la población?
b) ¿Cuál es la muestra?
c) ¿Cuál es la variable y de qué tipo es?
d) Grafique el diagrama circular.

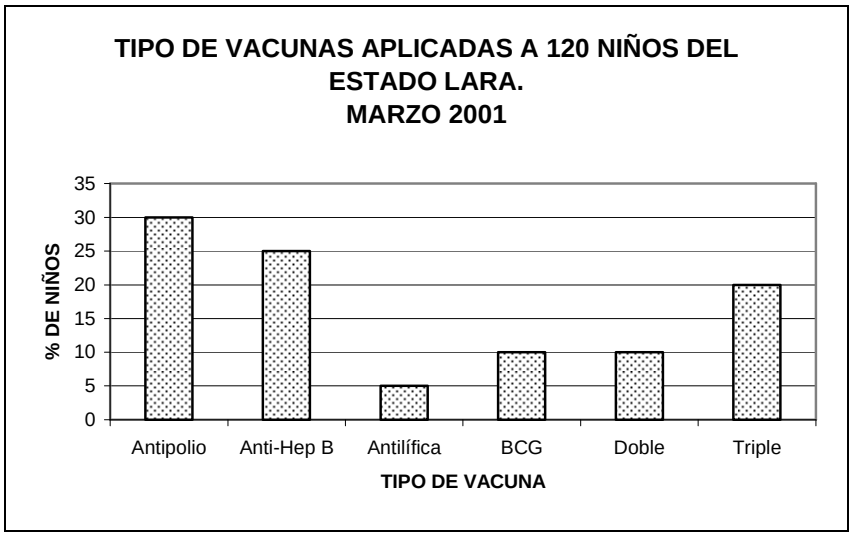
2.- Se ha realizado un estudio en 120 hogares de la clase media en la Ciudad de Maturin para conocer el tipo de aceite usado en la cocina. Los resultados de las encuestas fueron los siguientes:



Responda a las siguientes preguntas, justificando su respuesta:

a) ¿Cuál es la población?
b) ¿Cuál es la muestra?
c) ¿Cuál es la variable y de qué tipo es?
d) Realice la distribución de frecuencias.

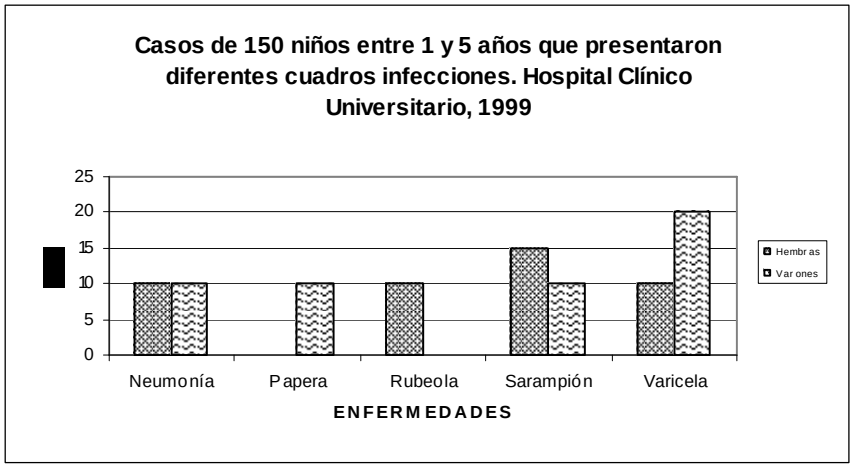
3.- Se desea estudiar el tipo de vacunación practicada en el Estado Lara. Para ello, se seleccionaron, al azar, 120 niños donde se recogieron los siguientes resultados:



Responda las siguientes preguntas, justificando su respuesta:

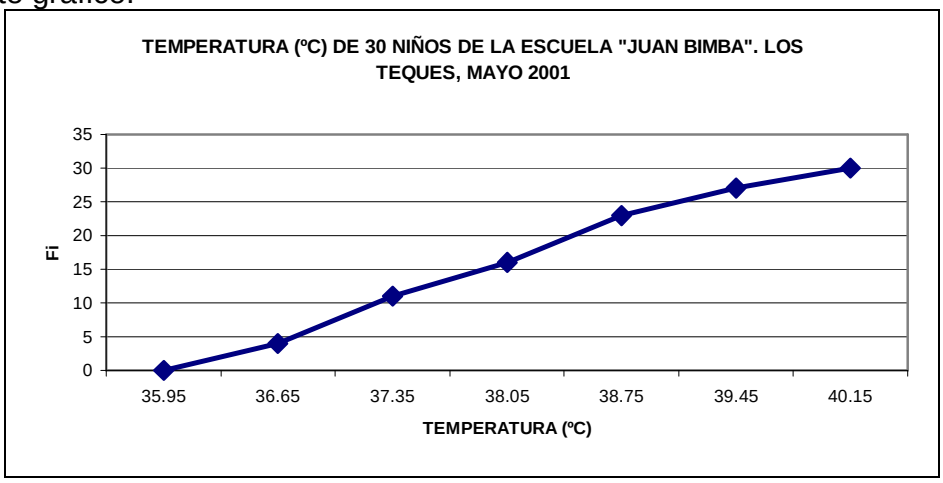
a) ¿Cuál es la población y cuál es la muestra?
b) ¿Cuál es la variable en estudio y de qué tipo es?
c) De las dos formas (tabulas y gráfica). ¿cuál escogería usted para presentar los datos?

4.- El siguiente gráfico representa casos de 150 niños, con edades comprendidas entre 1 y 5 años, que presentaron diferentes cuadros infecciosos que asistieron al Hospital Clínico Universitario en el año 1999.



Observando el gráfico:
a)¿Cuáles son las variables y de qué tipo son? b)¿Qué enfermedad en más frecuente en las hembras? ¿y en los varones? c)¿A qué enfermedad son más propensos los varones?

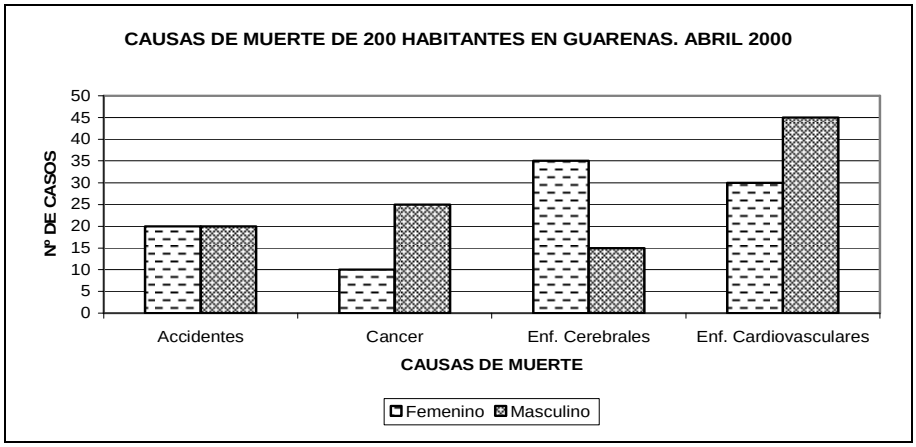
5.- En la Escuela “Juan Bimba” de los Teques, se tomaron al azar 30 niños para medirles la temperatura y los resultados fueron representados mediante el siguiente gráfico:



Se pide:

a) ¿Cuál es la variable y de qué tipo es?
b) ¿Cuál es la población y cuál es la muestra?
c) Construya la tabla de frecuencias.
d) Si a la enfermera le quedan 5 aspirinas y decide darlas a los niños que tienen las más altas temperaturas, ¿cuál debe ser la temperatura mínima que debe tener un niño para que se le suministre una pastilla?

6.- En Guanare, cada año mueren 1000 habitantes por diferentes razones. Un investigador desea estudiar las causas de estas muertes por sexo, por ello toma una muestra al azar de 200 defunciones del mes de abril 2000 y presenta los siguientes resultados:



Realice las tablas de frecuencia (por sexo) y conteste:

a) % de personas que murieron por enfermedades del corazón.
b) ¿Cuál es la población y cuál es la muestra?
c) ¿Cuál es la variable en estudio y de qué tipo es?
d) ¿Cuál es la causa de muerte más frecuente en los hombres? ¿y en las mujeres?

7.- La siguiente distribución pertenece a las notas obtenidas por un grupo de estudiantes regulares de Matemática I en el quizº #3 (ecuaciones) y quiz #4 (factorización) efectuado en el semestre SEG-00.

Calificación (puntos)	Quiz 3	Quiz 4
01 – 05	6	41
06 – 10	9	13
11 – 15	17	2
16 - 20	24	0

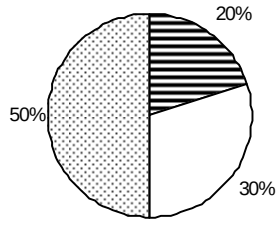
a) Construya la distribución de frecuencias
b) Comparar dichas calificaciones, con respecto a media, desviación y variación.
c) Si un estudiante obtuvo una calificación de 10 puntos, ¿en cuál de las dos evaluaciones es mejor estudiante?
d) Realice el Polígono de Frecuencias.

RESPUESTAS:

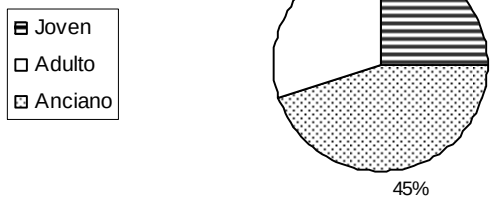
PREGUNTA	Quiz 3	Quiz 4
7. b)	Media= 13.2ptos	Media=4.5ptos
	Desviación=5.0ptos	Desviación=2.3ptos
	%CV=38.5%	%CV=40.0%
7.c)	K=26.7	K=96.1

8.- Los siguientes gráficos representan el porcentaje de personas hipertensas en función de la edad, separadas por sexo (35 femeninos y 45 masculinos), que asistieron a cuidados coronarios del Hospital José Gregorio Hernández:

% de mujeres hipertensas distribuidos por grupo etario que acudieron a cuidados coronarios del Hospital José Gregorio Hernández



% de hombres hipertensos distribuidos por grupo etario que acudieron a cuidados coronarios del Hospital José Gregorio Hernández



a) ¿Cuáles son las variables tratadas y de qué tipo son?
b) ¿Cuál es la población y cuál es la muestra?
c) Represente en una tabla de frecuencia los datos aportados por ambos gráficos.

9.- Se realizó un estudio para analizar los valores de colinesterasa en un recuento de eritrocitos (mol/min/ml) entre 35 trabajadores agrícolas expuestos a pesticidas:

10.6	9.9	12.6	15.2	12.3	11.7	12.3
12.5	11.8	12.4	10.2	11.3	9.4	11.4
11.0	11.6	12.2	13.4	9.9	9.8	10.2
9.2	15.3	10.9	9.0	11.0	8.6	12.5
11.6	12.6	16.7	7.7	10.9	10.1	8.7

Represente los datos en Forma tabular (datos agrupados) y gráfica (histograma de frecuencia)

10.- Se quiere probar la calidad de un jabón en cuanto a su duración; para ello, se encuestaron a 100 amas de casa y se obtuvieron los siguientes resultados

DURACIÓN (días)	fi
5 9	24
10 14	70
15 19	88
20 24	100

Los fabricantes del jabón tomaron el siguiente criterio: “Si más del 50% de las amas de casa afirman que el jabón dura entre 13 y 17 días, entonces se lanza el producto al mercado, de lo contrario se revisará su fabricación para mejorarla”. Estadísticamente hablando, ¿se podrá lanzar el producto al mercado? Justifique su respuesta.

RESPUESTAS:

$K_{(13DIAS)}$= 56.2%

$K_{(17DIAS)}$=79.0%

%AMAS DE CASA QUE AFIRMAN QUE EL JABON DURA ENTRE 13 Y 17 DIAS= 20.5%

$K_{(17DIAS)}$ - $K_{(13DIAS)}$ = 79.0%- 58.5% =22.8%

CONCLUSIÓN: NO SE LANZA EL PRODUCTO AL MERCADO, ES DECIR SE REVISARA SU FABRICACIÓN PARA MEJORARLA.

11.- La siguiente tabla se refiere a las estaturas de 50 estudiantes:

Li	Ls	fi
1.45	1.48	2
1.48	1.51	7
1.51	1.54	4
1.54	1.57	3
1.57	1.60	12
1.60	1.63	9
1.63	1.66	4
1.66	1.69	4
1.69	1.72	2
1.72	1.75	3

Calcule:

a) % de alumnos con estaturas igual o mayor a 1.60m, pero menor que 1.69m
b) ¿Cuántos alumnos miden menos de 1.52m?
c) ¿Cuántos alumnos miden más de 1.68m?
d) % de alumnos con estaturas comprendidas entre 1.67m y 1.73m
e) Estatura máxima del 20% de menor estatura
f) Estatura mínima del 15% de mayor estatura

RESPUESTAS:

PREGUNTA	RESPUESTA
11.a)	34 %
11.b)	11 ALUMNOS
11.c)	6 ALUMNOS
11.d)	11%
11.e)	1.52 m
11.f)	1.67 m

12.- Se seleccionó un grupo de 28 varones para analizar la creatinina (en mg%), tomada en muestras de orina de 24 horas; éstos fueron los resultados:

1.51	1.65	2.03	1.46	1.89	1.52	1.80
1.60	1.46	1.55	1.71	1.22	1.33	1.86
1.90	1.52	1.38	1.66	1.26	1.75	1.59
1.37	1.51	1.71	1.57	1.49	1.25	1.45

Se pide:

a) Realice la tabla de frecuencias con datos agrupados.
b) Si se considera que los varones que tienen sus valores de creatinina comprendidos entre 1.40mg% y 1.75mg% son normales, ¿cuántos hay en este grupo? **RESPUESTA**: 16 personas
c) Realice el polígono de frecuencia.

13.- Se quiere saber el tiempo (min) que emplea un estudiante ucevista, que habita en Caracas, en trasladarse desde su casa hasta la Universidad. Se escoge una muestra de 32 estudiantes de la Facultad de Medicina, que inician su horario de clase a las 8:00 a.m.; éstos fueron los resultados:

44	36	15	29	35	12	30	56
45	42	20	37	31	44	30	45
52	41	53	50	58	24	43	39
30	39	28	37	29	39	60	40

Se pide:

a) ¿Cuál es la población y cuál es la muestra?
b) ¿La muestra escogida es representativa de la población en estudio?
c) Elabore la tabla de datos agrupados con una amplitud igual a 8
d) Grafique el histograma de frecuencia
e) Si se desea saber las siguientes informaciones, ¿cuál método de representación escogería?
    - ¿Qué intervalo de tiempo emplea con más frecuencia los estudiantes, para su traslado?
    - ¿Qué % de estudiantes tardan en el trayecto un tiempo menor o igual a 36min?

14.- En una medición del colesterol (mg/dL) en el suero sanguíneo se han obtenido estos valores:

230	235	200	190	120	145	175	170	290	220
225	215	181	245	150	195	200	230	240	200
278	230	175	265	210	250	210	215	190	270

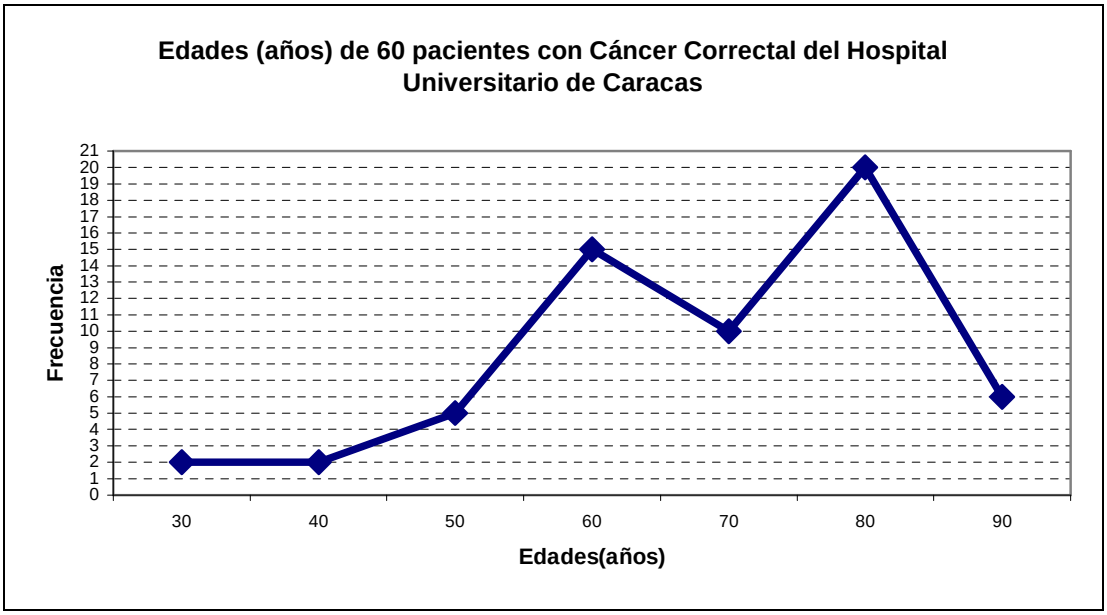
Tomando una amplitud de 30:

a) Construya la tabla de frecuencias
b) Dibuje el histograma de frecuencia y la ojiva correspondiente

15.- Se tomó una muestra aleatoria de 15 parejas de estudiantes de la Escuela de Karate para evaluar los latidos del corazón (pulsaciones/min) después de 10 minutos de intensa actividad física y se obtuvieron los siguientes resultados:

82	95	92	62	85	92	Se pide:
82	95	70	85	84	95	a) Construya la distribución de datos
91	82	94	76	88	91	agrupados
87	80	68	58	76	85	b) Grafique el polígono de frecuencia y la
110	60	75	88	67	74	ojiva

16.- Se realizó un estudio para evaluar las edades de pacientes con cáncer correctal en Venezuela. Se tomaron al azar 60 pacientes que asistieron al Hospital Universitario de Caracas. Los resultados fueron presentados mediante el siguiente gráfico:



Se pide:

a) ¿Cuál es la variable y de qué tipo es?
b) ¿Cuál es la población y cuál es la muestra?
c) Construya la tabla de distribución de frecuencias

17.- En un experimento para determinar el efecto de una droga en particular en el nivel de colesterol del suero (mg/dL) en adultos varones de 30 años de edad. Se registraron los siguientes valores en el grupo que se trató con droga:

230	235	200	175	170	289	181	245	150	190
120	145	220	225	215	195	200	230	165	265
210	250	210	215	190	270	250	158	270	268

a) Realizar una tabla distribución de frecuencia para datos agrupados con una amplitud de 20 para cada intervalo
b) ¿Cuántos adultos presentaron un nivel de colesterol superior a 140mg/dL? **RESPUESTA:** 29 Adultos
c) Qué % presenta un nivel de 210mg/dL o más? **RESPUESTA:** 52%
d) Grafique el histograma y la ojiva porcentual correspondiente

18.- La siguiente tabla se refiere a las estaturas (metros) de 80 estudiantes de la Escuela de Bioanálisis de la Universidad Central de Venezuela:

Li	Ls	fi
1.495	1.535	3
1.535	1.575	5
1.575	1.615	10
1.615	1.655	15
1.655	1.695	25
1.695	1.735	12
1.735	1.775	8
1.775	1.815	2

Hallar:

a) % de alumnos con estatura superior a 1.735m
b) La estatura mínima del 20% de los alumnos de mayor estatura
c) Coeficiente de Variación

RESPUESTAS:

PREGUNTA	RESPUESTA
18.a)	13%
18.b)	1.715 m
18.c)	4.33%

19.- A continuación, se indica la distribución de las puntuaciones obtenidas por 50 estudiantes en una prueba de admisión para ingresar a la Escuela de Bioanálisis:

50.3	61.0	77.0	85.9	57.8	75.2	85.3	65.1	74.2	77.3
71.2	75.1	52.1	77.0	74.2	88.1	63.1	89.3	86.3	64.3
82.5	67.2	63.4	91.1	53.1	78.5	99.3	74.2	96.1	87.3
68.4	71.3	83.7	78.3	84.2	95.3	76.4	57.5	78.1	72.4
74.2	54.1	74.8	84.2	55.0	76.2	97.2	77.3	76.4	66.5

Los criterios tomados para la admisión de los estudiantes fueron:

- Puntuación menor a 65,0, no se admiten
- Puntuación entre 65,0 y 78,0 (ambos inclusive), deben realizar un curso propedéutico
- Puntuación superior a 78,0, son admitidos
- Puntuación igual o superior a 90,0, son admitidos con beca

Se pide:

a) % de alumnos que fueron admitidos pero no ganaron la beca. **RESPUESTA:** 26%
b) Cantidad de alumnos que tienen que hacer el curso propedéutico. **RESPUESTA:** 20 alumnos
c) Calcular media, desviación, curtosis y sesgo sin agrupar datos.
RESPUESTAS:
Media = 74.9ptos
desviación = 12.3 ptos
curtosis = 0.218 = DISTRIBUCIÓN PLATICURTICA
sesgo = - 0,1057 = (ASIMETRÍA NEGATIVA)

20- Se tienen datos relativos a los ingresos y egresos diarios de una muestra de 20 padres de familia distribuida así:

INGRESO	fi	EGRESOS	fi
3000	2	2000	3
4000	4	3000	5
5000	7	4000	6
6000	5	5000	5
7000	2	6000	1

Realice la tabla correspondiente y responda las siguientes preguntas:

a) ¿Cuál es el ingreso que divide a la distribución en dos partes iguales?
b) ¿Cuál es el egreso más frecuente?
c) ¿Cuál es el valor del ahorro promedio por familia?

RESPUESTAS:

PREGUNTA	RESPUESTA
20.a)	Bs 5000
20.b)	Bs 4000
20.c)	Bs 1250

21.- Una persona quiere saber el sueldo promedio de 500 empleados de cierta compañía. Para ello decide encuestar a 15 trabajadores, escogidos al azar y obtiene los siguientes resultados (en Bs.)

940.000	104.000	92.000
85.000	75.000	1.090.000
108.000	85.000	98.000
1.200.000	85.000	85.000
90.000	1.000.000	999.000

Conteste:

a) ¿Cuál es la población y la muestra?
b) La muestra escogida,¿es representativa de la población?
c) Obtenga las tres medidas de tendencia central (sin agrupar los datos), ¿cuál de ellas representaría más el sueldo promedio de los empleados de esta compañía?

RESPUESTAS:

Media = Bs 409067
Mediana = Bs 98000
Moda = Bs 85000 (Distribución Unimodal)

22.- Un grupo de individuos padecieron de una inexplicable intoxicación con vitamina D (lo cual requirió de su hospitalización); se pensó que podría deberse a una dieta con ingestión excesiva de productos lácteos. Para estudiar esto, se tomaron dos grupos de personas: grupo A, constituido por 18 personas que padecieron de dicha intoxicación y el grupo B, por 20 individuos saludables. A los mismos se les realizó una prueba de laboratorio con la intención de determinar sus niveles de calcio (mmol/l)

Calcio (mmol/L) Li - Ls	Sanos	Enfermos
1.40 – 1.80	4	3
1.80 – 2.20	8	5
2.20 – 2.60	3	6
2.60 – 3.00	2	2
3.00 – 3.40	1	4

Si los valores de referencia en el nivel de calcio está entre 2.12 y 2.74 mmol/l:

a) Indique el número de individuos con calcio entre los valores de referencia para el grupo de los sanos
b) Para un individuo con un nivel de calcio de 2.50mmmol/l, ¿en cuál de los dos grupos se considera tiene un nivel más alto? Use percentiles
c) Indique límites del 50% central de los datos para el grupo de los enfermos

RESPUESTAS:

PREGUNTA	RESPUESTA
22.a)	5 INDIVIDUOS
22.b)	ES MAS ALTO EN EL GRUPO DE LOS SANOS
22.c)	LIMITE INFERIOR: 1.96 mmmol/l LIMITE SUPERIOR: 2.80 mmmol/l

23.- Se llevó a cabo un estudio en el que se compararon mujeres adolescentes que padecían Bulimia y mujeres Sanas con las mismas características corporales y niveles de actividad física. El siguiente cuadro muestra las medidas del consumo diario de calorías en kilocalorías por kilogramo:

Consumo de Calorías diario (Kc/Kg) Li - Ls	Mujeres Bulímicas	Mujeres Saludables
15 – 18	7	3
19 – 22	11	4
23 – 26	8	9
27 – 30	4	10
31 – 34	2	12

a) Calcule la media, moda y desviación típica para cada uno de los grupos
b) Calcule el nivel de asimetría y homogeneidad de cada grupo
c) ¿Qué puede concluir?

RESPUESTAS:

PREGUNTA	RESPUESTAS	
	MUJERES BULIMICAS	MUJERES SALUDABLES
23.a)	MEDIA = 22 Kc/Kg	MEDIA = 27 Kc/Kg
	MODA = 21 Kc/Kg	MODA = 31 Kc/Kg
	S = 5 Kc/Kg	S = 5 Kc/Kg
23.b)	SESGO = 0.2	SESGO = -0.8
	%CV = 23%	%CV = 19%

24.- Las siguientes distribuciones indican los sueldos diarios de 80 bioanalistas del laboratorio A y 90 bioanalistas del laboratorio B:

Xi	fi(A)	fi(B)
20000	8	5
30000	10	12
40000	23	28
50000	18	46
60000	12	71
70000	7	84
80000	2	90

Responda a las siguientes preguntas justificando la respuesta:

a) ¿Cuál de las dos distribuciones ofrece mayor variabilidad?
b) Un bioanalista que gane Bs. 50000 diarios, ¿en qué laboratorio debe considerarse peor pagado?
c) Halle la media aritmética de los 170 bioanalistas
d) ¿Cuál de las dos distribuciones se acerca más a la normal desde el punto de vista de la asimetría?

RESPUESTAS:

PREGUNTA	RESPUESTA
24.a)	La distribución de los sueldos del Laboratorio A ($\%CV_A > \%CV_B$)
24.b)	Se considera peor pagado en el Laboratorio B
24.c)	Media = Bs 49353
24.d)	La distribución de sueldo sen el Laboratorio A se acerca más a la distribución Normal (|Sesgo A|<|Sesgo B|)

25.- Se desea practicar un examen de sangre a dos grupos de personas para comparar sus niveles de colesterol (mg/dl). El grupo A está constituido por 25 vegetarianos y el grupo B por personas que consumen carnes. Una vez hecho el análisis, se presentaron los resultados:

Colesterol (mg/dL)		fi_A	fi_B
Li	Ls		
30	45	4	0
45	60	10	4
60	75	7	8
75	90	3	12
90	105	1	6

a) Realice las tablas respectivas
b) Hallar media, mediana y moda de cada distribución y compare los dos grupos ¿qué concluye?
c) Si una persona tiene un nivel de 70mg/dl, ¿en cuál de los dos grupos se consideraría con un nivel alto de colesterol? Use percentiles
d) Realice en un gráfico, el polígono de frecuencias.

RESPUESTAS:

PREGUNTA	RESPUESTAS	
	GRUPO A	GRUPO B
25.b)	MEDIA = 60 mg/dL	MEDIA = 78 mg/dL
	MEDIANA = 58 mg/Dl	MEDIANA = 79 mg/dL
	MODA = 55 mg/dL	MODA = 81 mg/dL
25.c)	En el grupo de los Vegetarianos el valor de 70 mg/dL se considera más alto.	

26.- Se desea comprobar la influencia del tabaquismo para producir Bajo Peso al Nacer (BPN). Se escogieron dos muestras de neonatos. Una de ellas (A) cuyas madres son fumadoras y la segunda (B) de madres no fumadoras. Para comprobar el BPN se tomó como indicador el peso (Kg) de los niños al nacer, teóricamente se considera un niño sano si su peso es superior a 2.50 Kgs y si su peso es menor o igual a 2.50 Kgs se considera con BPN. Los resultados se muestran en la tabla anexa.

Peso al nacer (Kg)		fiA	fiB
Li	Ls		
1.00	1.35	8	0
1.36	1.71	14	0
1.72	2.07	11	2
2.08	2.43	7	5
2.44	2.79	3	8
2.80	3.15	2	15
3.16	3.51	1	10
3.52	3.87	0	9

Se pide:

a) Construir la tabla de frecuencias para ambas muestras
b) Obtener el % de niños con BPN en cada muestra
c) Obtener el Coeficiente de Variación, el sesgo y la curtosis de cada muestra
d) Graficar el polígono de frecuencias, que refleje los datos de las dos muestras
e) En función de los resultados anteriores, ¿se podría afirmar que existe relación entre el tabaquismo y el BPN? Justifique

RESPUESTAS:

PREGUNTA	RESPUESTAS	
	FUMADORAS	NO FUMADORAS
26.b)	88.1%	17.2%
26.c)	CV = 27%	CV = 16.4%
	SESGO= 0.72	SESGO= -0.04
	CURTOSIS= PLATICURTICA	CURTOSIS= PLATICURTICA

27.- La siguiente distribución expresa el tiempo (min) que tardaron un grupo A de 90 estudiantes de la Facultad de Medicina y un grupo B de 130 estudiantes de la Facultad de Ingeniería en contestar una prueba de habilidades numéricas. Estos fueron los resultados:

Li	Ls	fiA	fiB
30	33	5	15
34	37	8	25
38	41	12	40
42	45	30	20
46	49	18	10
50	53	12	12
54	57	5	8

Se pide:

a) Determine la media aritmética, la mediana y la moda de las dos distribuciones
b) Interprete los resultados anteriores y compare las dos distribuciones
c) Determinar la media aritmética general del grupo de 220 estudiantes
d) Un estudiante que haya tardado 40 minutos, ¿en qué grupo se considera más rápido?
e) ¿Cuántos estudiantes, tomando en cuenta las dos Facultades, tardaron más de 52 minutos?

RESPUESTAS:

PREGUNTA	RESPUESTA	
	ESTUDIANTES MEDICINA (A)	ESTUDIANTES INGENIERIA (B)
27.a)	MEDIA = 44 min	MEDIA = 41 min
	MEDIANA = 44 min	MEDIANA = 40 min
	MODA = 44 min	MODA = 39 min
27.c)	MEDIA TOTAL= 42 min	
27.d)	EN EL GRUPO DE INGENIERIA.	
27.e)	22 ESTUDIANTES	

28.- A continuación se presenta un par de distribuciones que contienen los niveles de nicotina (ng/ml) en la sangre de un grupo de fumadores y un grupo de no fumadores. Estas mediciones se registraron como parte de un estudio de los diversos factores de riesgo de enfermedad cardiovascular:

Nivel de nicotina (ng/dL)		Fumadores	No Fum.
140	160	8	11
161	181	2	9
182	202	15	10
203	223	10	5
224	244	16	1

Se pide:

a) obtenga el grado de homogeneidad y de asimetría de cada grupo
b) Construya el polígono de frecuencia de los datos
c) Indique el número de personas con niveles entre 172 y 210 ng/ml en el grupo de los fumadores
d) ¿Qué puede concluir de acuerdo a los resultados obtenidos?

RESPUESTAS:

PREGUNTA	RESPUESTA	
	FUMADORES	NO FUMADORES
28.a)	CV= 14.3%	CV= 13.5%
	SESGO= - 0.2	SESGO= 0.875
28.c)	19 PERSONAS	

29.- Se han recogido las notas de Bioestadística I del primer semestre en las secciones A y B y éstos fueron los resultados:

Li	Ls	fiA	fiB
5	7	2	8
8	10	16	37
11	13	40	53
14	16	62	22
17	19	20	10

a) Compare la dos distribuciones en cuanto a homogeneidad y picuidez, ¿qué concluye?
b) Si un estudiante obtuvo una calificación de 09 puntos, ¿en qué sección se considera mejor estudiante?
c) Grafique la ojiva de las dos distribuciones y señale la mediana, ¿qué concluye?

RESPUESTAS:

PREGUNTA	RESPUESTA	
	SECCION A	SECCION B
29.a)	CV= 21%	CV= 25%
	CURTOSIS= LEPTOCURTICA	CURTOSIS= PLATICURTICA
29.b)	EN LA SECCION B	

30.- Los siguientes datos se refieren a la cantidad de Eritrocitos contenidos en la sangre (en millones), extraída de una muestra de varios niños:

Eritrocitos ($x10^6$)		fi
Li	Ls	
00	02	3
02	04	10
04	06	15
06	08	20
08	10	25
10	12	20
12	14	15
14	16	10
16	18	3

Se pide:

a) Determine el porcentaje de niños que obtuvieron una cantidad igual o menor que la media del grupo
b) Determinar las cantidades límites entre las cuales está el 50% central de los niños
c) En cuanto a la simetría de la distribución, ¿qué podría afirmar?

RESPUESTAS:

PREGUNTA	RESPUESTA
30.a)	50%
30.b)	(06-12) 10^6
30.c)	DISTRIBUCIÓN SIMETRICA

31.- Se quiere hacer un estudio acerca del efecto de la vitamina D sobre el crecimiento de las personas. Para ello se seleccionan dos muestras: una (A) formada por 35 adolescentes que no ingieren vitamina D, y la otra (B) constituida por 10 adolescentes, que se les suministra una dosis diaria de vitamina D por un mes. Después de este tiempo, se toman las estaturas (en mts) de ambas muestras:

ESTATURA (m)		fiA	fiB
Li	Ls		
1.45	1.50	2	2
1.50	1.55	6	1
1.55	1.60	9	0
1.60	1.65	10	3
1.68	1.70	5	1
1.70	1.75	2	2
1.75	1.80	1	1

Se pide:

a) Un adolescente que mide 1.69mts, ¿en qué grupo se considera de baja estatura?
b) Determine para ambos grupos la variabilidad y el grado de asimetría
c) Dibuje el histograma de frecuencia para los dos grupos
d) Interprete los resultados anteriores y diga si la vitamina D influye en el crecimiento. Justifique

RESPUESTAS:

PREGUNTA	RESPUESTA	
	GRUPO A	GRUPO B
31.a)	En el grupo B.	
31.b)	CV= 4.4%	CV= 6.8%
	SESGO= -0.14	SESGO= 0

32.- Se seleccionó un grupo de 70 señoras, de igual edad y estatura, para probar la eficiencia de una dieta para reducir del peso. Se desea comparar sus pesos (en Kgs), antes (momento A) y después (momento B) de someterse a dicho régimen, y estos fueron los resultados:

PESO (kg)		fiA	fiB
Li	Ls		
50	54	0	3
55	59	3	6
60	64	3	12
65	69	8	20
70	74	19	15
75	79	25	9
80	84	12	5

Se pide:

a) Realice las tablas de distribución
Si los médicos consideran que, para este tipo de mujer los valores que están entre (62 ± 9) Kgs son normales:
b) Obtenga el intervalo $\overline{X} \pm S$ de cada muestra y compare con el normal ¿qué concluye?
c) Obtenga el % de mujeres que entran en el intervalo normal, para cada muestra y compárelos
d) Compare las dos muestras en cuanto a homogeneidad

RESPUESTAS:

PREGUNTA	RESPUESTA	
	ANTES DE DIETA	DESPUÉS DE DIETA
32.b)	$74 \pm 6 \quad Kg$	$68 \pm 7 \quad Kg$
32.c)	28%	62%
32.d)	CV=8.1%	CV=10.3%

33.- Se quiere hacer un estudio acerca del efecto de la cantidad de alcohol ingerida sobre los niveles de triglicéridos en las personas. Para ello se seleccionaron dos muestras: una (A) formada por 40 adultos que ingieren licor frecuentemente, y la otra (B) constituida por 35 adultos, que no ingieren licor. Se hacen las pruebas y éstos fueron los resultados:

Li	Ls	fiA	fiB
80	100	2	4
100	120	1	8
120	140	3	11
140	160	4	7
160	180	7	3
180	200	10	2
200	220	8	0
220	240	5	0

Se pide:

a) Si los valores normales de los triglicéridos están comprendidos en el intervalo (115 ± 35) mg/100ml, determine cuántas personas de cada grupo tienen sus valores fuera de lo normal
b) Dibuje el histograma de frecuencia para los grupos
c) Obtenga el grado de asimetría de cada grupo
d) En función de los resultados anteriores, interprete los datos y diga si el nivel de licor ingerido influye en los valores de triglicéridos

RESPUESTAS:

PREGUNTA	RESPUESTA	
	INGIEREN LICOR	NO INGIEREN LICOR
33.a)	36 personas	9 personas
33.c)	SESGO= -0.324	SESGO= 0.4

34.- Se quiere registrar el peso (Kgs) de un grupo de 35 personas (hombres A, mujeres B) y éstos fueron los resultados:

PESO (Kg)		fi_A	fi_B
Li	Ls		
55	60	1	2
60	65	1	4
65	70	3	9
70	75	6	4
75	80	4	1

Hallar:

a) Media aritmética del grupo de 35 personas
RESPUESTA: 69 Kg
b) Compare las dos distribuciones en cuanto a simetría y picudez (interprete los resultados)
RESPUESTA: Distribución del grupo A es asimétrica negativa y platicúrtica; la distribución del grupo B es asimétrica positiva y platicúrtica.

35.- Considérese los siguientes valores de colesterol en mg/dl:

GRUPO 1	200	210	190	220	190	215	180	150
GRUPO 2	210	235	180	235	235	220	190	300

Seleccione la respuesta correcta:

- La desviación estándar del grupo 1 es:
  a) La misma que para el grupo 2
  b) Menor que la del grupo 2
  c) Mayor que la del grupo 2
  d) Diferente al grupo 2, pero con la misma media
  e) Indeterminable a partir de estos datos

- El coeficiente de asimetría es:
  a) Igual a cero para el grupo 1
  b) Positiva para el grupo 2
  c) Cola izquierda más larga para el grupo 1
  d) Igual a cero para el grupo 2
  e) Ninguna de las anteriores

36.- El contenido de hemoglobina (g/dl) en la sangre fue medido en una muestra de niños seleccionada aleatoriamente. Los resultados obtenidos expresados en fueron:

10.0	12.2	13.2	12.0	9.4	11.0	12.8	11.2
11.2	11.8	9.6	11.5	13.1	10.8	9.8	10.9
13.5	11.6	12.7	10.3	12.1	10.6	13.2	11.8
9.2	12.5	11.4	10.0	12.0	11.2	11.4	14.2

a) Construir la distribución de frecuencia
b) Con los datos sin agrupar calcule media, moda, mediana, coeficiente de variación y simetría. Interprete.

RESPUESTAS:
MEDIA: 11.5 g/dL
MODA: 11.2 g/dL
MEDIANA: 11.5 g/dL
CV: 10%
SIMETRÍA: ASIMETRÍA POSITIVA

37.- . La siguiente tabla muestra los valores de hemoglobina (gr/L) de un grupo de pacientes aparentemente sanos y anémicos que asistieron al Laboratorio del Hospital Clínico Universitario 14-nov-02

"Anémicos"					"Aparentemente Sanos"				
69	95	92	75	82	125	146	132	150	165
78	96	94	84	92	145	145	125	139	137
98	76	80	82	87	146	159	146	165	146

Con **los datos sin agrupar**, calcule, compare e interprete para cada grupos:
a) Media

RESPUESTA: ANÉMICOS: 85g/dL APARENTEMENTE SANOS: 145g/dL

b) Coeficiente de variación

RESPUESTA: ANÉMICOS: 10.6% APARENTEMENTE SANOS: 8.3%

c) Valores límites de hemoglobina del 90% central de las muestra

RESPUESTA: ANÉMICOS: (69-96)g/dL
APARENTEMENTE SANOS: (125-165)g/dL

38.- En el siguiente conjunto de números, se proporcionan los pesos (redondeados a la libra más próxima) de los bebés nacidos durante un cierto intervalo de tiempo en un hospital:

4, 8, 4, 6, 8, 6, 7, 7, 7, 8, 10, 9, 7, 6, 10, 8, 5, 9, 6, 3, 7, 6, 4, 7, 6, 9, 7, 4, 7, 6, 8, 8, 9, 11, 8, 7, 10, 8, 5, 7, 7, 6, 5, 10, 8, 9, 7, 5, 6, 5.

- Calcular las medidas de tendencia central, medidas de dispersión, medidas de forma.

RESPUESTAS:
MEDIA: 7 lbs
MODA: 7 lbs
MEDIANA: 7 lbs
DESVIACIÓN ESTANDAR: 2lbs
DESVIACIÓN MEDIA: 1 lb
VARIANZA 4 lbs^2
CV: 29%
SESGO: SIMETRÍCA
CURTOSIS: PLATICURTICA

- ¿Es esta una distribución sesgada? De ser así, ¿en qué dirección?

RESPUESTA: DISTRIBUCIÓN INSESGADA

- Encontrar el percentil 24.

RESPUESTA: 6 lbs

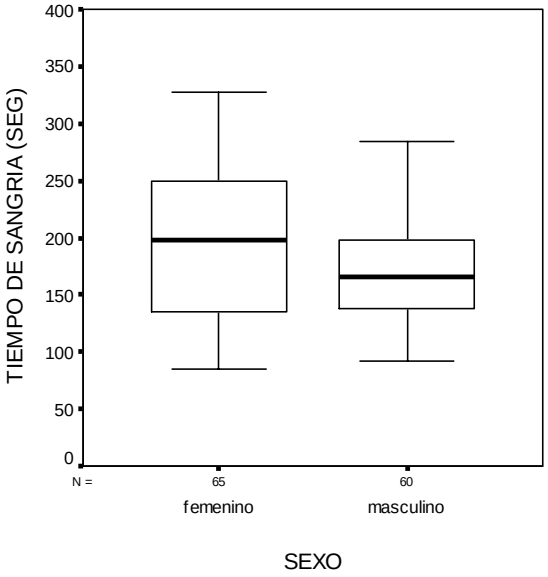
39.- A continuación se dan los resultados obtenidos en el tiempo de reacción (seg) ante un estímulo auditivo, en una muestra de 50 universitarios:

0,110	0,110	0,126	0,112	0,117	0,113	0,135	0,107	0,122
0,113	0,098	0,122	0,105	0,103	0,119	0,100	0,117	0,113
0,124	0,118	0,132	0,108	0,115	0,120	0,107	0,123	0,109
0,117	0,111	0,112	0,101	0,112	0,111	0,119	0,103	0,100
0,108	0,120	0,099	0,102	0,129	0,115	0,121	0,130	0,134
0,118	0,106	0,128	0,094	0,114				

- ¿Cuál es la amplitud total de la distribución de los datos?
- Obtenga la distribución de frecuencias absolutas y relativas.
- Obtenga la distribución de frecuencias acumuladas, absolutas y relativas.
- Calcular la media y la desviación con los intervalos de la tabla y después calcúlense las mismas magnitudes sin ordenar los datos en una tabla estadística.¿Con qué método se obtiene mayor precisión?
- Dibuje el polígono de frecuencias relativas.
- Dibuje el polígono de frecuencias relativas acumuladas.

40.- El siguiente grafico de caja fue construido en el SPSS y representa tiempos de sangría (seg) en un grupo de pacientes femeninos y masculinos del Banco Municipal de Sangre. Según esto, compare y explique bajo sus conocimientos estadísticos y de gráficos de caja, el comportamiento de las variables.

TIEMPO DE SANGRÍA (SEG) EN UN GRUPO DE PACIENTES FEMENINOS Y MASCULINOS DEL BANCO MUNICIPAL DE SANGRE



41.- Los siguientes datos corresponden a valores de Tiempo de sangría (seg) de un grupo de 32 pacientes que asistieron a donar sangre y a la Consulta del Banco de Sangre de Caracas el día 15 de marzo de 2003.

Donantes de Sangre				Consulta			
273	186	158	186	333	283	231	315
209	128	266	143	283	308	283	283
293	136	186	144	419	283	491	252
220	186	256	290	511	386	465	228

Con estos datos:

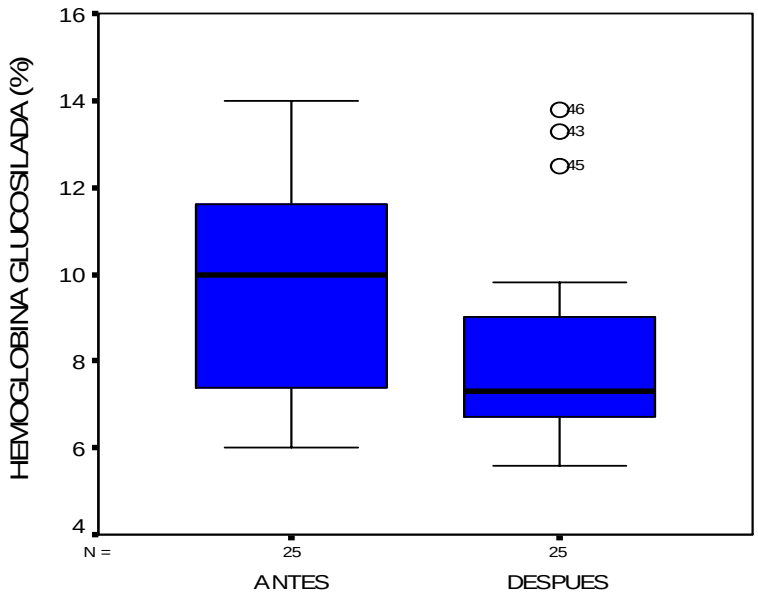
1.- Construya la distribución de frecuencias (calcule los límites aparentes y reales)

2.- Calcule e interprete los siguientes estadísticos con los datos sin agrupar:

a. Media
b. Desviación estándar
c. Moda
d. Sesgo
e. Rango Cuartílico

42.- En pacientes diabéticos la hemoglobina glucosilada es un parámetro de control durante 3-4 meses. Una persona no diabética debe tener alrededor del 6% de su hemoglobina glucosolida, un diabético controlado puede tener entre 7.5 y 8.0% de hemoglobina glucosilada, y valores mayores hablan de un mal control de estos pacientes. En el H.J.G.H. existe una Unidad de Diabetes que ayuda a los pacientes a controlar los factores de riesgo para la diabetes, evaluando sus valores de hemoglobina glucosilada al inicio y luego de tres meses en la Unidad. El siguiente gráfico de caja muestra algunos resultados de esta evaluación. Según todo lo expuesto y sus conocimientos estadísticos, se pide interpretar el gráfico de caja.

HEMOGLOBINA GLUCOSILADA EN HOMBRES ANTES Y DESPUES DE 3 MESES DE PARTICIPACIÓN EN LA UNIDAD DE DIABETES DEL H.J.G.H



Ejercicios de Probabilidades y Distribuciones de Probabilidades

1.- En una bolsa hay 4 pelotas negras y 3 blancas, mientras que en una segunda bolsa hay 3 pelotas blancas y 5 negras
a) Si se extrae al azar una pelota de la primera bolsa, y sin verla, se introduce a la segunda bolsa, ¿cuál es la probabilidad de sacar una bola negra de la segunda bolsa? **RESPUESTA:** P(2da NEGRA) = 0,619
b) Si se extrae una pelota al azar y salió blanca, ¿cuál es la probabilidad de que provenga de la primera bolsa? **RESPUESTA:** P(B/B) =0.500

2.- La caja A contiene 2 fichas blancas y 2 negras. La caja B contiene 3 blancas y 1 negra. La caja C contiene 1 blanca y 3 negras. Se juntan todas las fichas y se extrae una al azar:
a) Halle la probabilidad de obtener una ficha blanca **RESPUESTA:** P(B)=0.500
b) Una vez extraída una ficha blanca determinar la probabilidad de que sea de la caja A. **RESPUESTA:** P(A /B)=0.333

3.- En un hospital, hay dos laboratorios, A y B. El 10% de los análisis hechos por el laboratorio A salen con defecto, mientras que el 5% de los análisis realizados por el laboratorio B salen con defectos. Si en el laboratorio A se realizan 100.000 análisis al año y en el laboratorio B se realizan 50.000 análisis al año, cuál es la probabilidad de:
a) Seleccionar un análisis al azar y que sea defectuoso
RESPUESTA: P(D)=0.083
b) Si se realiza el examen y se comprueba que es defectuoso, hallar la probabilidad que haya sido realizado en el laboratorio A.
RESPUESTA: P(A/D)=0.803

4.- En una caja hay 10 rosas rojas, 20 rosas blancas y 30 rosas amarillas:
a) Si se extrae una rosa al azar, cuál es la probabilidad de que:
- a.1) Sea amarilla **RESPUESTA:** P(A)=0.500
- a.2) Sea roja o amarilla **RESPUESTA:** P(R$\cup$A)=0.667
- a.3) No sea ni roja ni amarilla **RESPUESTA:** P(B)=0.333

b) Si se extraen dos rosas al azar, cuál es la probabilidad de que:
- b.1) La primera sea blanca y la segunda roja **RESPUESTA:** P(1ºB$\cap$2ºR)=0.056
- b.2) Una de ellas sea amarilla **RESPUESTA:** P(1.SEA.A)=0.508
- b.3) Al menos salga una roja **RESPUESTA:** P(AL MENOS 1 R)=0.308

5.- En una cierta universidad el 4% de los hombres y el 1% de las mujeres miden más de 1.80mts. El 60% de los estudiantes son mujeres. Si se selecciona uno al azar y es de estatura mayor a 1.80mts, ¿cuál es la probabilidad de que sea mujer?
RESPUESTA: P(MUJER/>1.80m)=0.273

6.- En los accidentes de tránsito, las autoridades siempre revisan a los conductores involucrados un examen para determinar la posible ingesta de

alcohol. El 97% de los exámenes da resultados positivos, 2% son negativos y el resto son resultados dudosos. Suponiendo que en el 1% de los accidentes el conductor haya ingerido alcohol, ¿cuál es la probabilidad de que un conductor que resultó positivo para el consumo de alcohol, efectivamente lo haya ingerido?
RESPUESTA: P(OH/+)= 0,01

7.- Para diagnosticar la Disfunción Timpánica (DT) se usa la Timpanometría. Esta técnica tiene una sensibilidad de 83% y especificidad de 94%. En la población que nos ocupa la probabilidad de que una persona sufra una DT es del 8%. Si a una persona se le realiza una Timpanometría y da resultado positivo, cuál es la probabilidad de que no sufra una DT?
RESPUESTA: P($\overline{DT}$/+)=0.454
8.- Suponiendo que con la administración de un nuevo antibiótico a 5 enfermos, la probabilidad de que se curen todos es 0.00098. ¿Cuál es la probabilidad de que en un grupo de 7 enfermos, se curen menos de 2 o más de 5?
RESPUESTA: P(2>X>5)=0.446

9.- De 100 sujetos que se utilizaron para probar un medicamento nuevo, dos mostraron efectos nocivos. Si el nuevo medicamento se administra a 80 sujetos, determinar la probabilidad de que manifiesten efectos nocivos:
a) A lo sumo 3 sujetos **RESPUESTA:** P($X \leq 3$)=0.921
b) Por lo menos 2 sujetos **RESPUESTA:** P($X \geq 2$)=0.475

10.- La presión sanguínea se distribuye normalmente en una muestra de 5000 adultos con una media de 140mm de Hg. Sabiendo que el 0.62% son hipotensos (menos de 110 mm de Hg), calcular:
a) Número esperado de adultos con tensión igual a 125 mm de Hg.
RESPUESTA: Nº DE PERSONAS CON X=125mmHg= APROXIMADAMENTE 64
b) Porcentaje de adultos con presión sanguínea comprendida entre 118 y 145 mm de Hg.
RESPUESTA: % DE PERSONAS CON PRESION ENTRE 118 Y 145 mmHg= 63%
c) Probabilidad de que un adulto elegido al azar tenga una presión sanguínea entre 116 y 136 mm de Hg
RESPUESTA: P($116 < X < 136$)=0.348
d) Porcentaje de adultos con presión sanguínea menor que 138 mmHg o mayor que 152 mm de Hg
RESPUESTA: % DE ADULTOS CON PRENSIÓN <138 O >152 mmHg= 59%
e) Mínima presión sanguínea del 25% de los hipertensos.
RESPUESTA: X= 148mmHg

11.- Varias fábricas de detergentes, cuya cantidad de producción en el mercado es la misma, deciden colocar un cupón dentro de la caja de jabón, quien logre encontrarlo se gana Bs. 1.000.000. Sin embargo, el fabricante del JABON X, quien

no está tan dispuesto a pagar tanto dinero, decide que por cada 50 cajas solo 2 tendrán el cupón, el fabricante de JABON B decide dar más oportunidad colocando el cupón a 5 cajas por cada 20 que produzca, el fabricante de JABON C, quien desea aumentar su venta, le coloca 7 cajas del premio por cada 12. Una ama de casa compró un detergente, y al abrirlo descubre el cupón, ¿cuál es la probabilidad de que la marca del detergente sea JABON X?
RESPUESTA: P($JABON \quad X / P$)=0.046

12.- Un análisis para descubrir una enfermedad venérea arroja un índice del 6% de resultados falsos positivos y un índice del 20% de resultados falsos negativos. En una población, el 2% de las personas padecen dicha enfermedad. ¿Cuál es la probabilidad de:
a) que una persona con resultados positivos padezca de esta enfermedad?
RESPUESTA: P(E/+)=0.214
b) una persona con resultados negativos padezca la enfermedad?
RESPUESTA: P(E/-)=0.004

13.- En un laboratorio hay tres cajas (A, B y C) con ampolletas de agua. La gaveta A contiene 3 con agua destilada y 2 sin destilar. La B tiene 6 con agua destilada y 4 sin destilar. La C tiene 5 con agua destilada y 1 sin destilar. Si se tomó una ampolleta al azar y resultó con agua no destilada ¿Cuál es la probabilidad de que provenga de la gaveta B?
RESPUESTA: P($B / \overline{D}$)=0.571

14.- Diversas enfermedades producen los mismos síntomas, por ejemplo, dolor de cabeza. Suponiendo que en un grupo de personas, el 35% sufre miopía, el 20% tiene resfriado y el 15% tuvo mala digestión. Imaginemos que el dolor de cabeza se presenta a los miopes en un 90%, a los resfriados en un 67% y a los que tienen mala digestión en un 65%. ¿Cuál es la probabilidad de que si se selecciona una persona al azar con dolor de cabeza, tenga resfriado?
RESPUESTA: P(R/DC)=0.245

15.- Un cirujano desarrolla una técnica quirúrgica para una enfermedad en la cual la mortalidad post-operatoria usual es de 20%. Si en este mes debe operar a 10 personas que tienen dicha enfermedad, calcular la probabilidad de que:
a) ninguna se muera después de la operación
RESPUESTA: P(x=0)=0.107
b) a lo sumo 2 personas se mueren
RESPUESTA: P($X \leq 2$)=0.678

16.- Un defecto metabólico sucede un caso en cada 10000. En un hospital se reciben diariamente 80 pacientes ¿Qué probabilidad hay de que:

a) por lo menos 2 sufran ese defecto?
RESPUESTA: P($X \geq 2$)= 6,4. 10^{-5}
b) a lo sumo se den tres casos?
RESPUESTA: P($X \leq 3$)=0.99996825

17.- El 1.2% de los exámenes efectuados en cierto laboratorio resultan defectuosos. Halle la probabilidad de que de 100 exámenes realizados en ese laboratorio sean defectuosos:
a) Exactamente 40
RESPUESTA: P(x=40)=$5.43.10^{-46} \approx 0$
b) A lo sumo 10
RESPUESTA: P($x \leq 10$) ≈ 1

18.- Se han inyectado 5 enfermos elegidos al azar con nuevo antibiótico. Suponiendo que la probabilidad de que ninguno se cure es de 0.3456. Determinar la probabilidad de que, aplicado ese antibiótico a un grupo de 4 personas, se curen menos de 2 personas o más de 3 personas.
RESPUESTA: P(2>x>3)=0.8237

19.- Los niveles de calcio en suero, en una muestra de 400 adultos se distribuyen normalmente con una media de 10gr%. y una desviación de 2.5gr%. Calcule:
a) Número de adultos con un nivel de calcio menor a 7.8gr%.
RESPUESTA: 76 PERSONAS
b) Porcentaje de adultos con niveles entre 8.5 y 10.8gr%.
RESPUESTA: 35%
c) Probabilidad de elegir una persona con un nivel de calcio mayor que 12gr%.
RESPUESTA: P(x>12gr%)=0.212
d) El 60% central de las personas, ¿qué niveles de calcio les corresponden?
RESPUESTA: LE CORRESPONDEN NIVELES DESDE 7.9gr% Y 12.1gr%

20.- El 2.8% de los exámenes efectuados en un laboratorio son de heces. Si el laboratorio en cuestión procesa 100 exámenes diarios, determinar la probabilidad de que:
a) A lo sumo 4 exámenes sean de heces
RESPUESTA: P($X \leq 4$)=0.848
b) Por lo menos 4 lo sean
RESPUESTA: P($X \geq 4$)=0.308

21.- Los niveles de calcio en suero en una muestra de 500 adultos se distribuyen normalmente con una media de 10.0mg%. Sabiendo que P(X<11.0 mg%) =0.9772, se pide:

a) Número esperado de adultos con un nivel de calcio en suero menor que 8.9mg%
RESPUESTA: 7 ADULTOS
b) Porcentaje de adultos con nivel de calcio en suero igual a 9.0mg%

RESPUESTA: 11.4% APROXIMADAMENTE

22.- En un sorteo especial de Lotería hay una emisión total de 10000 billetes, desde el 0000 al 9999. Carlos compra los números: 4641, 3828, 6828, 6840. El sorteo se utilizará escogiendo dos números al azar: el primero que salga ganará el segundo premio y el segundo número que salga, ganará el primer premio. Carlos desea saber:
a) ¿Cuál es la probabilidad de ganar el 1º y 2º premio?
RESPUESTA: $P(1º \cap 2º)=1,2.10^{-7}$
b) ¿Cuál es la probabilidad de ganar el 1º o 2º premio?
RESPUESTA: $P(1º \cup 2º)=8,0.10^{-4}$

23.- La evaluaciones de 150 trabajadores de una empresa se distribuyen normalmente con una media de 12.0 puntos y una desviación de 2.5 puntos. Al observar los resultados el gerente toma las siguientes decisiones:
- las personas que tienen de 10.8 a 13.0 puntos permanecen en el mismo cargo
- de 14.0 a 16.0 se le subirá el sueldo
- el 5% de máximo puntaje ascenderán de cargo
- el 7% de mínimo puntaje será despedido

Si se selecciona un trabajador al azar, hallar:
a) la probabilidad de permanecer en el mismo cargo

RESPUESTA: $P(10.8<x<13.0)=0.3398$

b) la máxima puntuación que hay que tener para ser despedido

RESPUESTA: 8.3 PTOS

24.- Un análisis para descubrir una enfermedad contagiosa arroja un índice del 8% de resultados positivos falsos y un índice del 15% de resultados negativos falsos. Si se escoge una muestra donde se sabe que el 5% posee dicha enfermedad, ¿cuál es la probabilidad de que una persona con resultado negativo padezca de la enfermedad?
RESPUESTA: P(E/-)=0.009 revisar redacción

25.- Se lanzan dos dados una sola vez. Calcule la probabilidad de que:
a) Salgan los dos números iguales
RESPUESTA: P(2#IGUALES)=1/6
b) salga en los dados, números menores que 4
RESPUESTA: P(#<4)=1/4
c) Si salen, en los dos dados números menores de 4, ¿cuál es la probabilidad de que los dos sean impares?
RESPUESTA: P(IMPAR /#<4)=0.44

26.- Supongamos que en un cierto hospital cada niño que nace tiene una probabilidad de 0.55 de ser varón. Encuentre la probabilidad de que si nacen 5 niños:
a) más de 1 sea varón

RESPUESTA: P(X>1)=0.869

b) exactamente hayan 2 hembras
RESPUESTA: P(X=3)=0.337

27.- Un cardiólogo estima que la distribución del tiempo de duración de un tipo de marcapaso, desde su instalación hasta que éste comienza a fallar, sigue una distribución normal con media 5.2 años y una desviación de 0.8 años. Si se ensaya este tipo de marcapaso en un grupo de 120 personas, calcular:
a) la probabilidad de escoger una persona al azar, cuyo marcapaso tiene un tiempo de duración entre 4.5 y 6.2 años
RESPUESTA: P(4.5<X<6.2)=0.705
b) número esperado de personas cuyo tiempo de duración del marcapaso sea mayor a 7.0 años
RESPUESTA: 2 PERSONAS

28.- Se quiere estudiar la incidencia del cigarro sobre el cáncer pulmonar. Después de una serie de investigaciones se determinará que de 300 personas, 120 eran fumadores y el resto no fumadores; el 85% de los fumadores y el 20% de los no fumadores presentaban cáncer pulmonar. Si se escoge una persona al azar y resultó con cáncer. ¿Cuál es la probabilidad de que no sea fumador?
RESPUESTA: P($\overline{F}/C$)=0.261

29.- La probabilidad de que un niño nazca con una anomalía congénita es de 0.02. En la Maternidad Concepción Palacios, nacen diariamente, un estimado de 125 neonatos. Se quiere saber la probabilidad de que en un día determinado:
a) ningún neonato tenga la anomalía
RESPUESTA: P(X=0)=0.082
b) a lo sumo nazcan 3 neonatos con la anomalía
RESPUESTA: P($X \leq 3$)=0.758

30.- Un 15% de los pacientes atendidos en un hospital son hipertensos y un 10% son obesos, y de estos grupos 3% son hipertensos y obesos. ¿Qué probabilidad hay de elegir un paciente al azar que sea obeso o hipertenso?
RESPUESTA: P(H$\cup$O)=0.220

31.- La probabilidad de que una madre diabética transmita su enfermedad a su primer hijo es 0.65. Si el primer hijo es diabético, la probabilidad de que el segundo hijo también lo sea es 0.23; pero si el primer hijo no heredó dicha enfermedad, la probabilidad de que el segundo si la herede es 0.86. Hallar:
a) la probabilidad de que el segundo hijo herede la enfermedad
RESPUESTA: P(2º HIJO HEREDE)=0.451
b) Si el segundo hijo es diabético, ¿cuál es la probabilidad de que el primer hijo también posea dicha enfermedad?
RESPUESTA: P(1ºD/2ºD)=0.332

32.- En un hospital, acuden 200 enfermos de dengue, cuyos valores de plaquetas en la sangre siguen una distribución normal de media 110000 y una desviación típica de 20000. El médico tratante toma la siguiente determinación:

- Si el paciente tiene más de 150000 plaquetas, se dirigirá a su domicilio con un tratamiento adecuado
- Si el paciente tienen entre 75000 y 150000 plaquetas, deberá regresar en 24 horas para realizarse otro examen de sangre
- Si el paciente tiene menos de 75000 plaquetas, deberá hospitalizarse
- Si el paciente tiene menos de 50000 plaquetas, deberá hospitalizarse y hacerle una transfusión de sangre

Se quiere saber:

a) Número de pacientes esperados que serán hospitalizados sin necesidad de transfusión de sangre

RESPUESTA: APROXIMADAMENTE 8 PACIENTES

b) Probabilidad de que un paciente deba acudir a las 24 horas para otro examen de sangre

RESPUESTA: $P(75.000<X<15.0000)=0.937$

33.- En el mes de noviembre, acudieron 220 pacientes al laboratorio, solicitando exámenes de Hematología Completa y plaquetas, por presentan cansancio, debilidad, malestar general. Según los resultados se determinaron que 105 pacientes tenían un simple cuadro gripal, 50 pacientes tenían dengue y dentro de estos dos grupos se encontró que 20 pacientes tenían ambas enfermedades. Si se escoge un paciente al azar, ¿cuál es la probabilidad de:

a) tener dengue?

RESPUESTA: $P(D)=0.227$

b) tener solo gripe?

RESPUESTA: $P(SOLO\ G)=0.386$

c) tener dengue o gripe?

RESPUESTA: $P(D \cup G)=0.614$

d) si tiene gripe, hallar la probabilidad de que tenga dengue

RESPUESTA: $P(D/G)=0.1905$

34.- El colesterol en adultos se distribuye normalmente en una muestra de 3000 adultos con una media de 140mg/dl. Sabiendo que el 99.79% de esta muestra tiene más de 35mg/dl de colesterol, calcular:

a) número esperado de adultos con colesterol con 120mg/dl

RESPUESTA: APROXIMADAMENTE 21 PERSONAS

b) Porcentaje de adultos con colesterol menor que 125mg/dl o mayor que 190mg/dl

RESPUESTA: 43%

c) Máximo valor de colesterol del 40% que tiene los menores niveles de colesterol

RESPUESTA: 131 mg/dL

35.- En el Banco de Sangre del Hospital Clínico Universitario hay disponibles los siguientes tipos de sangre, debidamente identificados:

28 bolsas tipo A
17 bolsas tipo B
20 bolsas tipo AB
35 bolsas tipo O (la sangre tipo O puede donar a cualquier otro tipo)

En cierto momento, acuden a emergencia 9 pacientes que requieren de sangre tipo AB. Rápidamente y sin fijarse en la etiqueta, una enfermera toma las bolsas necesarias (una por persona) para atender a los pacientes. Se desea saber:

a) ¿Cuál es la probabilidad de que a todos los pacientes se les dé el tipo de sangre correcto?

RESPUESTA: P($X = 9$)=0.005

b) ¿Cuál es la probabilidad de que por lo menos tres de los pacientes reciban la correcta transfusión?

RESPUESTA: P($X \geq 3$)=0.950

36.- La incidencia de pseudotrombocitopenias EDTA-dependientes (PCTP) es de 1 en 400 hematologías realizadas. En el Hospital Clínico de Caracas, el día 15 de marzo de 2000, se hicieron un total de 150 hematologías, y se quiere conocer para ese día la probabilidad de que:

a) ningún paciente presente en su hematología una PCTP

RESPUESTA: P($X = 0$)=0.687

b) menos de dos pacientes tengan una PCTP

RESPUESTA: P($X < 2$)=0.945

37.- La hemofilia (trastorno de la coagulación sanguínea) es una enfermedad hereditaria que sólo padecen los hombres, mientras que las mujeres son portadoras. La probabilidad de que una familia con antecedentes transmita la enfermedad a su primer hijo es de 0.151. Si el primer hijo, heredó la enfermedad, la probabilidad de que el segundo la herede es de 0.025. Pero si el primer hijo no heredó la enfermedad, la probabilidad de que el segundo si la herede es de 0.432.

a) Realizar el diagrama de árbol correspondiente

b) Hallar la probabilidad de que el segundo hijo hembra herede la enfermedad

RESPUESTA: P(2ºH.H)=0.182

NOTA: Considerar que la probabilidad de ser varón es igual a 0.51 y el complemento es la probabilidad de ser hembra

38.- En las competencias para cualquier deporte, las autoridades siempre hacen un examen anti-doping. El 45% de los exámenes da positivo del cual el 8% da erróneo, es decir, da doping positivo, sin serlo. Suponiendo que el 10% de los deportistas haya ingerido drogas de abuso, ¿cuál es la probabilidad de que un deportista que resultó positivo, haya consumido drogas realmente?

RESPUESTA: P($D/+$)=0.385

39.- La hemoglobina en hombres se distribuye normalmente en una muestra de 4000 adultos, con una media de 14.0gr/dl. Sabiendo que el 2.17% tiene menos de 9.0gr/dl, calcular:
a) Número esperado de adultos con hemoglobina de 12.5gr/dl
RESPUESTA: 54 ADULTOS APROXIMADAMENTE
b) Porcentaje de adultos con hemoglobina menor que 13.8gr/dl o mayor que 15.0gr/dl
RESPUESTA: 81.3 %
c) mínimo valor de hemoglobina del 35% que tiene mayores niveles.
RESPUESTA: 15.0 gr/dL

40.- Un traumatólogo estima que el tiempo de duración de un tipo de prótesis de cadera desde su instalación hasta que su deterioro, sigue una distribución normal con una media de 5 años y 3 meses, y una desviación de 11 meses. Se pide:
a) si se escoge una prótesis al azar, ¿cuál es la probabilidad de que falle después de 4 años?
RESPUESTA: P($X > 4$)=0.913
b) Si el traumatólogo trató a 50 personas, ¿cuántos se estima que le dure la prótesis entre 6 y 8 años?
RESPUESTA: 10 PERSONAS

41.- En ocasiones, algunos pacientes que desean realizarse un examen de orina, traen sus muestras en los recolectores pero mal sellados; este hecho puede producir un resultado positivo cuando, en realidad no hay infección. Suponiendo que, en un determinado laboratorio, llegan 120 muestras de orina y el bioanalista se da cuenta que 30 están mal selladas, de todas maneras, se practica el examen para todas y se obtuvo que el 40% de las muestras mal selladas y el 25% de las que estaban bien selladas resultaron positivas. Si se selecciona un examen al azar y resultó ser positivo, ¿cuál es la probabilidad de que la muestra de orina venía en el frasco mal sellado?
RESPUESTA: P(MS/+)=0.348

42.- Supongamos que en cierto hospital cada niño tienen una probabilidad de 0.55 de ser varón. Encuentre la probabilidad de que si nacen 6 niños:
a) A lo sumo 3 sean varones
RESPUESTA: P($X \leq 3$)=0.558
b) Nazcan entre 2 y 4 varones
RESPUESTA: P($2 \leq X \leq 4$)=0.767

43.- En el Banco Municipal de sangre, a todas las muestras de sangre donadas para transfusiones se les realiza serología para Hepatitis B, Hepatitis C, HIV y VDRL. Los siguientes son los resultados obtenidos el día 15 / 06 / 01:

45 con serología negativa para Hepatitis B y C, HIV y VDRL
10 con Hepatitis B positivo
12 con VDRL positivo
23 con Hepatitis C positivo

a) En cierto momento, se requieren 7 muestras de sangre con serología completamente negativa. Si se seleccionan al azar las muestras:

- ☞ ¿Cuál es la probabilidad de que todas las muestras se seleccionen correctamente?

RESPUESTA: P(X=7)=0.008

- ☞ ¿Cuál es la probabilidad de que por lo menos tres de las muestras sean las correctas?

RESPUESTA: P($X \geq 3$)=0.773

b) Si se requiere separar todas las muestras VDRL positivo, ¿Cuál es la probabilidad de que escogiendo al azar se tomen exactamente esas muestras?

RESPUESTA: P(X=12) = 3,147 x 10^{-11}

44.- En el mes de diciembre 2000, acudieron 232 pacientes al laboratorio, solicitando exámenes de Hemoglobina Completa y Plaquetas, por presentar cansancio, debilidad, malestar general. Según los resultados se determinaron que 105 pacientes tenían faringitis, 50 pacientes tenían dengue y dentro de estos dos grupos se encontró que 20 pacientes tenían ambas enfermedades. Si se escoge un paciente al azar, ¿Cuál es la probabilidad de:

a) tener dengue?

RESPUESTA: P(D)=0.216

b) tener solo faringitis?

RESPUESTA: P(SÓLO F)=0.366

c) tener dengue o faringitis?

RESPUESTA: P(DUF)=0.582

d) Si tiene faringitis, hallar la probabilidad de que tenga dengue.

RESPUESTA: P(D/F)=0.1905

45.- En un campus universitario existen 3 carreras sanitarias. Se sabe que el 50% cursan estudios de Enfermería, el 30% Medicina y el 20% Veterinaria. Los que finalizaron sus estudios son el 20, 10 y 5% respectivamente. Elegido un estudiante al azar, hállese la probabilidad de que haya acabado la carrera.

RESPUESTA: P(FE)=0.140

46.- La siguiente tabla muestra los resultados de la evaluación de la prueba de detección en la que participaron una muestra aleatoria de 650 individuos con la enfermedad y una segunda muestra aleatoria independiente de 1200 individuos sin enfermedad.

RESULTADO DE LA PRUEBA	**ENFERMEDAD**	
	PRESENTE	AUSENTE
POSITIVO	490	70
NEGATIVO	160	1130

- Calcule la sensibilidad y especificidad de la prueba.

RESPUESTA: SENSIBILIDAD 75.4% Y ESPECIFICIDAD 94.2%

- Si la tasa de enfermedad en la población general es 0.002, ¿Cuál es el valor que predice la positividad de la prueba?

RESPUESTA: VPP=2.5%

47.- La prueba de detección PPD tiene una sensibilidad de 98% y una especificidad del 80%. Si la tasa de tuberculosis es de 0,12 calcular los valores predictivos de positividad y negatividad.
RESPUESTA: VPP = 40.1% Y VPN = 99.7%

Ejercicios de Estadística Inferencial

1.- Un estudio sociológico de una región mostró que por lo general el 40% de los habitantes son menores de edad. En un pequeño estudio, se mostró que de 80 habitantes 30 eran menores de edad. Estimar la proporción de menores de edad en esa región mediante un intervalo de confianza del 0.84
RESPUESTA: 30% Y 45%

2.- En una cierta fecha, el 20% de la población suele comprar carros. Una muestra de 1500 personas revela un número de 300 personas que planean comprar carro el próximo año. Estimar el intervalo de confianza de 90% para el porcentaje de personas de la población que intentan comprar carros el año próximo
RESPUESTA: 18% y 22%

3.- Normalmente la desviación de la calificaciones de una población estudiantil es de $\sigma = 4$. Una muestra al azar de 80 estudiantes varones obtuvo una calificación media de 14.0 puntos con una desviación típica de 3.5 puntos. Otra muestra aleatoria de 50 hembras, que realizaban los mismos estudios, logró una calificación de 15.0 puntos con una desviación de 6.5 puntos. Estimar el intervalo de confianza a un nivel de 0.95 la diferencia entre ambos grupos.
RESPUESTA: - 0,94 a 2,94 ptos.

4.- Al medir la aceleración de la gravedad un grupo de 20 estudiantes elegidos al azar reportaron un valor medio de 9.84m/seg^2 con una desviación típica de 0.15m/seg^2 . Determinar los límites de la aceleración mediante un intervalo de confianza del 0.90
RESPUESTA: 9.78 m/seg^2 a 9.90 m/seg^2

5.- El promedio de bacterias contadas en 10 placas de Pietri escogidas al azar es de $3x10^4$ en mm^3 con una desviación de $2x10^3$ en mm^3. Estime los límites de confianza para el número de bacterias promedio en el grupo de cultivos de donde se tomaron las muestras: a) al 92%; b) al 86%
RESPUESTAS
a) de 28780 a 31220
b) de 29593 a 30407

6.- La probabilidad de encontrar personas con cáncer es 1 sobre 1000. Se ha estudiado la posibilidad de, si al suministrar la BCG a un grupo de personas, se puede prevenir dicha enfermedad. De los 14000 personas que se les suministró la vacuna, solo 7 presentaron cáncer. ¿Es cierto que la BCG pueda prevenir el cáncer?. Probar a un nivel de significación del 1%.
RESPUESTA: Acepto H_o con un nivel de confianza de 99%, es decir, la BCG no previene el cáncer.

7.- En una ciudad, el 20% de los ciudadanos se enferman de cólera. Para evitar el contagio de dicha enfermedad, se realiza una campaña preventiva y se escoge una muestra de 200 personas, donde 14 presentaron la enfermedad. ¿Fue efectiva la campaña como medida de prevención contra el cólera?. Tome $\alpha = 1\%$.
RESPUESTA: Rechazo H_o con un nivel de confianza de 99%, es decir, fue efectiva la campaña contra el cólera.

8.- La nota media entre los estudiantes de idiomas de una universidad es 12 puntos con una desviación típica igual a 3 puntos. Mediante un nuevo método de enseñanza se espera que el rendimiento escolar sea mayor. Para ensayar esta aspiración se utiliza el nuevo método en una muestra al azar de 64 estudiantes, obteniéndose una puntuación media de 14 puntos. ¿Puede afirmarse que el nuevo método realmente es de mayor eficacia que el tradicional? Pruébelo con un nivel de confianza del 95%, con 99% y con 88%.
RESPUESTAS:
Rechazo H_o con un nivel de confianza de 95%, lo que quiere decir que fue efectiva la campaña contra el cólera.
Rechazo H_o con un nivel de confianza de 99%, lo que quiere decir que fue efectiva la campaña contra el cólera.
Rechazo H_o con un nivel de confianza de 88%, lo que quiere decir que fue efectiva la campaña contra el cólera.

9.- En un examen dado a un gran número de estudiantes de muchas escuelas, la puntuación media fue 13.0 puntos con desviación típica igual a 2.0 puntos. En una determinada escuela con 200 estudiantes, la puntuación media para el mismo examen fue de 14.4 puntos. ¿Se puede afirmar que existe diferencia significativa en el aprovechamiento de los alumnos de esta escuela con relación a las otras? (Use $\alpha = 0.05; \alpha = 0.01; \alpha = 0.06$)
RESPUESTAS:
Rechazo H_o con un nivel de significación de 5%, lo que quiere decir que existe diferencia significativa en el aprovechamiento del examen de los alumnos de la citada escuela, con respecto a la puntuación media general.
Rechazo H_o con un nivel de significación de 1%, lo que quiere decir que existe diferencia significativa en el aprovechamiento del examen de los alumnos de la citada escuela, con respecto a la puntuación media general.
Rechazo H_o con un nivel de significación de 6%, lo que quiere decir que existe diferencia significativa en el aprovechamiento del examen de los alumnos de la citada escuela, con respecto a la puntuación media general.

10.- Un laboratorio afirma que un antihistamínico de su invención tiene un 90% de efectividad en el alivio de afecciones alérgicas. En una muestra de 200 individuos que tenían alergia, la medicina suministrada alivió a 160 personas. Determinar si la aseveración del laboratorio es cierta (Use $\alpha = 0.05; \alpha = 0.01; \alpha = 0.10$)

RESPUESTAS:

Rechazo H_o , Acepto H_i , con un nivel de significación de 5%, lo que quiere decir que existe que el antihistamínico no tiene efectividad del 90%.

Rechazo H_o , Acepto H_i , con un nivel de significación de 1%, lo que quiere decir que existe que el antihistamínico no tiene efectividad del 90%.

Rechazo H_o , Acepto H_i , con un nivel de significación de 10%, lo que quiere decir que existe que el antihistamínico no tiene efectividad del 90%.

11.- En un examen de admisión a cierta carrera aprueban 3 de cada 5 candidatos. En una muestra de 50 aspirantes se han registrado 28 aprobados. Decidir si hay diferencia significativa entre la proporción de aprobados antiguos y los recientes. Use $\alpha = 0.05$

RESPUESTA:

Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, no existe diferencia significativa entre la proporción de aprobados antigua y recientemente.

12.- En un examen de ortografía en una escuela elemental, la puntuación media fue de 12 puntos con una desviación típica igual a 3 puntos; mientras que la puntuación media de 36 niñas de una sección del a escuela fue de 13 puntos con una desviación típica de 3 puntos. Ensayar las hipótesis de que las puntuaciones medias de las niñas y la escuela no presentan diferencia significativa con un nivel de 0.08.

RESPUESTA:

Rechazo H_o, Acepto H_i , con un nivel de significación de 8%, si existe diferencia significativa entre las puntuaciones medias de las niñas y la escuela.

13.- Muestras al azar de 200 piezas fabricadas por la máquina A y 100 piezas fabricadas por la máquina B dieron 19 y 5 piezas defectuosas respectivamente. Ensayar la hipótesis de que las dos máquinas no presentan diferencia significativa con un nivel de 0.12

RESPUESTA:

Acepto H_o, Rechazo H_i , con un nivel de significación de 12%, no existe diferencia entre la producción del as máquinas.

14.- Un dado se lanza 200 veces y se observa que la cara seis sale 53 veces. Ensayar la hipótesis de que el dado está bien hecho. Use $\alpha = 0.05$

RESPUESTA:

Rechazo H_o, Acepto H_i , con un nivel de significación de 5%, es decir, el dado no está bien hecho.

15.- En un hospital, a una muestra de 12 individuos con artritis se les determinó concentración de ácido úrico en sangre, obteniendo una media de 6.5 mg/dl y una desviación de 0.7 mg/dl. En un Ambulatorio, se encontró que, en una muestra de 15 individuos aparentemente sanos de la misma edad y sexo, tenían niveles medios de ácido úrico de 5.4 mg/dl y una desviación de 0.5 mg/dl. ¿Proporcionan estas muestras evidencia suficiente como para indicar una diferencia significativa en los niveles de ácido úrico en el suero de los pacientes del Hospital y el Ambulatorio? Use $\alpha = 0.05$.

RESPUESTA:

Rechazo H_o, Acepto H_i , con un nivel de significación de 5%, es decir, si existe diferencia significativa en los niveles de ácido úrico en suero de pacientes del hospital y del ambulatorio.

16.- Para comparar dos cremas dentales A y B, se toma una muestra de 20 niños y una muestra de 25 niños que utilizaron los productos A y B respectivamente, durante un año. La primera muestra revela un número medio de 2.3 caries con una desviación típica de 0.2, mientras que la segunda muestra revela un número medio de 1.8 caries con una desviación típica de 0.4. Determinar si hay diferencia entre las cremas dentales con un nivel de significación de 0.05. Se conoce que la desviación típica poblacional es de 0.35.

RESPUESTA:

Rechazo H_o, Acepto H_i , con un nivel de significación de 5%, es decir, si existe diferencia significativa entre las dos cremas dentales.

17.- En un laboratorio se estudia la acción de dos antibióticos sobre el crecimiento de una misma bacteria. La muestra del antibiótico A, que consta de 10 envases, da una media de 30000ufc y una desviación de 10000ufc, la muestra del antibiótico B, con 15 envases tiene una media de 16000ufc y una desviación de 8000ufc. Estadísticamente hablando, ¿habrá alguna diferencia significativa entre los dos antibióticos? (nivel de significación: 1%)

RESPUESTA:

Rechazo H_o, Acepto H_i , con un nivel de significación de 1%, es decir, si existe diferencia significativa entre los dos antibióticos.

18.- La desviación típica de la duración de las bombillas debe ser 150 horas. Una muestra de 30 bombillas dio una desviación típica de 130 horas. Ensayar la hipótesis de que las bombillas tienen la desviación requerida, con un nivel de confianza del 10%.

RESPUESTA:

Acepto H_o, Rechazo H_i , con un nivel de significación de 10%, las bombillas tienen la desviación requerida.

19.- Para comparar la ortografía de los alumnos de dos escuelas A y B, se toman dos muestras de 15 alumnos y 24 alumnos respectivamente. La primera muestra dio una nota media de 13.5 puntos con una desviación típica de 2.3 puntos; mientras que la segunda muestra dio una nota media de 12.7 puntos con una

desviación típica de 2.6 puntos. Determinar si los alumnos de la escuela A tienen mejor ortografía que los de la escuela B, con un nivel de significación del 0.05.
RESPUESTA:
Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, quiere decir que los alumnos de la Escuela no tienen mejor ortografía que los de la Escuela A.

20.- Una cierta Revista Médica establece que uno de cada 40 adultos con cáncer de pulmón, expresan la enfermedad después de exponerse a gases tóxicos del ambiente. Una muestra de 400 personas tomadas al azar de un total que trabaja en un ambiente contaminado, arroja que 19 de éstas muestran señales de cáncer pulmonar. Estadísticamente hablando, ¿qué diría ud. sobre la influencia del ambiente contaminado a la manifestación de cáncer?. Tome $\alpha = 0.18$
RESPUESTA:
Rechazo H_o, Acepto H_i , con un nivel de significación de 18%, es decir, quiere decir que el ambiente contaminado aumenta la aparición de cáncer de pulmón.

21.- Los contenidos de las botellas de aceite deben tener una desviación típica de 20ml. Una muestra de 20 botellas dio una desviación típica de 28ml. Determinar si las botellas tienen la desviación requerida con un nivel de significación de 0.1.
RESPUESTA:
Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de significación de 10%, que las botellas no tienen la desviación requerida.

22.- Al medir la estatura de 100 alumnos de una escuela, de los cuales 50 alumnos pertenecen a un grupo A y practican un deporte, y el resto pertenecen a un grupo B que no sienten interés alguno por ningún deporte. Se encontró que el grupo A tiene una media de 68.2 pulgadas y desviación de 2.5 pulgadas, y el grupo B tiene una media 67.5 pulgadas y una desviación de 2.8 pulgadas. ¿Se podría decir que el deporte influye en la estatura de los alumnos, sabiendo que la desviación de la población es de 2.7 pulgadas? Use $\alpha = 0.05$
RESPUESTA:
Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, quiere decir que el deporte no influye en la estatura de los alumnos.

23.- Dos dietas, una baja en grasas y otra normal, produce los siguientes resultados en individuos escogidos al azar en cuanto a contenidos de colesterol. La baja en grasa, se tienen 19 individuos, los cuales tienen un valor medio de 170 con una desviación de 14.07; mientras que 24 individuos hicieron la dieta normal teniendo un valor medio de 196 con una desviación de 20.85. Tome un nivel de confianza de 90%
a) Estadísticamente hablando, ¿qué diría ud. de las dietas?
b) ¿Cómo variarían los valores en el caso de la dieta baja en grasa en toda la población? Se sabe que la desviación de la población es de 18.
RESPUESTAS:
a) Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de confianza de 90%, que si existe diferencia entre las dos dietas.

b) Rechazo H_o, Acepto H_i , con un nivel de confianza de 90% se puede asegurar que el colesterol de las personas con dieta baja en grasa son significativamente menores que los de la población.

24.- Se quiere probar el efecto de un antibiótico sobre ciertas bacterias. Se tratan 40 cultivos con dicho antibiótico y en 32 se inhibe el crecimiento bacteriano. Mientras que en otro grupo control de 60 cultivos no son tratados con el citado antibiótico y en ellos 5 presentan crecimiento bacteriano. ¿Podemos afirmar que el antibiótico es efectivo frente a tales bacterias? Demuéstrelo con un nivel $\alpha = 0.05$.

RESPUESTA:
Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, quiere decir que el antibiótico no es efectivo frente a tales bacterias.

25.- Dos grupos A y B formados de 80 y 120 individuos respectivamente, padecen una enfermedad. Se administra un suero al grupo A, pero no al grupo B (que se llama de control). Se encuentra que en los grupos A y B, 62 y 70 individuos, respectivamente, se han recuperado de la enfermedad. Ensayar la hipótesis de que el suero ayuda a curar la enfermedad, a un nivel de significación de 0.01.
RESPUESTA:
Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de significación de 1%, que el suero si ayuda a curar la enfermedad.

26.- Los archivos de un hospital muestran que en una muestra de 1000 varones 52 ingresaron por ataque cardíaco, y en una muestra similar de mujeres 23 ingresaron por el mismo motivo. ¿Estos datos arrojan evidencia estadística de diferencia entre hombres y mujeres en cuanto a enfermedades del corazón a un nivel del 93%?
RESPUESTA:
Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de significación de 7%, que si existe diferencia significativa entre los hombres y mujeres.

27.- La resistencia media a la rotura de las cuerdas debe ser de 8000 lb. Una muestra de 8 cuerdas fabricadas por una compañía dio una resistencia media de 7759 lb. con la desviación típica de 140 lb. Determinar si las cuerdas fabricadas por esa compañía tienen la resistencia requerida, a un nivel de significación de 0.01.
RESPUESTA:
Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 99%, que las cuerdas no tienen la resistencia requerida.

28.- Una moneda se lanza 100 veces y se obtiene la cara 35 veces. Ensayar la hipótesis de que la moneda está bien hecha, a un nivel de significación de 0.05.
RESPUESTA:
Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 95%, quiere decir que la moneda no está bien hecha.

29.- De mil pacientes que se examinaron en un hospital (427 hombres y 573 mujeres) se tomaron aquellos que presentaban anemia. SE observó que de los hombres examinados, 323 tenían anemia y en el grupo de las mujeres había 375. ¿Se puede afirmar que la anemia se presenta con la misma frecuencia entre hombres y mujeres? Tome $\alpha = 1\%$

RESPUESTA:

Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 99%, quiere decir que la anemia no se presenta con la misma frecuencia en hombre y mujeres.

30.- El coeficiente de inteligencia de 16 estudiantes de una universidad A dio una media de 107 con desviación típica igual a 10, mientras que en una universidad B, un grupo de 13 estudiantes tiene un coeficiente de inteligencia media de 112 con una desviación típica de 9. ¿Existe alguna diferencia significativa entre estos dos grupos de estudiantes, con respecto a su coeficiente de inteligencia? Tome $\alpha = 0.05$

RESPUESTA:

Acepto H_o, Rechazo H_i , se puede asegurar con un nivel de de confianza de 95%, que no existe diferencia significativa entre ambos grupos.

31.- El promedio de calificaciones en la asignatura de Bioestadística I, en la Escuela de Bioanálisis es de 12.5 puntos. En el semestre PRI-94, se extrajo una muestra al azar de 20 estudiantes del primer semestre y se observó que su promedio en dicha asignatura fue de 13.2 puntos con una desviación de 2 puntos. El jefe de cátedra decide realizar un curso introductoria antes de cursar la materia para una mejor comprensión de la misma; para esto, selecciona 15 estudiantes del semestre SEG-94. Al finalizar el estudio de la asignatura, obtuvo los siguientes resultados un promedio en dicha asignatura de 15 puntos con una desviación de 3 puntos.

a) ¿existe alguna diferencia significativa entre los alumnos del PRI-94 y el promedio normal de la asignatura?

b) ¿fue efectivo el curso introductoria para los estudiantes del SEG-94 con respecto a los del PRI-94?. Use un nivel de confianza del 95%

RESPUESTAS:

a) Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, quiere decir que no existe diferencia significativa entre los promedios.

b) Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 95%, que fue efectivo el curso introductorio.

32.- En el pasado, la desviación típica de los pesos de ciertos paquetes de 40 onzas, llenados por una máquina era de 0.25 onzas. Una muestra al azar de 20 paquetes dio una desviación típica de 0.32 onzas. ¿Es significativo el incremento de variabilidad? Use un nivel de confianza del 95%

RESPUESTA:

Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 95%, que es significativo el incremento de variabilidad en el peso de los paquetes.

33.- Durante procesos de transfusión de sangre suelen ocurrir infecciones. Se condujo un experimento para probar si la inyección de anticuerpos reducía el riesgo de infección, dando los siguientes resultados: con anticuerpos 4 se infectaron y 78 no se infectaron, y a los que no se les administraron anticuerpos tuvieron infección 11 pacientes y no tuvieron infección 45. ¿A qué conclusión estadística le llevaría estos datos? Tome $\alpha = 0.12$

RESPUESTA:

Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de confianza de 88%, quiere decir que la inyección de anticuerpos reduce el riesgo de infección.

34.- Un nuevo estudio sugiere que la toma de aspirina protege de la formación de coágulos de sangre en las venas después de las operaciones. Así que se administran 4 pastillas de aspirina a 33 pacientes de los cuales 5 desarrollaron coágulos. A otro grupo de 55 personas en las mismas condiciones y no se les administró ninguna aspirina 14 presentaron coágulos. Estadísticamente hablando, ¿qué diría Ud.? Tome $\alpha = 0.05$

RESPUESTA:

Acepto H_o, Rechazo H_i , con un nivel de significación de 5%, quiere decir que tomar aspirina no protege de la formación de coágulos.

35.- Una encuesta ha revelado que el 20% de los niños de un barrio padecen cierto tipo de anemia. Se aplicó un tratamiento antianémico a un grupo de 80 niños de los que 18 niños no se curaron. ¿Qué piensa Ud. de este tratamiento antianémico estadísticamente hablando? Use un nivel de significación del 0.01.

RESPUESTA:

Acepto H_o, Rechazo H_i , con un nivel de significación de 1%, quiere decir que el tratamiento antianémico no fue efectivo.

36.- Actualmente en el mercado, existen varios medicamentos que logran disminuir la temperatura en casos de fiebres muy altas en un tiempo promedio de 2 horas. Se quiere probar la eficacia de una nueva droga que produzca el mismo efecto en menor tiempo. Se escogió una muestra de 25 personas con fiebre alta y se les suministró el nuevo medicamento y se observó que en un tiempo promedio de 1 hora y 15 minutos con una desviación de 7 minutos, se reduce la temperatura. ¿La nueva droga es más eficaz que los medicamentos actuales del mercado? Probar con un nivel de significación del 10%

RESPUESTA:

Rechazo H_o, Acepto H_i , se puede asegurar con un nivel de de significación de 10%, quiere que la nueva droga es más eficaz que los medicamentos actuales del mercado.

37.- Se desea saber si el ayuno afecta los resultados de algunos exámenes clínicos. Para ello, se escogen dos muestras de personas normales: la primera muestra consta de personas que respetaron las 14 horas de ayuno antes de practicarse el examen; y la segunda muestra, por personas que hicieron sus comidas cotidianas. Después de aplicar los exámenes se obtuvieron los siguientes resultados:

- ☞ En la primera muestra, de 80 personas, 50 exámenes tienen los valores dentro de los intervalos normales
- ☞ En la segunda muestra, de 90 personas, 45 exámenes tiene los valores dentro de los intervalos normales

¿Existe alguna diferencia significativa entre las dos muestras? ¿Influirá el ayuno en los resultados de los exámenes clínicos? Use un nivel de significación del 8%

RESPUESTA:
Acepto H_o, Rechazo H_i , con un nivel de significación de 8%, quiere decir que no existe diferencia significativa entre las dos muestras, por l oque el ayuno no influye en los resultados de exámenes clínicos.

38.- Dos tipos de soluciones químicas A y B, fueron ensayadas para determinar su pH (grado de acidez de la solución). Un análisis de 6 soluciones tipo A dieron un pH medio de 7.52 y una desviación de 0.24; y 5 soluciones tipo B dieron un pH promedio de 7.02 con una desviación de 0.32. ¿Habrá diferencia significativa entre las dos muestra? Use un nivel $\alpha = 0.05$

39.- Se desea comprobar la incidencia de dengue en las parroquias Antemano y Sucre. Para ello se toman dos muestras de 102 pacientes (una para cada parroquia), resultado que 28 pacientes en Antemano y 17 en Sucre, padecí8an la enfermedad. ¿Existe alguna diferencia significativa en la incidencia de dengue entre los pacientes de estas zonas? Utilice un nivel de significación del 3%

40.- La Oficina Sectorial de Laboratorios, desea comparar sus reportes de Hemoglobina Glicosilada (en %) con los de un laboratorio “A”. Para ello, proporciona al laboratorio “A” una muestra de 23 controles para que determinen el valor de Hemoglobina Glicosilada. Este laboratorio reportó valores con una media de 8.2% con una desviación típica de 1.9%; mientras que en la Oficina Sectorial de Laboratorios, anteriormente se reportaba una media de 7.2% con una desviación de 0.8%. ¿Considera usted que es significativo el aumento de variabilidad en reportes de Hemoglobina Glicosilada del Laboratorio “A” y la Oficina Sectorial? Use un nivel de significación del 4%

41.- Dos soluciones A y B, fueron ensayadas para determinar su densidad. Un total de 8 muestras de A dieron una densidad media de 1026 con una desviación de 23; y 6 muestras de B dieron una densidad media de 1016 con una desviación de 18. Determinar si existe diferencia significativa entre las dos soluciones, para un nivel de significación del 7%.

42.- Antes de comenzar un estudio sobre la capacidad de la Heparina para prevenir la broncoconstricción, se midieron valores de referencia de la función pulmonar de una muestra de 12 individuos con un historial de asma inducida. El

valor medio de la Capacidad Vital Forzada (CVF) de la muestra en un tiempo inicial tiene media de 4.49 litros con una desviación de 0.83 litros. Después de 10 minutos, se tomó otra medición obteniendo un valor medio de 3.71 litros con una desviación de 0.62 litros. El médico investigador considera que el tiempo es un factor que influye en la disminución del valor de la CVF. Se pide:

- ☞ Verifique si la hipótesis del médico es cierta con un nivel de confianza del 95%
- ☞ Construya el intervalo de confianza correspondiente al 90% de confianza.

43.- Se desea evaluar los resultados obtenidos por un laboratorio A. Se tomaron 25 muestras de adultos normales, y se determinaron los valores de plaquetas, obteniéndose un valor medio de $270x10^3/\mu l$, con una desviación estándar de $60x10^3/\mu l$. La Oficina Sectorial del laboratorio reporta una media de $145x10^3/\mu l$ con una desviación de $70x10^3/\mu l$. ¿Considera usted que es significativa la diferencia en la variabilidad de los reportes de plaquetas del laboratorio A y la Oficina Sectorial? Use un nivel de significación del 5%.

44.- El Instituto Nacional de Higiene, desea comparar sus reportes de Inmunofluorescencia (IFI) anti-leishmania con los de un laboratorio X. Para ello, proporciona al laboratorio X una muestra de 29 controles para que realicen el IFI. Este laboratorio reportó valores con una media de 2.023 con una desviación típica de 0.052, mientras que en el Instituto Nacional de Higiene, anteriormente se reportaba una media de 2.523 con una desviación típica de 0.095. ¿Considera Ud. que es significativa la disminución de variabilidad entre los reportes del IFI del laboratorio X y el Instituto Nacional de Higiene?. Use un nivel de significación del 10%

Ejercicios de Regresión y correlación lineal

1.- Un grupo de 12 sujetos fue sometido a la acción de una droga. En la siguiente tabla se recogen la dosis de droga en mg y la tensión arterial en mm de Hg:

Droga	18	20	24	19	24	25	20	20	30	18	19	18
Tensión	150	130	120	140	120	110	110	110	100	150	140	150

a) Determine el coeficiente de correlación. Interprete el resultado.
RESPUESTA: r = -0.78
b) ¿Qué tensión arterial corresponderá un tratamiento de 22mg?
RESPUETA: 125 mmHg
c) Dibuje el diagrama de dispersión y la recta de regresión encontrada.

2.- La siguiente tabla muestra la edad (X) y la presión sanguínea (mmHg) (Y) de 10 mujeres:

Edad	56	42	72	36	63	47	55	49	38	42
Presión Sanguínea	147	125	160	118	149	128	150	145	115	140

a) ¿Qué tipo de interacción existe entre estas dos variables? Explique su respuesta.
RESPUESTA: r = 0.89

b) Estimar la presión sanguínea de una mujer de 45 años.
RESPUESTA: 132 mmHg

3.- Interesa establecer la dependencia entre la dieta de comer carne y los mg de colesterol presentes en la sangre. Se escoge a 10 personas y se les somete a un control diario. Los resultados fueron los siguientes:

Carne(mg)	100	120	135	140	160	120	130	180	170	200
Colesterol(mg)	80	90	100	120	110	100	110	140	140	170

a) ¿Cuántos mg de colesterol se esperaría para una dosis de carne de 195mg?
RESPUESTA: 157mg

b) ¿Tiene alguna relación el consumo de carne y los mg de colesterol presentes en la sangre? ¿de qué tipo?

c) Dibuje el diagrama de dispersión y la recta de regresión encontrada

4.- La siguiente tabla muestra la edad (en años) y el tiempo de reacción (en seg) a un estímulo observado a una muestra al azar de 20 niños.

Edad	6	3	6	4	7	7	3	4	6	4	6	4	4	3	4	4	4	6	6	6
Tiempo	9	10	9	10	8	9	10	9	9	10	10	9	10	10	10	9	9	8	10	9

a) Construir el diagrama de dispersión.
b) ¿Qué tipo de interacción existe entre estas dos variables? Explique su respuesta
RESPUESTA: r = -0.69

c) Estimar el tiempo de reacción de un niño de 5 años
RESPUESTA: 9 seg

5.- Los datos de tensión arterial sistólica “pre” y “post” tratamiento, dados en cm. de Hg, de una muestra, son respectivamente:

X	16	16	17	17	18	18	19	19	19	20
Y	13	14	14	15	14	15	15	15	16	16

a) Calcular el coeficiente de correlación lineal de Pearson de las dos variables.
RESPUESTA: r = 0.83

b) Hallar la ecuación de la recta de mínimo cuadrados de las variables.
RESPUESTA: a = 0.57 b = 4.43

c) Dibuje la recta de mínimo cuadrados.

6.- La siguiente tabla muestra el tiempo que duraba un deportista en realizar un ejercicio físico y su respectiva cantidad de triglicéridos:

Trig (mg/dl)	43	65	78	73	71	69	67	45	69	60	65	59
Tiempo (min)	8.0	7.5	6.9	6.1	7.0	6.6	7.2	7.7	6.8	8.2	4.9	6.2

a) ¿Qué tipo de interacción existe entre estas dos variables? Explique su respuesta
RESPUESTA: r = -0.46

b) Estimar el tiempo que tardaría una persona cuya cantidad de triglicéridos es 73mg/dl

7.- Clínicamente se ha determinado que existe relación entre el peso y los niveles de glucosa en la sangre en personas que sufren de diabetes. Se quiere analizar el tipo de relación entre estas dos variables y para ello se selecciona un grupo de 14 diabéticos y se registraron los siguientes datos:

Peso(kg)	58	69	75	67	71	59	72	78	77	70	68	65	80	76
Glucosa	160	192	199	178	197	165	198	198	199	195	190	175	210	197

Se desea saber:

a) ¿Qué tipo de relación existe entre el peso y los niveles de glucosa? Justifique su respuesta

RESPUESTA: r = 0.94

b) Estime el nivel de glucosa que puede tener un diabético que pesa 90kgs.

RESPUESTA: No se puede estimar el nivel de glucosa de un diabético que pesa 90kgs con estos datos.

8.- Clínicamente se ha determinado que en casos de infarto al miocardio, existe relación entre las enzimas CK TOTAL y CKMB (fracción de la CK TOTAL). Se desea analizar qué tipo de relación cumplen estas variables en una muestra de 20 individuos que sufrieron de infarto al miocardio, según los siguientes resultados:

Ck total	250	178	456	201	85	62	45	697	100	1120	278	456	200	156	159	62	124	32
Ckmb	38	12	48	20	12	14	8	59	15	110	16	29	20	16	17	8	9	38

RESPUESTA: r = 0.97

9.- Los siguientes datos corresponden al peso y estatura de recién nacidos del Hospital J.M. de los Ríos en el mes de Enero de este año, los cuales se caracterizan por nacer con peso extremadamente bajo:

Estatura(cms)	41	40	38	30	45	32	33	37	39	43
Peso (grs)	1360	1490	1450	1320	1390	1120	1413	1215	1720	1750

a) ¿Qué tipo de relación existe entre el Peso y la Estatura? Justifique su respuesta

RESPUESTA: r = 0.52

b) Estime el peso de un recién nacido con estatura de 34 cms.

RESPUESTA: 1342 gms

10.- Clínicamente se ha determinado que en pacientes diabéticos (personas con niveles de glucosa por encima de 115mg/dl en sangre) existe relación entre la glucosa y la hemoglobina glucosilada (fracción de hemoglobina que transporta glucosa). Se desea analizar qué tipo de relación cumplen estas variables, en una muestra de individuos diabéticos, según los siguientes resultados:

Glucosa																

(mg/dl)	250	175	460	201	126	120	258	300	150	420	278	456	200	156	168	620
Hb-gli (%)	10.2	9.3	13.2	9.3	8.1	7.9	10.3	10.9	8.9	12.0	10.5	13.1	9.0	8.2	8.3	14.2

a) Explique la relación de dependencia existente entre las variables en estudio
b) ¿Qué tipo de interacción (si existe) se establece entre las variables?
c) Personas con valores de glucosa de 228mg/dl y 622mg/dL, ¿cuánto tendrá de hemoglobina glucosilada?

RESPUESTAS: para un nivel de glucosa de 228mg/dl se espera 8.9% de hemoglobina glucosilada. No se puede estimar el valor de Hemoglobina glucosilada de una persona con 622mg/dL de glucosa

11. Se quiere probar una nueva metodología para determinar transaminasa oxalacética (TGO) U/L , para ello se compara los resultados arrojados con muestras de pacientes corridas con el método nuevo y un método de referencia. A continuación se presentan los resultados:

Método de Referencia	180	342	90	200	201	400	500	30	40	62
Método de Prueba	176	345	88	197	205	405	498	32	45	59

Para que se acepte el nuevo método, se debe comprobar que el índice de correlación entre los resultados de los dos métodos es mayor de 0,95. ¿Se puede aceptar el método nuevo? Justifique su respuesta

ANEXOS