

ESTADÍSTICA desde cero

La estadística descriptiva es una gran parte de la estadística que se dedica a recolectar, ordenar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente las características de este. Este análisis es muy básico. Aunque hay tendencia a generalizar a toda la población

Albert Maguiña Rojas
Profesor de
Matemática y
Estadística..

SEMANA 1 – NOCIONES BÁSICAS DE ESTADÍSTICA

■ Definición

La ciencia que se ocupa de la recopilación, tabulación, análisis, interpretación y presentación de datos.

■ Población y muestra

Población es el conjunto de individuos, con alguna característica común, sobre el que se hace un estudio estadístico.

En la práctica es frecuente tener que recurrir a una muestra para inferir datos de la población. La **muestra** es un subconjunto de la población, seleccionada de modo que ponga de manifiesto las características de la misma, de ahí que la propiedad más importante de las muestras es su representatividad.

El proceso seguido en la extracción de la muestra se llama **muestreo**.



Ejemplo

Identificar la población y la muestra en la siguiente situación:

En una institución educativa se quiere saber la ocupación de los egresados de la última década. Para esto se convoca a una reunión de egresados y de los asistentes, se encuesta a diez egresados de cada año. Determina la población y la muestra.

Solución

Población. Todos los egresados de la última década.

Muestra. Los 100 estudiantes seleccionados, 10 de cada promoción.

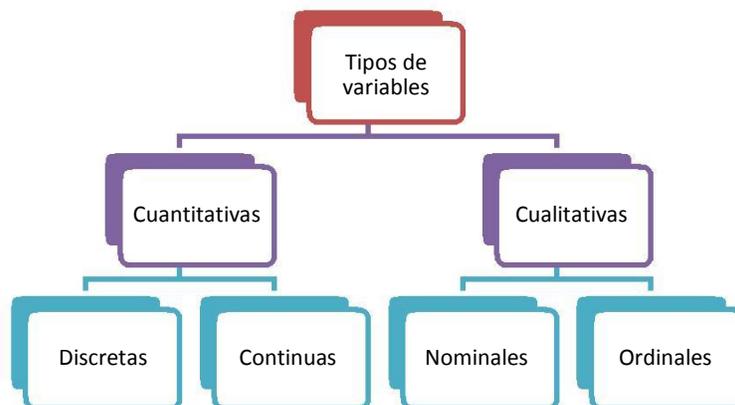
Luego de establecer la población y la muestra, es importante determinar las características a estudiar en la población: temperatura, peso, asistencia, gusto por algo, ocupación, etc. Estas características se denominan **variables** y se clasifican en cualitativas y cuantitativas.

■ Variables estadísticas

Las variables estadísticas pueden ser esencialmente de dos tipos **cualitativas y cuantitativas**.

Las variables cualitativas son las que no aparecen en forma numérica sino como una categoría o atributo.

Las variables cuantitativas son las que pueden expresarse numéricamente



Ejemplo

Tipo de variable	Definición	Ejemplo
Nominal	Está asociada a nombres.	Marca de auto, Sexo, Religión.
Ordinal	Tiene asociado un orden.	Nivel educacional, Estado nutricional, Nivel Socioeconómico.
Discreta	Sólo puede tomar un número finito (o contable) de posible valores.	El número de respuestas correctas en una prueba de 5 preguntas de V o F.
Continua	Puede tomar cualquier valor en un intervalo(s).	Cantidad de agua en un vaso de 50 ml.



Ejercitándose

Determine qué tipo son las siguientes variables. Si son variables cualitativas (nominal u ordinal) o cuantitativas (discretas o continuas).

- a) Marca de automóvil.
- b) Duración de una canción.
- c) Número de temas de un CD o DVD.
- d) Estado civil (soltero, casado, divorciado).
- e) Cantidad de lluvia en un año en Tacna (mm^3).
- f) Nivel educacional (básica, media, universitaria).
- g) Temperatura al mediodía en Tumbes (grados Celsius).

TRABAJO PRACTICO 1

- I. Determina la población y sugiere la muestra para cada una de las siguientes situaciones.
 - a. Una empresa de telefonía celular quiere realizar en la ciudad de Chimbote un estudio sobre el celular que prefieren los jóvenes entre 18 y 22 años de edad.
 - b. Una empresa de software quiere determinar cuál es el tiempo promedio que los jóvenes de la ciudad de Chimbote emplean en internet para diseñar un nuevo juego que se desarrolle en ese tiempo.
- II. Clasifica las siguientes variables en cualitativas o en cuantitativas.
 - a. El ingreso mensual de un trabajador.
 - b. El número de estudiantes clasificados por el grado que cursan.
 - c. El código de identificación de una persona en un centro médico.
 - d. Los números que llevan en sus camisetas los jugadores de un equipo.
 - e. Los números que indican las posiciones de llegada de los caballos en una carrera.

SEMANA 2 – ORGANIZACIÓN DE DATOS

Existen muchas formas de organizar los datos. Podemos sólo coleccionarlos y mantenerlos en orden; o si las observaciones están hechas con números, entonces podemos hacer una lista de los puntos de los datos, de menor a mayor según su valor numérico. Pero si los datos son trabajadores especializados (como carpinteros, albañiles o soldadores) de una construcción; o los distintos tipos de automóviles que ensamblan todos los fabricantes; o los diferentes colores de suéteres fabricados por una empresa dada, debemos de organizarlos de manera distinta. Necesitaremos presentar los puntos de datos en orden alfabético o mediante algún principio de organización. Una forma común de organizar los datos consiste en dividirlos en categorías o clases parecidas y luego contar el número de observaciones que quedan dentro de cada categoría. Este método produce una

Distribución de frecuencias.

■ Distribución de frecuencias

$X_i = \text{Variable}$	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$
x_1	f_1	f_1	$\frac{f_1}{n}$	h_1	$h_1 (100)$	I
x_2	f_2	$f_1 + f_2$	$\frac{f_1 + f_2}{n}$	$h_1 + h_2$	$h_2 (100)$	D
x	f	$f_1 + f_2 + f_3$	$\frac{f_1 + f_2 + f_3}{n}$	$h_1 + h_2 + h_3$	$h_3 (100)$	E
x	f	$f_1 + f_2 + f_3 + \dots$	$\frac{f_1 + f_2 + f_3 + \dots}{n}$	$h_1 + h_2 + h_3 + \dots$	$h_i (100)$	M
$n = \text{Total}$	$\sum f_i = n$	-	$\sum h_i = 1$	-	$\sum h_i \% = 100$	H_i



Ejemplo

Los siguientes datos representan el número de hijos por familia encuestada.

0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2								
2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5	6

A partir de estos datos, construya una tabla de frecuencias.

Solución:

Para construir la tabla de frecuencias hay que tener en cuenta que la variable en estudio es el número de hijos (discreta), que toma los valores existentes entre 0 y 6 hijos y las frecuencias son el conjunto de familias, de esta forma tenemos:

X_i	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$
0	2	2	0,04	0,04	4	4
1	4	6	0,08	0,12	8	12
2	21	27	0,42	0,54	42	54
3	15	42	0,30	0,84	30	84
4	6	48	0,12	0,96	12	96
5	1	49	0,02	0,98	2	98
6	1	50	0,02	1	2	100
$n = 50$	$\sum f_i = 50$	-	$\sum h_i = 1$	-	$\sum h_i \% = 100$	-



Ejercitándose

Se realizó un sondeo entre 25 miembros de una clase de psicología acerca del número de hermanos que tenían en sus familias. A partir de estos datos, elabore una tabla de distribuciones de frecuencias.

2	3	1	3	3
5	2	3	3	1
1	4	2	4	2
5	4	3	6	5
1	6	2	2	2

A menudo, los conjuntos de datos que contienen una gran cantidad de elementos se organizan en grupos o clases. Todos los datos son asignados a la clase que les corresponde; luego, se elabora una **Distribución de frecuencias para datos agrupados**. Estos intervalos (grupos o clases) tienen un punto medio que recibe el nombre de **marca de clase**. La marca de clase se obtiene calculando el promedio entre los límites inferior y superior de cada intervalo.



Ejemplo

$X_i = \text{Variable}$	$Y_i = \text{marca de clase}$	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$
$[x_1 - x_2 >$	$y_1 = \frac{x_1 + x_2}{2}$	I	I	I	I	I	I
$[x_2 - x_3 >$	$y_2 = \frac{x_2 + x_3}{2}$	D	D	D	D	D	D
$[x_3 - x_4 >$	$y_3 = \frac{x_3 + x_4}{2}$	E	E	E	E	E	E
$[x_{n-1} - x_n >$	$y_n = \frac{x_{n-1} + x_n}{2}$	M	M	M	M	M	M
n = Total	-	$\sum f_i = n$	-	$\sum h_i = 1$	-	$\sum h_i \% = 100$	-



Ejercitándose

En una prueba tomada a 50 alumnos, se registraron los siguientes puntajes:

83	82	87	64	63	75	83	62	67	83
68	85	66	61	83	76	83	67	78	76
83	72	70	84	71	77	82	79	83	72
77	74	67	80	84	75	73	75	83	84
77	72	89	80	87	77	63	72	84	78

Antes de construir la tabla de frecuencias debemos construir los intervalos. Para ello, debemos tener en cuenta las siguientes indicaciones.

❖ Determinar la número de intervalos, para ello utilizaremos la **Regla de Sturges**: $m = 1 + 3,3 \log n$

❖ Determinar la amplitud de los intervalos, para ello utilizaremos la siguiente regla: $C = \frac{x_{\text{máx}} - x_{\text{mín}}}{m}$

En nuestro ejemplo, $n = 50 \wedge x_{\text{máx}} = 89 \wedge x_{\text{mín}} = 61$ En consecuencia:

$$m = 1 + 3,3 \log 50 \rightarrow m = 6,6 \rightarrow m = 7 \qquad C = \frac{89 - 61}{7} \Rightarrow C = 4$$

X_i	Y_i	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$
[61 – 65 >	63	5	5	0,10	0,10	10	10
[65 – 69 >	67	5	10	0,10	0,20	10	20
[69 – 73 >	71	6	16	0,12	0,32	12	32
[73 – 77 >	75	7	23	0,14	0,46	14	46
[77 – 81 >	79	9	32	0,18	0,64	18	64
[81 – 85 >	83	14	46	0,28	0,92	28	92
[85 – 89]	87	4	50	0,08	1	8	100
Total	-	$\sum f_i = 50$	-	$\sum h_i = 1$	-	$\sum h_i \% = 100$	-



Ejercicio: los siguientes datos representan las calificaciones de 50 estudiantes de la facultad de derecho en el curso de estadística.

12 7 17 10 13 15 17 13 11 15
 14 16 7 10 7 13 15 9 14 11
 16 8 12 18 11 15 12 11 9 15
 8 16 11 10 14 17 13 8 15 12
 19 13 20 13 15 8 13 19 18 18

A partir de esta información construya una tabla de distribuciones.

TRABAJO PRACTICO 2

- I. En una empresa, se hizo el estudio sobre las edades de los empleados y se obtuvo la siguiente tabla:

Edades	N° de empleados
[20 – 25 >	12
[25 – 30 >	15
[30 – 35 >	23
[35 – 40 >	11
[40 – 45 >	9

Donde “A” es el porcentaje de empleados con 30 años o más; “B” es el porcentaje de empleados con menos de 40 años. Calcular el valor de “A + B”

- II. La siguiente tabla muestra el número de jóvenes que obtuvieron los puntajes señalados en una prueba de ingreso.

Puntaje	N° de jóvenes
[10 – 15 >	10
[15 – 20 >	15
[20 – 25 >	28
[25 – 30 >	20
[30 – 35 >	17

Donde “A” es el porcentaje de jóvenes con puntajes mayores a 20; “B” es el porcentaje de jóvenes con puntajes menores a 15. Halle el valor de “A – B”

- III. Para cada una de las tablas, calcular lo siguiente:

a) $\frac{f_4 + F_3}{H_4}$

b) $\frac{H_2\% + h_4\%}{h_2}$

c) $\frac{H_3\% + h_2\%}{Y_3}$

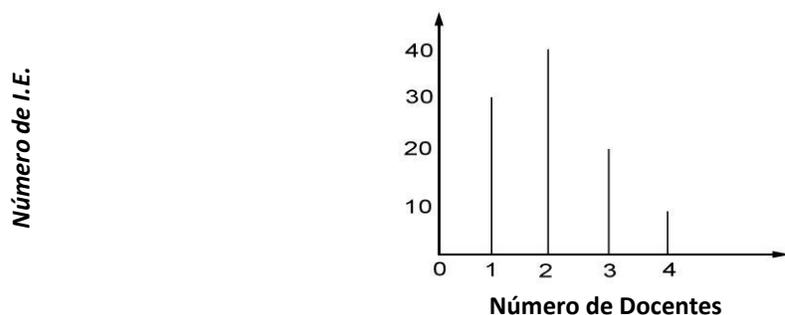
d) $\frac{f_2 + F_1}{H_5}$

SEMANA 3 – PRESENTACIÓN DE DATOS

Las tablas de frecuencias de los datos estadísticos muestran una información ordenada del hecho que se analiza y estudia. Además de esta forma de presentación es útil conocer la forma de presentarlos gráficamente para obtener una apreciación global, rápida y visual de la información señalada.

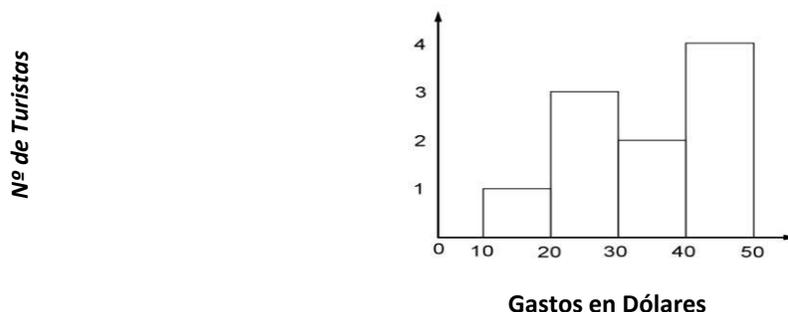
■ *Gráfico de Bastones*

Se utiliza para describir datos cuando la variable es discreta. Su construcción se hace levantando segmentos perpendiculares al eje de la variable y con una altura proporcional a su frecuencia absoluta o relativa porcentual.



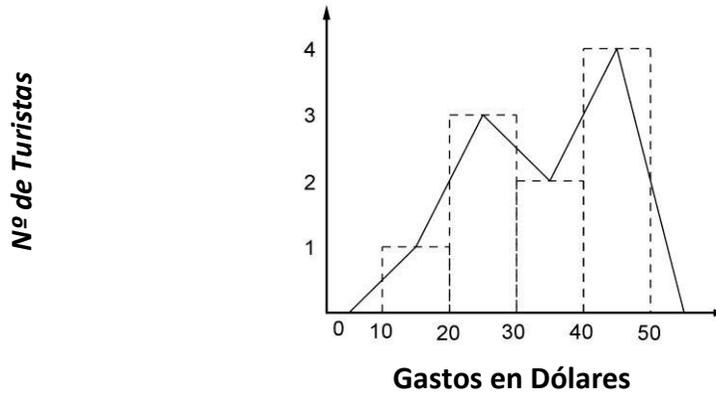
■ *Histograma de Frecuencias*

Se utiliza para describir datos cuando la variable es continua. Su construcción se hace levantando sobre el eje de la variable rectángulos que tengan por base la amplitud del intervalo de clase y una altura proporcional a su frecuencia absoluta o relativa porcentual.



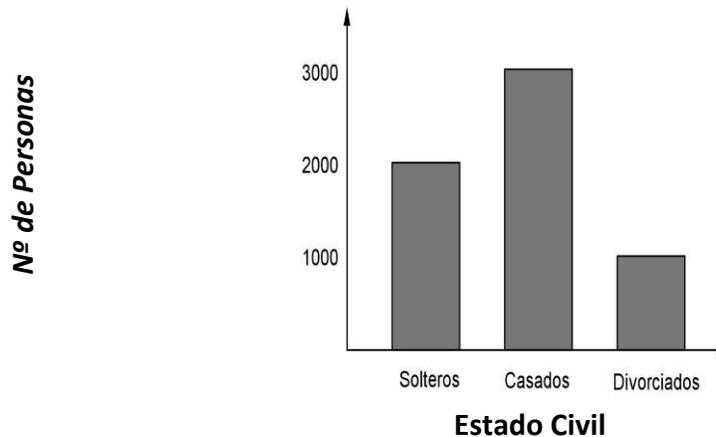
■ **Polígono de Frecuencias**

Se utiliza también para describir datos cuando la variable es continua. Su construcción se hace uniendo los puntos medios superiores de los rectángulos en el histograma.



■ **Gráfico de Barras**

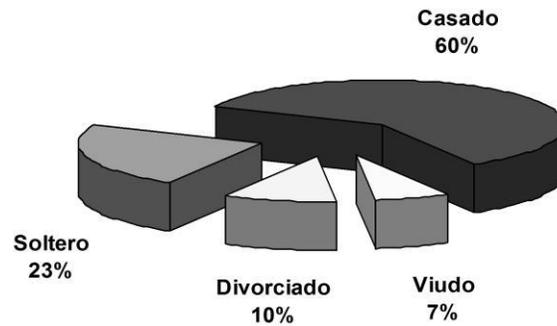
Se utiliza para describir datos cuando la variable es cualitativa. Su construcción se hace levantando barras proporcionales a su frecuencia absoluta o relativa porcentual.



■ **Gráfico de Sectores Circulares**

Se utiliza también para describir datos cuando la variable es cualitativa. Se usa frecuentemente cuando se desea comparar cada categoría de la variable con respecto al total. Para su elaboración se utiliza una circunferencia, siendo

necesario que los valores absolutos y/o porcentuales sean traducidos en grados sexagesimales.



TRABAJO PRACTICO 3

- I. Los siguientes datos corresponden a una muestra de pequeñas empresas según su número de trabajadores afiliados al sistema privado de pensiones.

1	2	5	4	1	3
3	4	4	4	3	5
3	3	4	5	4	4
3	3	4	3	5	3
2	2	3	3	3	2

Construir un Gráfico y Comentar

- II. Los siguientes datos corresponden a una muestra aleatoria de los gastos semanales en dólares de 20 turistas que se alojaron en el hotel “El Delfín” de la ciudad de Lima.

400	500	550	600
680	750	780	630
640	650	700	740
750	800	850	750
1000	890	850	950

Construir un Gráfico y Comentar

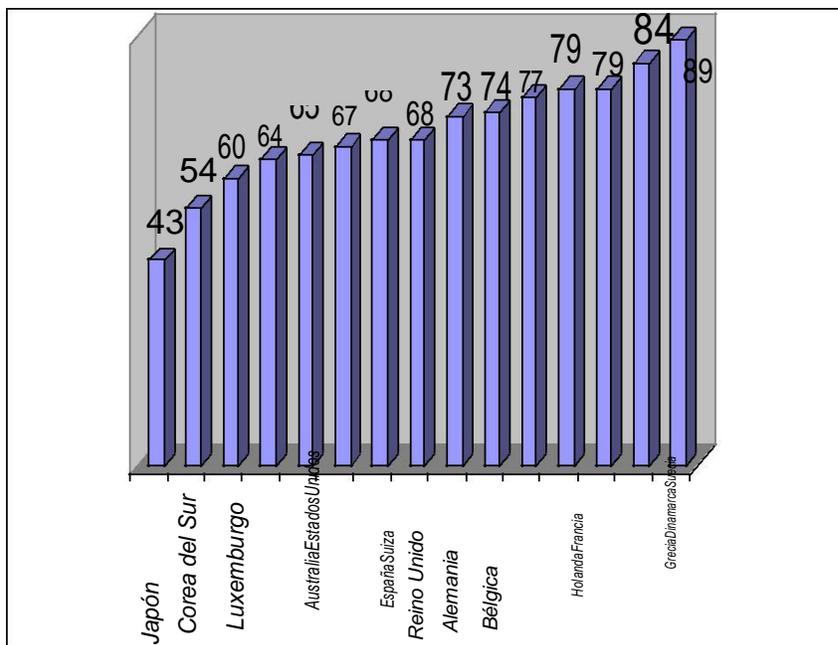
- III. Los siguientes datos corresponden a una muestra aleatoria de 30 docentes de la Universidad Católica los Ángeles según su estado civil del semestre 2010 – I

S	C	S	C	C	D	V	C	D	C
C	D	C	C	C	S	V	C	S	S
C	C	S	S	C	C	C	C	C	C

a) Construir un Gráfico y Comentar

- IV. En un estudio sobre deudas incobrables entre los trabajadores estatales de cierta urbanización de Chimbote, se determinó que a través de los años, la dinámica de dicho fenómeno fue la siguiente: 1985: 10 hombres y 5 mujeres. 1990: 20 hombres y 16 mujeres. 1995: 45 hombres y 32 mujeres, 2000: 80 hombres y 45 mujeres. Se pide: Presentar la información en una tabla de frecuencias e indicar la escala de medición de la variable utilizada.

- V. En el siguiente estudio se analizan los sueldos que ganan las mujeres en la industria en diversos países del mundo, en porcentaje sobre lo que gana los hombres:



mujer en Suiza gana 1300
¿cuánto gana un hombre
mismo puesto y con la
categoría profesional?

en promedio, gana
en un sueldo mensual de
euros netos. ¿Cuánto si
fuese mujer?

SEMANA 4 – MEDIDAS DE TENDENCIA CENTRAL

A veces, de los datos recolectados ya organizados en alguna de las formas vistas en capítulos anteriores, se desea encontrar una especie de punto central en función de sus frecuencias. En Estadística se conocen tres diferentes, llamadas **Medidas de Tendencia Central**, cuya utilización varía de acuerdo con lo que se desee del conjunto de datos recolectados. Esas tres medidas de tendencia central son **la media, la mediana y la moda**.

Cada una de ellas se estudiará en dos partes: primero, cuando los datos están organizados en tablas de distribución de frecuencias simples y, segundo, cuando están organizados en intervalos. Además, a veces difieren las fórmulas para calcular alguna de ellas si se trata de poblaciones o de muestras. En caso de que no se diga nada, deberá entenderse que la fórmula es la misma para ambas.

■ La Media

La media, llamada también *media aritmética*, es la medida de tendencia central conocida popularmente como “promedio”.

■ La media para frecuencias simples

Cuando los datos recolectados han sido organizados en una tabla de distribución de frecuencias simples, la media, para poblaciones como para muestras, se puede calcular por medio de la fórmula

$$\bar{x} = \frac{\sum f_i x_i}{n}$$

✓ Ejemplo

Calificaciones	f_i		Calificaciones	f_i
1	3		6	17
2	3		7	22
3	6		8	10
4	8		9	6
5	9		10	5

Calificaciones	f_i	$f_i \cdot x_i$
1	3	3
2	3	6
3	6	18
4	8	32
5	9	45
6	17	102
7	22	154
8	10	80
9	6	54
10	5	50
Total	89	544

Tenemos:

$$n = \sum f_i = 89$$

$$\sum f_i x_i = 544$$

$$\bar{x} = \frac{\sum f_i x_i}{n}$$

$$\bar{x} = \frac{544}{89} \Rightarrow \bar{x} = 6,11$$

■ La media para frecuencias por intervalos

Cuando los datos recolectados han sido organizados en una tabla de frecuencias por intervalos, la media para poblaciones como para muestras se puede calcular por medio de la fórmula

$$\bar{x} = \frac{\sum f_i y_i}{n}$$

✓ Ejemplo

Intervalos	f_i
[0 – 2]	12
[3 – 5]	13
[6 – 8]	23
[9 – 11]	16
[12 – 14]	18
Total	82

Intervalos	f_i	y_i	$f_i \cdot y_i$
[0 – 2]	12	1	12
[3 – 5]	13	4	52
[6 – 8]	23	7	161
[9 – 11]	16	10	160
[12 – 14]	18	13	234
Total	82	-	619

Tenemos:

$$n = \sum f_i = 82$$

$$\sum f_i y_i = 619$$

$$\bar{x} = \frac{\sum f_i y_i}{n}$$

$$\bar{x} = \frac{619}{82} \Rightarrow \bar{x} = 7,55$$

■ La Mediana

La mediana es la medida de tendencia central que se define como aquel valor nominal que tiene, dentro de un conjunto de datos ordenados, arriba y abajo de él, el mismo número de datos nominales. En otras palabras, es el dato que está a la mitad, es el dato que divide en dos partes iguales a un conjunto de datos.

■ La Mediana para frecuencias simples

Cuando los datos recolectados han sido organizados en una tabla de distribución de frecuencias simples, la mediana, para poblaciones como para muestras, se puede calcular por medio de la fórmula

$$F_i > \frac{n}{2} \Rightarrow x_i \in F_i \Rightarrow x_i = Me$$



Ejemplo

X_i	f_i	F_i	
0	1	1	
1	1	2	
2	3	5	
3	5	10	
4	6	16	
5	7	23	
6	11	34	
7	15	49	
8	25	74	$F_i > \frac{117}{2} \Rightarrow F_i > 58,5 \Rightarrow \boxed{F_i = 74}$
9	20	94	
10	23	117	
Total	117	-	$x_i \in F_{74} \Rightarrow x_i = 8 \Rightarrow \boxed{Me = 8}$

■ La Mediana para frecuencias por intervalos

Cuando los datos recolectados han sido organizados en una tabla de frecuencias por intervalos, la mediana para poblaciones como para muestras se puede calcular por medio de la fórmula

$$Me = L_{inf} + \left(\frac{\frac{n - F_{i-1}}{2}}{J_i} \right) (C) \Leftrightarrow \boxed{F_i > \frac{n}{2} \Rightarrow [L_{inf} - L_{sup}] \in F_i \Rightarrow Me \in [L_{inf} - L_{sup}]}$$

✓

Ejemplo

$L_{inf} - L_{sup}$	f_i	F_i		
[1 - 30]	1	1		
[31 - 60]	1	2		
[61 - 90]	3	5	$F_i > \frac{49}{2} \Rightarrow F_i > 24,5 \Rightarrow F_i = 34$ $[181 - 210] \in (F_i = 34)$ $Me \in [181 - 210]$	$Me = 181 + \left(\frac{24,5 - 23}{11} \right) (29)$ $Me = 184,95$
[91 - 120]	5	10		
[121 - 150]	6	16		
[151 - 180]	7	23		
[181 - 210]	11	34		
[211 - 240]	15	49		
Total	49	-		

■ La Moda

La moda es la medida de tendencia central que se define como aquel valor nominal que tiene la frecuencia mayor. Por lo tanto, una distribución de frecuencias puede tener más de una moda o, inclusive, no tener moda cuando todos los datos tienen frecuencia 1.

■ La Moda para frecuencias simples

Cuando los datos recolectados han sido organizados en una tabla de distribución de frecuencias simples, la moda, para poblaciones como para muestras, se puede calcular por medio de la fórmula

$$f_i > f_x \Rightarrow x_i \in f_i \Rightarrow Md = x_i$$

✓ **Ejemplo**

X_i	f_i		
1	8	$f_i > f_x$ $17 > f_x \Rightarrow f_i = 17$	$x_i \in f_i$ $x_i \in 17 \Rightarrow x_i = 4$ $Me = x_i \Rightarrow Me = 4$
2	5		
3	13		
4	17		
5	10		
6	7		
Total	60		

■ **La Moda para frecuencias por intervalos**

Cuando los datos recolectados han sido organizados en una tabla de frecuencias por intervalos, la moda para poblaciones como para muestras se puede calcular por medio de la fórmula

$$Md = L_{inf} + \left(\frac{a_1}{d_1 + d_2} \right) (C) \Leftrightarrow \boxed{f_i > f_x \Rightarrow [L_{inf} - L_{sup}] \in f_i \Rightarrow Md \in [L_{inf} - L_{sup}]}$$

Donde : $d_1 = f_i - f_{i-1} \wedge d_2 = f_i - f_{i+1}$

✓ **Ejemplo**

$L_{inf} - L_{sup}$	f_i		
[10 - 20 >	8	$17 > f_x \Rightarrow [40 - 50 > \in 17$ $Md \in [40 - 50 >$ $\left\{ \begin{array}{l} d_1 = 17 - 13 \rightarrow d_1 = 4 \\ d_2 = 17 - 10 \rightarrow d_2 = 7 \end{array} \right.$	$Md = 40 + \left(\frac{4}{4 + 7} \right) 10$ $Md = 40 + \left(\frac{40}{11} \right)$ $Md = 40 + (3,6)$
[20 - 30 >	5		
[30 - 40 >	13		
[40 - 50 >	17		
[50 - 60 >	10		
[60 - 70 >	7		
Total	60		

TRABAJO PRACTICO 4

- I. Halle la Media, la Mediana y la Moda de los siguientes datos:

X_i	f_i
0	1
1	4
2	7
3	6
4	2
Total	20

- II. Halle la Media, la Mediana y la Moda de los siguientes datos:

$L_{inf} - L_{sup}$	f_i
[26 – 34 >	1
[34 – 42 >	2
[42 – 50 >	4
[50 – 58 >	10
[58 – 66 >	16
[66 – 74 >	8
[74 – 82 >	3
[82 – 90 >	7
Total	51

- III. De las edades de cuatro personas, se sabe que la media es igual a 24 años, la mediana es 23 y la moda es 22. Encuentre las edades de las cuatro personas.
- IV. Al calcular la media de 125 datos, resultó 42. Un chequeo posterior mostró que en lugar del valor 12,4 se introdujo 124. Corregir la media.
- V. De una central telefónica salieron 70 llamadas de menos de 3 minutos, promediando 2,3 minutos; 40 llamadas de menos de 10 minutos pero no menos de 3 minutos, promediando 6,4 minutos y 10 llamadas de al menos 10 minutos, promediando 15 minutos. Calcular la duración promedio de todas las llamadas.

SEMANA 5 – MEDIDAS DE DISPERSIÓN

La media es un buen indicador de la tendencia central de un conjunto de valores, pero no da la información completa de los datos. Para ver por qué, compare la distribución A con la distribución B en la tabla que se muestra a continuación:

	A	B
	5	1
	6	2
	7	7
	8	12
	9	13
Media	7	7
Mediana	7	7

Ambas distribuciones de números tienen la misma media (y también la misma mediana), pero fuera de esto, son completamente diferentes. En la primera, el 7 es un valor característico aceptable, pero en la segunda, la mayor parte de los valores difieren bastante de 7. Lo que se necesita aquí es alguna medida de **dispersión**, de los datos.

Una de las medidas más útiles de dispersión es la **desviación estándar**, la cual se basa en las **desviaciones de la media** que presentan los datos. Para calcular la desviación estándar se aplica la siguiente fórmula:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$



Ejemplo

Encuentre la desviación estándar muestral para la siguiente distribución de frecuencias.

X_i	f_i
2	5
3	8
4	10
5	2

X_i	f_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	5	$\bar{x} = \frac{2(5) + 3(8) + 4(10) + 5(2)}{25}$ $\bar{x} = 3,36$	$2 - 3,36 = -1,36$	1,8496
3	8		$3 - 3,36 = -0,36$	0,1296
4	10		$4 - 3,36 = 0,64$	0,4096
5	2		$5 - 3,36 = 1,64$	2,6896
Total	25		-	-

Pero como los datos no son únicos, es decir, hay repeticiones por cada dato (frecuencias), se tendrá que multiplicar cada valor obtenido por su frecuencia respectiva. Esto es:

$(x_i - \bar{x})^2 \cdot f_i$	$\sum (x_i - \bar{x})^2 f_i$ $9,2480 + 1,0368 + 4,0960 + 5,3792$ $\sum (x_i - \bar{x})^2 f_i = 19,76$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{n - 1}}$ $s = \sqrt{\frac{19,76}{25 - 1}}$ $s = 0,91$
1,8496 (5) = 9,2480		
0,1296 (8) = 1,0368		
0,4096 (10) = 4,0960		
2,6896 (2) = 5,3792		
-		

En el caso de que los datos sean agrupados (con intervalos) se utilizará la marca de clase como el valor que representará a cada dato.

TRABAJO PRACTICO 5

- I. Hallar la desviación estándar para cada una de las siguientes distribuciones.

X_i	f_i
0	2
1	4
2	21
3	15
4	6
5	1
6	1

X_i	f_i
[61 – 65 >	5
[65 – 69 >	5
[69 – 73 >	6
[73 – 77 >	7
[77 – 81 >	9
[81 – 85 >	14
[85 – 89]	4

- II. Calcular la desviación típica (estándar) del dinero que gastan mensualmente 30 alumnos de 4° cuyos datos se han recogido en la siguiente distribución:

Intervalo	5 – 9	9 – 13	13 – 17	17 – 21	21 – 25
Frecuencia	10	8	5	4	3

- III. Una aplicación a nuestro mundo real: **Marketing**



La empresa *Gloria* comercializa mucho tres de sus productos a nivel nacional. Uno de los objetivos fundamentales de la publicidad de cada producto consiste en lograr que los consumidores reconozcan que *Gloria* es la que elabora el producto. Para medir qué tan bien cada anuncio publicitario logra tal reconocimiento, se le pidió a un grupo de consumidores que identificara lo más rápido posible a la compañía responsable de una larga lista de productos. El primer producto de *Gloria* obtuvo un tiempo promedio, antes de ser reconocido, de 2,5 segundos, con una desviación estándar de 0,004 segundos. El segundo producto obtuvo un tiempo promedio, antes de ser reconocido, de 2,8 segundos, con una desviación estándar de 0,006 segundos. El tercer producto obtuvo un tiempo promedio, antes de ser reconocido, de 3,7 segundos, con una desviación estándar de 0,09 segundos. ¿Para cuál de los productos estuvo el consumidor más alejado del desempeño promedio?

SEMANA 6 – PROBABILIDADES

En general, la probabilidad es la posibilidad de que algo pase. Las probabilidades se

expresa como fracciones $\left(\frac{1}{6}, \frac{1}{2}, \frac{8}{9}\right)$ o como decimales (0,167 ; 0,500 ; 0,889) que están entre

cero y uno. Tener una probabilidad de cero significa que algo nunca va a suceder; una probabilidad de uno indica que algo va a suceder siempre.

En la teoría de probabilidad, un **evento** es uno o más de los posibles resultados de hacer algo. Al lanzar una moneda al aire, si cae cruz es un **evento**, y si cae cara es otro. De manera análoga si sacamos una carta de un mazo de naipes, el tomar el as de espadas es un **evento**. Un ejemplo de evento que quizás esté más cercano a su quehacer diario es ser elegido de entre 40 estudiantes para que responda una pregunta. Cuando escuchamos las pocas gratas predicciones del índice de mortalidad en accidentes de tránsito, esperamos no ser uno de tales **eventos**.

En la teoría de probabilidad, la actividad que origine uno de dichos eventos se llama **experimento**. Utilizando un lenguaje formal, podríamos hacer la siguiente pregunta: ¿En un **experimento** de lanzar una moneda, cuál es la probabilidad del evento cara? Y desde luego, si la moneda no está cargada y tiene la misma probabilidad de caer en cualquiera de sus dos lados (sin posibilidades de que caiga parada), podríamos responder, $\frac{1}{2}$ - ó 0,5

Al conjunto de todos los resultados posibles de un experimento se llama **espacio muestral**

del experimento. En el de lanzar una moneda, el **espacio muestral** es: $S = \{cara, cruz\}$

■ Teorema de Laplace

Este teorema debe usarse sólo para eventos equiprobables (cada evento tiene la misma posibilidad de que suceda). Laplace menciona que la probabilidad de un evento está dada por la siguiente fórmula:

$$P(x_i) = \frac{\text{Casos Favorables}}{\text{Casos Posibles}}$$

✓ **Ejemplo**

Imaginemos que deseamos obtener un número impar al lanzar un dado. ¿Cuáles serían los casos favorables y cuáles los casos posibles? Naturalmente, tendríamos:

Casos favorables: {1,3,5}

Casos posibles: {1, 2,3, 4,5,6}

Entonces, la probabilidad de obtener un número impar al lanzar un dado es:

$$P(x_i : \text{impar}) = \frac{3}{6} \rightarrow P(x_i : \text{impar}) = \frac{1}{2} \text{ Formalizando:}$$

Experimento: Lanzar un dado.

Espacio muestral: {1, 2, 3, 4, 5, 6}

Evento: obtener un # impar, {1, 3, 5}

✓ **Ejercitándose**

Resuelva los siguientes problemas, justificando su razonamiento.

- La probabilidad de extraer una bola roja de una caja es $\frac{1}{3}$ ¿Cuál es la probabilidad de sacar una bola que no sea roja?
- Se lanzan dos dados. ¿Cuál es la probabilidad de que sumen 3 ó 4?
- Una rueda está dividida en 8 sectores iguales, numeradas del 1 al 8. ¿Cuál es la probabilidad de obtener un número impar y mayor que 3?
- Se tienen 10 fichas con los números 44, 44, 45, 46, 46, 46, 47, 48, 49. ¿Cuál es la probabilidad de sacar una ficha con un número mayor que 46?
- En una caja hay 50 fichas de igual peso y tamaño. 12 son rojas, 20 son cafés y 18 son amarillas. ¿Cuál es la probabilidad de sacar una roja, una café, una amarilla y nuevamente una roja, en ese orden y sin reposición?
- Se depositan en una caja tarjetas del mismo tipo con las letras de la palabra HERMANITOS, luego se saca de la caja una tarjeta al azar, la probabilidad de que en ésta esté escrita una vocal es:

TRABAJO PRACTICO 6

- I. Resuelva cada uno de los problemas que se le propone. No olvide que debe justificar cada uno de sus razonamientos.

- ❖ Una máquina produce 100 tornillos de los que 3 son defectuosos. Si se cogen dos tornillos, halla la probabilidad de que al coger el segundo sea defectuoso, con la condición de que el primero también haya sido defectuoso.
- ❖ Una familia tiene tres hijos. Hallar la probabilidad de que los tres sean varones.
- ❖ Se extrae una bola de una urna que contiene 6 bolas rojas y 4 verdes, se observa si ha sido roja y se vuelve a introducir; luego se extrae otra bola. ¿Cuál es la probabilidad de que las dos sean rojas?
- ❖ Se extraen de una vez dos bolas de una urna que contiene 6 bolas rojas y 4 verdes. ¿Cuál es la probabilidad de que las dos sean rojas?

II. Una aplicación a nuestro mundo real: ***El dilema del prisionero***

- ❖ En una cárcel hay 3 prisioneros (A; B; C) con historiales similares. En un momento dado, los tres solicitan el indulto a un tribunal, y sin conocerse más detalles llega la información al prisionero A de que han concedido el indulto a 2 de los 3 prisioneros. El prisionero A conoce a uno de los miembros del tribunal y puede intentar hacerle una pregunta para obtener algo de información. Sabe que no puede preguntar si él es uno de los dos indultados, pero si puede pedir que le den el nombre de uno de los otros dos (nunca él) que esté indultado. Pensando un poco concluye que si no hace tal pregunta, entonces la probabilidad de ser uno de los dos indultados es $2/3$, mientras que si la hace obtendrá respuesta y entonces la probabilidad de ser el otro indultado es $1/2$. Por ello, concluye que es mejor no hacer tal pregunta, porque sea cual sea la respuesta, sólo le servirá para disminuir la probabilidad de ser uno de los dos indultados. ¿Dónde está el error de su razonamiento?

SEMANA 7 – PROBABILIDAD BINOMIAL

Se denomina prueba o ensayo de Bernoulli a todo experimento aleatorio que consiste de solo dos resultados posibles mutuamente excluyentes, generalmente llamados: éxito (E) y fracaso (F). Por ejemplo, son ensayos de Bernoulli, lanzar una moneda al aire con los resultados: cara o sello. Elegir al azar un objeto fabricado, con los resultados: defectuoso o no defectuoso. Se denomina **Experimento Binomial** a un número fijo “ n ”, de repeticiones independientes de un experimento aleatorio de Bernoulli. Este se caracteriza por:

1. Las n pruebas son estadísticamente independientes.
2. Los resultados de cada prueba son dos, mutuamente excluyentes: éxito y fracaso.
3. La probabilidad p de éxito es invariante en cada una de las pruebas.

Para calcular la probabilidad Binomial, se usa la siguiente fórmula:

$$P(X=k) = \frac{\binom{n}{k} p^k q^{n-k}}{1}$$

Ejemplo

La última novela de un autor ha tenido un gran éxito, hasta el punto de que el 80% de los lectores ya la han leído. Un grupo de 4 amigos son aficionados a la lectura.

a) ¿Cuál es la probabilidad de que el grupo hayan leído la novela 2 personas?

Tenemos los siguientes datos: $n = 4$ $k = 2$ $p = 0,8$ $q = 0,2$

$$P(X=2) = \binom{4}{2} (0,8)^2 (0,2)^2 \rightarrow P(X=2) = 0,1536$$

b) ¿Cuál es la probabilidad de que el grupo hayan leído la novela a lo más 2 personas?

$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$$

$$\left[\binom{4}{2} (0,8)^2 (0,2)^2 + \binom{4}{3} (0,8)^3 (0,2)^1 + \binom{4}{4} (0,8)^4 (0,2)^0 \right] = 0,9728$$



Ejercitándose

- a) Se afirma que el 30% de la producción de ciertos instrumentos se realiza con material nacional y los demás con material importado. Si se toma una muestra aleatoria con reemplazo de 25 de estos instrumentos, calcular:
- ¿Cuál es la probabilidad de que 3 de ellos sean de material nacional?
 - ¿Cuál es la probabilidad de que no más de 3 de ellos sean de material nacional?
 - ¿Cuál es la probabilidad de que al menos 3 sean de material nacional?
 - ¿Cuántos instrumentos fabricados con material nacional se espera encontrar en la muestra?
- b) Un examen consta de 6 preguntas con 4 posibles respuestas cada una, de las que sólo una de ellas es correcta. Un estudiante que no se había preparado la materia responde completamente al azar marcando una respuesta aleatoria. Calcular la probabilidad de que acierte 4 o más preguntas.
- c) La probabilidad de que un cazador novato cobre una pieza es 0,4. Si lo intenta 5 veces, calcula la probabilidad de que cobre una pieza al menos 3 veces.
- d) El 53% de los trabajadores de una determinada empresa son mujeres. Si elegimos 8 personas de esa empresa al azar, calcula la probabilidad de que:
- Haya más de 6 mujeres
 - Hallar la media y la desviación estándar.
- e) Un examen tipo test consta de 100 preguntas, cada una de las cuales se acompaña de cuatro respuestas, una de ellas correcta y erróneas las otras tres. Si un estudiante contesta al azar, ¿cuál es la probabilidad de que acierte más de 30 preguntas? ¿y menos de 15?

TRABAJO PRACTICO 7

- I. Una urna contiene 6 bolas con números pares y 9 bolas con números impares. Si hacemos diez extracciones con reemplazamiento, calcula la probabilidad de obtener número impar:
 - Alguna vez
 - Más de 8 veces

- II. La probabilidad de que un determinado juguete salga defectuoso es de 0,03. Calcular la probabilidad de que en un lote de 60 de estos juguetes haya:
 - Alguno defectuoso
 - Menos de dos defectuosos

- III. La probabilidad de que un cierto experimento tenga éxito es 0,4. Si repetimos el experimento 15 veces, calcular la probabilidad de que tenga éxito:
 - Alguna vez
 - Menos de dos veces

- IV. Lourdes Flores es la alcaldesa de una ciudad grande. Últimamente, se ha estado preocupando acerca de la posibilidad de que grandes cantidades de personas que cobran el seguro de desempleo en realidad tengan un trabajo en secreto. Sus asistentes estiman que el 40% de los beneficiarios del seguro de desempleo entran en esta categoría. Pero la señora Flores no está convencida. Le pide a uno de sus ayudantes que haga una investigación de 10 beneficiarios del seguro tomados al azar, a partir de esta información determine:
 - Si los asistentes de la alcaldesa tienen razón, ¿cuál es la probabilidad de que los individuos investigados tengan un empleo? ¿Cuál es la probabilidad de que sólo tres de los individuos investigados tengan trabajo?

SEMANA 8 – PRUEBA CHI CUADRADO – χ^2

La **Prueba Chi Cuadrado** es un instrumento que se utiliza para determinar el grado de dependencia que existe entre dos variables cualitativas.

Esta prueba está dada por la siguiente fórmula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Donde: $\begin{cases} f_o = \text{Una frecuencia observada} \\ f_e = \text{Una frecuencia esperada} \end{cases}$

Proceso estadístico

Para aplicar la Prueba Chi Cuadrado se debe tomar en cuenta los siguientes pasos:

1. Formulación de la hipótesis

H_0 = Hipótesis nula (no existe dependencia) H_1

= Hipótesis alternativa (existe dependencia)

2. Nivel de significancia

$$N.C = x \% \quad \rightarrow \quad \alpha = 1 - \frac{x}{100}$$

3. Grados de libertad

$$v = (r - 1)(c - 1)$$

donde: $\begin{cases} r = \text{número de filas} \\ c = \text{número de columnas} \end{cases}$

4. Estadístico de prueba

$$\chi^2_{Tabla} \quad \text{con} \quad \alpha \wedge v$$

5. Establecimiento de los criterios de decisión

6. Cálculos

f_o = frecuencia absoluta de cada celda de la tabla

$f_e = \frac{r_i \cdot c_j}{n}$ donde r_i es el total de filas para la fila que contiene dicha celda n de la tabla y c_j es el total de columnas para la columna que contiene dicha celda de la tabla.

7. Decisión

$$\text{Si: } \begin{cases} X^2 \leq X_{\text{Tabla}}^2 & \text{Se acepta } H_o \\ X^2 > X_{\text{Tabla}}^2 & \text{Se rechaza } H_o \end{cases}$$

8. Coeficiente de contingencia

$$C = \frac{X^2}{X^2 + n} \quad \text{además} \quad C \in \left[0, \sqrt{\frac{k-1}{k}} \right]$$

Donde $k = \min(|i|, |j|)$ el mínimo de entre la cantidad de formas posibles de la característica en las variables estudiadas.



Ejemplo

El señor Althomaro, presidente de la Compañía Nacional General Aseguradora de Salud, se opone al seguro de salubridad nacional. Argumenta que sería muy costoso de implantar, en particular debido a que la existencia de este sistema, entre otras cosas, tendería a fomentar en la gente permanecer más tiempo en los hospitales. Althomaro tiene la creencia de que las hospitalizaciones dependen del tipo de seguro de salud que tengan las personas. Le pide a Raúl Mendoza, el especialista en estadística de la empresa, que verifique el asunto. Mendoza recogió datos de una muestra aleatoria de 660 hospitalizaciones y la información la resumió en la tabla 01. Mendoza desea probar las hipótesis:

H_0 = tiempo de estancia y tipo de seguro son independientes.

H_1 = el tiempo de estancia depende del tipo de seguro.

$\alpha = 0,01$ (nivel de significancia para la prueba de estas hipótesis)

Tabla 01
Datos de hospitalizaciones clasificados según el tipo de cobertura del seguro y el tiempo de estancia

% de costos cubiertos por el seguro	Días en el hospital			Total
	< 5	5 – 10	> 10	
< 25 %	40	75	65	180
25 – 50 %	30	45	75	150
> 50 %	40	100	190	330
Total	110	220	330	660

Solución

Paso 1 – formulación de la hipótesis

H_0 = tiempo de estancia y tipo de seguro son independientes.

H_1 = el tiempo de estancia depende del tipo de seguro.

Paso 2 – nivel de significancia

N.C = 99%, lo cual implica $\alpha = 0,01$

Paso 3 – grados de libertad

$$v = (3 - 1)(3 - 1) \rightarrow v = 4$$

Paso 4 – Estadístico de prueba

$$\chi^2_{Tabla} \text{ con } \alpha = 0,01 \wedge v = 4 \qquad \lambda^2_{Tabla} = 13,28$$

Paso 5 – Establecimiento de los criterios de decisión

Paso 6 – Cálculos

Calculando las frecuencias esperadas de cada celda de la tabla:

$$\begin{aligned}
 f_e |_{c_1=40} &= \frac{180(110)}{660} = 30 & f_e |_{c_2=75} &= \frac{180(220)}{660} = 60 & f_e |_{c_3=65} &= \frac{180(330)}{660} = 90 \\
 f_e |_{c_4=30} &= \frac{150(110)}{660} = 25 & f_e |_{c_5=45} &= \frac{150(220)}{660} = 50 & f_e |_{c_6=75} &= \frac{150(330)}{660} = 75 \\
 f_e |_{c_7=40} &= \frac{330(110)}{660} = 55 & f_e |_{c_8=100} &= \frac{330(220)}{660} = 110 & f_e |_{c_9=190} &= \frac{330(330)}{660} = 165
 \end{aligned}$$

Agregando estos valores en la tabla, tenemos:

% de costos cubiertos por el seguro	Días en el hospital			Total
	< 5	5 – 10	> 10	
< 25 %	40 30	75 60	65 90	180
25 – 50 %	30 25	45 50	75 75	150
> 50 %	40 55	100 110	190 165	330
Total	110	220	330	660

$$\chi^2 = \left[\frac{(40-30)^2}{30} + \frac{(75-60)^2}{60} + \frac{(65-90)^2}{90} + \frac{(30-25)^2}{25} + \frac{(45-50)^2}{50} + \frac{(75-75)^2}{75} + \frac{(40-55)^2}{55} + \frac{(100-110)^2}{110} + \frac{(190-165)^2}{165} \right] \Rightarrow \chi^2 = 24,32$$

Paso 7 – Decisión

Como $\chi^2 = 24,32 > 13,28$ rechazamos H_0 . En este sentido, Mendoza debe rechazar la hipótesis nula e informar al señor Althomaro que la evidencia refuerza su creencia de que la duración de las hospitalizaciones y la cobertura de los seguros son *dependientes* entre sí.

Paso 8 – Coeficiente de contingencia

$$C = \sqrt{\frac{24,32}{24,32 + 660}} \rightarrow C = 0,19 \quad , \text{ además el valor máximo que puede tomar } C \text{ es:}$$

$$C_{\text{máx}} = \sqrt{\frac{3-1}{3}} \rightarrow C_{\text{máx}} = 0,82 \quad , \text{ gráficamente podemos observar:}$$



Dado que el valor obtenido está muy lejos del valor deseado (0,6) se concluye que la dependencia entre las variables es baja (débil)



Ejercitándose

- a) Estamos interesados en estudiar la fiabilidad de cierto componente informático con relación al distribuidor que nos lo suministra. Para realizar esto, tomamos una muestra de 100 componentes de cada uno de los 3 distribuidores que nos sirven el producto comprobando el número de defectuosos en cada lote. La siguiente tabla muestra el número de defectuosos en para cada uno de los distribuidores.

	Componentes defectuosos	Componentes correctos	Total
Distribuidor 1	16	94	100
Distribuidor 2	24	76	100
Distribuidor 3	9	81	100
Total	49	251	300

- b) Estamos interesados en estudiar la relación entre cierta enfermedad y la adicción al tabaco. Para realizar esto seleccionamos una muestra de 150 individuos, 100 individuos no fumadores y 50 fumadores. La siguiente tabla muestra las frecuencias de enfermedad en cada grupo.

	Padecen la Enfermedad	No Padecen la enfermedad	Total
Fumadores	12	88	100
No Fumadores	25	25	50
Total	37	113	150

TRABAJO PRACTICO 8

- I. Lan Perú desea determinar si existe alguna relación entre el número de vuelos que las personas toman y su ingreso. ¿A qué conclusión llega al nivel del 1% con base en los datos para 100 viajeros en la tabla de contingencia?

Ingreso	Frecuencia de vuelos		
	Nunca	Rara vez	Con frecuencia
Menos de \$ 30 000	20	15	2
30 000 – 50 000	8	5	1
50 000 – 70 000	7	8	12
Más de \$ 70 000	2	5	15

- II. Un editor de periódicos, que trata de determinar con precisión las características de mercado de su periódico, se pregunta si la costumbre de la gente de la comunidad de leer diarios está relacionada con el nivel educativo de los lectores. Se aplica una encuesta a los adultos del área referente a su nivel educativo y la frecuencia con que leen el periódico. Los resultados se muestran a continuación:

Frecuencia con la que leen	NIVEL EDUCATIVO				Total
	Profesional o posgrado	Pasante de licenciatura	Preparatoria	No terminó la preparatoria	
Nunca	10	17	11	21	59
Algunas veces	12	23	8	5	48
Mañana o tarde	35	38	16	7	96
Ambas ediciones	28	19	6	13	66
Total	85	97	41	46	269

- III. Un educador tiene la opinión de que las calificaciones que obtienen los alumnos de preparatoria dependen de la cantidad de tiempo que ellos pasan escuchando música. Será cierta su opinión.

Horas consumidas escuchando música	PROMEDIO DE CALIFICACIONES					Total
	A	B	C	D	E	
< 5	13	10	11	16	5	55
5 – 10	20	27	27	19	2	95
11 – 20	9	27	71	16	32	155
> 20	8	11	41	24	11	95
Total	50	75	150	75	50	400

RESPUESTAS DE LOS TRABAJOS PRÁCTICOS

TRABAJO PRÁCTICO 1 – Nociones Básicas de Estadística

1.

- a) Población: los jóvenes de la ciudad de Chimbote, entre 18 y 22 años de edad.
Muestra: 1000 jóvenes (50 de cada barrio) de la ciudad de Chimbote, entre 18 y 22 años de edad. Recuerde que la muestra es una parte representativa de la población, por ello, debe mantenerse la misma característica. En este caso la característica es tener entre 18 y 22 años de edad.
- b) Población: los jóvenes de la ciudad de Chimbote. Muestra: 500 jóvenes de la ciudad de Chimbote, elegidos de manera aleatoria.

2.

- a) El ingreso mensual (sueldo) es por excelencia una variable cuantitativa (continua)
- b) Lo que aquí importa es el grado que cursan. Por ejemplo, en primer grado hay 25 alumnos, en segundo grado hay 37 alumnos, etc. Por lo tanto, la variable es cualitativa (nominal)
- c) El código de identificación es como el DNI (34003247, 76293045, etc.). En consecuencia, la variable es cuantitativa (discreta), dado que, el número de DNI no puede tomar decimales.
- d) Los números de las camisetas pueden ser 1, 2, 3, etc. Por lo tanto, la variable es cuantitativa (discreta),
- e) Dado que se refiere a la posición, por ejemplo, el 1 significa primer puesto, el 2 significa segundo puesto y 3 el tercer puesto, es decir, posee un orden invariable. por lo tanto, la variable es cualitativa (ordinal)

TRABAJO PRÁCTICO 2 – Organización de Datos

1.

Edades	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$	Y_i
[20 – 25 >	12	12	0,17	0,17	17	17	22,5
[25 – 30 >	15	27	0,21	0,38	21	38	27,5
[30 – 35 >	23	50	0,33	0,71	33	71	32,5
[35 – 40 >	11	61	0,16	0,87	16	87	37,5
[40 – 45 >	9	70	0,13	1	13	100	42,5
Total	70	-	1	-	100	-	-

A = 30 años o más, es decir, [30 – 35 > + [35 – 40 > + [40 – 45 > = 33+16+13 = 62%

B: menos de 40 años, es decir, [20 – 25 > + [25 – 30 > + [30 – 35 > + [35 – 40 >, lo cual es igual a 17 + 21 + 33 + 16 = 87%. Luego, A + B = 149%

2.

Puntaje	f_i	F_i	h_i	H_i	$h_i \%$	$H_i \%$	Y_i
[10 – 15 >	10	10	0,11	0,11	11	11	12,5
[15 – 20 >	15	25	0,17	0,28	17	28	17,5
[20 – 25 >	28	53	0,31	0,59	31	59	22,5
[25 – 30 >	20	73	0,22	0,81	22	81	27,5
[30 – 35 >	17	90	0,19	1	19	100	32,5
Total	90	-	1	-	100	-	-

A = [20 – 35 > = 31+22+19 = 72% y B = [10 – 15 > = 11%, por lo tanto, A – B = 61%

3.

Tabla 1

- a) 70,11
- b) 257, 14
- c) 2, 83
- d) 27

Tabla 2

- a) 90, 12
- b) 294, 12
- c) 3, 38
- d) 25

TRABAJO PRÁCTICO 3 – Presentación de Datos

1. La variable: número de trabajadores es discreta, por lo tanto, el gráfico que le corresponde es el gráfico de bastones.

x_i	f_i
1	2
2	4
3	12
4	8
5	4

Recuerde que el eje x debe ser mayor al eje y

2. La variable: gastos semanales es continua, por lo tanto, el gráfico que le corresponde es el histograma o en su defecto el polígono de frecuencias.

Teniendo en cuenta que

$$R = 1000 - 400 = 600 \text{ m}$$

$$= 1 + 3,3 \log 20 = 5$$

$$C = 600/5 = 120$$

LI – LS	f_i
[400 – 520 >	2
[520 – 640 >	3
[640 – 760 >	8
[760 – 880 >	4
[880 – 1000 >	3

3. La variable: estado civil es ordinal, por lo tanto, el gráfico que le corresponde es el gráfico circular o en su defecto el gráfico de barras.

Estado civil	frecuencia
Soltero	7
Casado	18
Divorciado	3
Viudo	2
Total	30

■ Soltero ■ Casado
 ■ Divorciado ■ Viudo

Cuando la variable es cualitativa, evite usar abreviaturas (iniciales). Se recomienda escribir el nombre completo de la categoría.

4. Dado que el problema trata, por un lado, de fechas las cuales tienen un orden invariable y, por otro lado, del género o sexo; podemos afirmar lo siguiente: la fecha corresponde a una variable cualitativa ordinal y el sexo a una variable cualitativa nominal. En consecuencia, el gráfico que le corresponde es el gráfico de barras compuestas, dado que están presentes dos variables.

5. Este problema trata sobre el sueldo que ganan las mujeres respecto a lo que ganan los hombres. Por ejemplo, en Japón, una mujer gana el 43% de lo que gana un hombre (ver gráfico). En este sentido si una mujer, en Suiza, tiene un sueldo de 1300, este valor equivale al 68% del sueldo de un hombre, es decir, un suizo ganaría 1911,76 (ítem 1). De manera análoga, en España una mujer ganaría el 67% de 1102, esto es 738,34 (ítem 2)

TRABAJO PRÁCTICO 4 – Medidas de Tendencia Central

1.

X_i	f_i	F_i	Media	Mediana	Moda
0	1	1	$x = \frac{0(1) + \dots}{20}$ $x = 2,2$	$F_i > \frac{20}{2} \rightarrow F_i > 10$ $F_i = 12 \rightarrow x_i = 2$ $Me = x_i \rightarrow Me = 2$	La mayor f_i en este caso es $f_3 = 7$, en consecuencia, $Md = 2$
1	4	5			
2	7	12			
3	6	18			
4	2	20			
Total	20	-			

2.

$L_i - L_s$	f_i	F_i	Y_i	Media	Mediana
[26 – 34 >	1	1	30	$x = \frac{30(1) + \dots}{51}$ $x = 63,10$	$F_i > \frac{51}{2} \rightarrow F_i > 26,5$ $r_i = 33 \rightarrow [58 - 66 >$ $Me = 58 + 8 \left(\frac{26,5 - 17}{16} \right)$ $Me = 62,75$
[34 – 42 >	2	3	38		
[42 – 50 >	4	7	46		
[50 – 58 >	10	17	54		
[58 – 66 >	16	33	62		
[66 – 74 >	8	41	70		
[74 – 82 >	3	44	78		
[82 – 90 >	7	51	86		
Total	51	-	-		
Moda					
La mayor f_i en este caso es $f_5 = 16$, en consecuencia $Md \in [58 - 66 >$ $Md = 58 + 8 \left(\frac{6}{6 + 8} \right) \rightarrow Md = 61,41$					

3. Sean las edades: a, b c y d. Entonces, $x = \frac{a + b + c + d}{4} = 24 \Rightarrow a + b + c + d = 96$ y

$$Me = \frac{b + c}{2} = 23 \Rightarrow b + c = 46 \rightarrow a + d = 50, \text{ en este caso a y d no podrían ser las}$$

modas ya que por dato la moda es 22, en consecuencia sumarían 44.

Por este motivo, a y c son las modas o en su defecto b y d. Asumamos que a y c son

las modas, es decir: $a = c = 22$. Reemplazando estos valores en las ecuaciones

anteriores ($b + c = 46 \wedge a + d = 50$), tendríamos: $b = 24$ y $d = 28$.

4. Sean los datos: a_1, a_2, a_3 $\rightarrow \frac{a_1 + a_2 + a_3 + \dots + a_n}{125} = 42$ Pero se introdujo 124 en

lugar de 12, 4 es decir, hay un exceso de: $E = 124 - 12,4 = 111,6$ pero, este exceso se ha promediado, de acuerdo al dato del problema. En consecuencia:

$$\frac{111,6}{125} = 0,8928 (\text{exceso promedial}) \rightarrow x = 42 - 0,8928 \rightarrow x = 41,11 (\text{corregido})$$

5. En este caso, nos solicitan el promedio ponderado, esto es:

$$\frac{70(2,3) + 40(6,4) + 10(15)}{70 + 40 + 10} \Rightarrow 4,7525$$

TRABAJO PRÁCTICO 5 – Medidas de Dispersión

1.

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
0	2	0 - 2,52	6,35	12,70
1	4	1 - 2,52	2,31	9,24
2	21	2 - 2,52	0,27	5,67
3	15	3 - 2,52	0,23	3,45
4	6	4 - 2,52	2,19	13,14
5	1	5 - 2,52	6,15	6,15
6	1	6 - 2,52	12,11	12,11

$$\sum (x_i - \bar{x})^2 f_i = 62,46 \wedge n = 50$$

$$\frac{62,46}{50 - 1} \Rightarrow s = 1,13$$

Li - Ls	f_i	Y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
[61 - 65 >	5	63	63 - 76,44	180,63	903,15
[65 - 69 >	5	67	67 - 76,44	89,11	445,55
[69 - 73 >	6	71	71 - 76,44	29,59	177,54
[73 - 77 >	7	75	75 - 76,44	2,07	14,49
[77 - 81 >	9	79	79 - 76,44	6,55	58,95
[81 - 85 >	14	83	83 - 76,44	43,03	602,42
[85 - 89]	4	87	87 - 76,44	111,51	446,04

$$\sum (x_i - \bar{x})^2 f_i = 2648,14$$

$$\frac{2648,14}{50 - 1} \Rightarrow s = 7,35$$

2. Por comodidad escribiremos la siguiente tabla de manera vertical

Intervalo	5 - 9	9 - 13	13 - 17	17 - 21	21 - 25
Frecuencia	10	8	5	4	3

Li - Ls	f_i	Y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
5-9	10	7	7-12,6	31,36	313,6
9-13	8	11	11-12,6	2,56	20,48
13-17	5	15	15-12,6	5,76	28,80
17-21	4	19	19-12,6	40,96	163,84
21-25	3	23	23-12,6	108,16	324,48

$$\sum (x_i - \bar{x})^2 f_i = 851,2 \wedge n = 30$$

$$\frac{851,2}{30-1} \Rightarrow s = 5,42$$

3.

1 ^{er} producto	2 ^{do} producto	3 ^{er} producto
$\bar{x} = 2,5 \rightarrow s = 0,004$	$\bar{x} = 2,8 \rightarrow s = 0,006$	$\bar{x} = 3,7 \rightarrow s = 0,09$

Si los promedios fueran iguales en los tres productos, se compararían de manera directa sus desviaciones estándar. Pero en este caso, las medias son distintas. En tal sentido, es necesario hallar el coeficiente de variación, el cual relaciona las medias con sus desviaciones estándar. El coeficiente de variación se define como:

$$CV = \frac{s}{\bar{x}}$$

1 ^{er} producto	2 ^{do} producto	3 ^{er} producto
0,0016 <> 0,16%	0,002 <> 0,2%	0,024 <> 2,4%

1^{er} producto posee el menor coeficiente de variación. En consecuencia, es el más estable. Por otro lado, el 3^{er} producto es el que está más alejado del desempeño promedio.

TRABAJO PRÁCTICO 6 – Probabilidades

1. Veamos cada caso:

- a) Hay 100 tornillos, de los cuales 3 son defectuosos. Por ende, hay 97 tornillos sin defecto. Es decir, la probabilidad de que sea defectuoso es 3/100 y la probabilidad de que no tenga defecto es 97/100. Pero, al elegir el 2^{do} tornillo quedarían 99 tornillos, de los cuales solo dos son defectuosos, dado que ya se escogió uno de ellos. En consecuencia, la probabilidad de que el segundo tornillo resulte defectuoso es 2/99

- b) Los casos posibles del evento son: { HHH, HHM, HMH, HMM, MHH, MHM, MMH, MMM }. Entonces, la probabilidad de que sus tres hijos sean varones es { HHH } = $1/8$
- c) En total hay 10 bolas (entre rojas y verdes). En este sentido, la probabilidad de que resulte roja es $6/10$, tanto para 1^{ra} como para la 2^{da} extracción. Esto ocurre porque la bola se vuelve a introducir. Entonces, la probabilidad de que las dos bolas sean rojas es $(6/10)(6/10) = 36/100$
- d) En este caso es sin reposición (no se vuelve a introducir). Entonces, la probabilidad de que en la 1^{ra} sea roja es $6/10$ pero la probabilidad de que la 2^{da} sea roja es $5/9$ (en total quedan 9 bolas, de las cuales 5 son rojas, dado que ya elegimos una roja previamente). Es decir, la probabilidad de que ambas sean rojas es $(6/10)(5/9) = 30/90$
2. La probabilidad de que un prisionero sea indultado es $1/3$, inicialmente. Como son dos los que adquirirán el indulto, se puede dar los siguientes casos: { AB, AC, BC }, es decir, si el prisionero A no preguntara la probabilidad de que él sea uno de los que adquieran el indulto sería { AB o AC } = $2/3$. Ahora bien, si preguntara (naturalmente no le dirán que es uno de los elegidos) el miembro del tribunal le diría es B, pero este puede estar acompañado por A o por C. En su defecto, si el miembro del tribunal diría que es C, este puede estar acompañado por A o por B, en consecuencia habrían 4 casos { BA, BC, CA, CB } sin embargo, $BC = CB$. Entonces, solo tenemos tres casos posibles { BA, BC, CA}. Nótese que A participa en dos de ellas, en consecuencia, la probabilidad de que él sea uno de los indultados, en caso de preguntar, es $2/3$. De ello, se concluye que da lo mismo que le haga o no la pregunta al miembro del tribunal.

TRABAJO PRÁCTICO 7 – Probabilidad Binomial

1.

Alguna vez (n = 10, p = 9/15 <> 0,6; q = 6/15 <> 0,4; k = 0)

$$P(X=0) = \binom{10}{0} (0,6)^0 (0,4)^{10} = 0,00010 \Rightarrow P(x) = 1 - P(X=0) = 0,99999$$

Más de 8 veces (n = 10, p = 0,6, q = 0,4, k = 9 y 10)

$$P(X=9) + P(X=10) = \binom{10}{9} (0,6)^9 (0,4)^1 + \binom{10}{10} (0,6)^{10} (0,4)^0 = 0,04031 + 0,00605 = 0,04636$$

2.

Alguno defectuoso (n = 60, p = 0,03; q = 0,97; k = 0)

$$P(X=0) = \binom{60}{0} (0,03)^0 (0,97)^{60} = 0,16080 \Rightarrow P(x) = 1 - P(X=0) = 0,83920$$

Menos de dos defectuosos (n = 60, p = 0,03, q = 0,97, k = 1 y 0)

$$P(X=1) + P(X=0) = \binom{60}{1} (0,03)^1 (0,97)^{59} + \binom{60}{0} (0,03)^0 (0,97)^{60} = 0,2984 + 0,1608 = 0,4592$$

3.

Alguna vez (n = 15, p = 0,4; q = 0,6; k = 0)

$$P(X=0) = \binom{15}{0} (0,4)^0 (0,6)^{15} = 0,00047 \Rightarrow P(x) = 1 - P(X=0) = 0,99953$$

Menos de dos veces (n = 15, p = 0,4, q = 0,6, k = 1 y 0)

$$P(X=1) + P(X=0) = \binom{15}{1} (0,4)^1 (0,6)^{14} + \binom{15}{0} (0,4)^0 (0,6)^{15} = 0,00047 + 0,00047 = 0,00094$$

4. De manera análoga: $P(X=3) = \binom{10}{3} (0,4)^3 (0,6)^7 = 0,2150$

TRABAJO PRÁCTICO 8 – Prueba Chi Cuadrado

1.

Ingreso	Frecuencia de vuelos		
	Nunca	Rara vez	Con frecuencia
Menos de \$ 30 000	20 13,69	15 12,21	2 11,10
30 000 – 50 000	8 5,18	5 4,62	1 4,20
50 000 – 70 000	7 9,99	8 8,91	12 8,10
Más de \$ 70 000	2 8,14	5 7,26	15 6,60

$$\chi^2 = \left[\begin{aligned} & \left(\frac{20 - 13,69}{13,69} \right)^2 + \left(\frac{15 - 12,21}{12,21} \right)^2 + \left(\frac{2 - 11,10}{11,10} \right)^2 + \left(\frac{8 - 5,18}{5,18} \right)^2 \\ & \left(\frac{5 - 4,62}{4,62} \right)^2 + \left(\frac{1 - 4,20}{4,20} \right)^2 + \left(\frac{7 - 9,99}{9,99} \right)^2 + \left(\frac{8 - 8,91}{8,91} \right)^2 \\ & \left(\frac{12 - 8,10}{8,10} \right)^2 + \left(\frac{2 - 8,14}{8,14} \right)^2 + \left(\frac{5 - 7,26}{7,26} \right)^2 + \left(\frac{15 - 6,60}{6,60} \right)^2 \end{aligned} \right] \Rightarrow \chi^2 = 33,90$$

Se sabe que: $\chi^2 \Big|_{\alpha=0,01}^{v=6} = 16,81$ pero $33,90 > 16,81$ entonces se rechaza H_0

2.

Frecuencia con la que leen	NIVEL EDUCATIVO				Total
	Profesional o posgrado	Pasante de licenciatura	Preparatoria	No terminó la preparatoria	
Nunca	10 18,64	17 21,28	11 8,99	21 10,09	59
Algunas veces	12 15,17	23 17,31	8 7,32	5 8,21	48
Mañana o tarde	35 30,33	38 34,62	16 14,63	7 16,42	96
Ambas ediciones	28 20,86	19 23,80	6 10,06	13 11,29	66
Total	85	97	41	46	269

$$X^2 = \left[\begin{aligned} & \left(\frac{10-18,64}{18,64} \right)^2 + \left(\frac{17-21,28}{21,28} \right)^2 + \left(\frac{11-8,99}{8,99} \right)^2 + \left(\frac{21-10,09}{10,09} \right)^2 \\ & \left(\frac{12-15,17}{15,17} \right)^2 + \left(\frac{23-17,31}{17,31} \right)^2 + \left(\frac{8-7,32}{7,32} \right)^2 + \left(\frac{5-8,21}{8,21} \right)^2 \\ & \left(\frac{35-30,33}{30,33} \right)^2 + \left(\frac{38-34,62}{34,62} \right)^2 + \left(\frac{16-14,63}{14,63} \right)^2 + \left(\frac{7-16,42}{16,42} \right)^2 \\ & \left(\frac{28-20,86}{20,86} \right)^2 + \left(\frac{19-23,80}{23,80} \right)^2 + \left(\frac{6-10,06}{10,06} \right)^2 + \left(\frac{13-11,29}{11,29} \right)^2 \end{aligned} \right] \Rightarrow X^2 = 32,86$$

Se sabe que: $X^2 \Big|_{\alpha=0,05} = 16,92$ pero $32,86 > 16,92$ entonces se rechaza H_0

3.

Horas consumidas escuchando música	PROMEDIO DE CALIFICACIONES					Total
	A	B	C	D	E	
< 5	13 6,88	10 10,31	11 20,63	16 10,31	5 6,88	55
5 – 10	20 11,88	27 17,81	27 35,63	19 17,81	2 11,88	95
11 – 20	9 19,38	27 29,06	71 58,13	16 29,06	32 19,38	155
> 20	8 6,88	11 10,31	41 20,63	24 10,31	11 6,88	95
Total	50	75	150	75	50	400

De manera análoga (ver problema 1 y 2), tenemos:

$$X^2 = \left[\frac{13-6,88}{6,88} \right]^2 + \left[\frac{11-6,88}{6,88} \right]^2 \Rightarrow X^2 = 97,91$$

Se sabe que: $X^2 \Big|_{\alpha=0,01} = 26,22$ pero $97,91 > 26,22$ entonces se rechaza H_0

Fin

PREGUNTAS SOBRE CONCEPTO Y TEORIA

1. En comparación con un arreglo de datos, la distribución de frecuencias tiene la ventaja de representar los datos de una manera comprimida. V F
2. Un histograma es una serie de rectángulos, cada uno proporcional en ancho al número de elementos que caen dentro de una clase específica de datos. V F
3. Cuando una muestra contiene las características importantes de cierta población en las mismas proporciones como se encuentran en ésta, se dice que se trata de una muestra representativa. V F
4. Si uniéramos los puntos medios de las barras consecutivas de un histograma de frecuencias con una serie de rectas, estaríamos graficando un polígono de frecuencias. V F
5. Una desventaja del ordenamiento de datos es que no nos permite hallar fácilmente los valores mayores y menores del conjunto de datos. V F
6. El valor de cada observación del conjunto de datos se toma en cuenta cuando calculamos su mediana. V F
7. Cuando la población está sesgada positiva o negativamente, a menudo es preferible utilizar la mediana como mejor medida de posición, debido a que siempre está entre la media y la moda. V F
8. Las medidas de tendencia central de un conjunto de datos se refieren al grado en que las observaciones están dispersas. V F
9. El valor que más se repite en un conjunto de datos se conoce como media aritmética. V F

10. Si organizamos las observaciones de un conjunto de datos en orden V F descendente, el punto de datos que se encuentra en medio es la mediana del conjunto de datos.
11. ¿Cuál de los siguientes no es un ejemplo de datos comprimidos?
- a) Distribución de frecuencias
 - b) Arreglo de datos
 - c) Histograma
 - d) Ojiva
12. ¿Cuál de las afirmaciones acerca de los rectángulos de un histograma es correcta?
- a) Los rectángulos tienen una altura proporcional al número de elementos que entra en cada una de las clases.
 - b) Por lo general existen 5 rectángulos en cada histograma
 - c) El área de un rectángulo depende solo del número de elementos de la clase
13. ¿Cuál de los siguientes no es una prueba acerca de la utilidad de los datos?
- a) La fuente de datos
 - b) La contradicción con respecto a otra evidencia
 - c) La falta de evidencia
 - d) El número de observaciones
 - e) N.A.
14. Las gráficas de distribuciones de frecuencias se utiliza debido a que:
- a) Tiene una larga historia en aplicaciones prácticas
 - b) Atraen la atención sobre los patrones que siguen los datos
 - c) Toman en cuenta los datos parciales o incompletos
 - d) Permiten una fácil estimación de los datos
 - e) Incisos b y d
15. Los datos continuos se diferencian de los datos discretos en que:
- a) Las clases de datos discretos están representadas por fracciones

- b) Las clases de datos continuos pueden representarse por fracciones
- c) Los datos continuos solo toman valores enteros
- d) Los datos discretos pueden tomar cualquier valor real

16. ¿Cuál de las afirmaciones siguientes no es correcta?

- a) Algunos conjuntos de datos no poseen media
- b) El cálculo de una media se ve afectado por los valores extremos del conjunto de datos
- c) Una media pesada se debe utilizar cuando es necesario tomar en consideración la importancia de cada valor
- d) T.A.

17. Cuando una distribución es simétrica y posee solamente una moda, el punto más alto de la curva de distribución se conoce como:

- a) La moda
- b) La media
- c) La mediana
- d) T.A.

18. Cuando nos referimos a que una curva está cargada hacia el extremo izquierdo, podemos decir que es:

- a) Simétrica
- b) Sesgada hacia la derecha
- c) Positivamente sesgada
- d) N. A.

19. Si un evento no se ve afectado por el resultado de otro evento, se dice que ambos eventos son:

- a) Dependientes
- b) Independientes
- c) Mutuamente excluyentes
- d) Tanto b como c

20. Suponga que se lanza un dado dos veces consecutivas y que usted tiene que trazar el árbol de probabilidades que muestra todos los resultados posibles de los dos lanzamientos ¿Cuántas ramas tendrá el árbol?

- a) 6
- b) 12
- c) 36
- d) 42
- e) 48

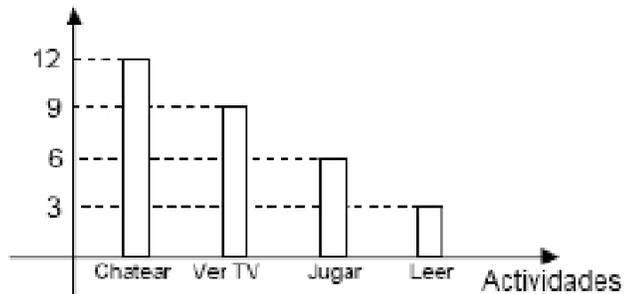
MISCELÁNEA – PROBLEMAS SELECTOS

1. La tabla adjunta muestra las edades de 220 alumnos de un colegio. ¿Cuál(es) de las siguientes afirmaciones es (son) verdadera(s)?
 - i. La moda es 17 años
 - ii. La mediana es mayor que la media
 - iii. La mitad de los alumnos tiene 17 o 18 años

2. Las fichas del peso de 10 niños, marcan en promedio 20 kg. En la oficina de control se pierde una ficha y se sabe que el promedio del resto es 19 kg. ¿Cuál es el peso del niño al que le perdieron la ficha?

3. Si se tabularan las frecuencias de las estaturas y color de ojos de los alumnos de un curso, ¿Cuál de las opciones siguientes es **siempre** verdadera?
 - i. Con la moda de las estaturas se determina la estatura promedio del curso
 - ii. Con la mediana del color de ojos se determina el color de ojos que predomina
 - iii. Con el promedio de las estaturas se determina la estatura más frecuente
 - iv. Con la mediana de las estaturas se determina la estatura más frecuente
 - v. Con la moda del color de ojos se determina el color de ojos que predomina

4. Se pregunta a los alumnos acerca de lo que más les gusta hacer en vacaciones y sus respuestas están en el gráfico de la figura. ¿Cuál(es) de las siguientes afirmaciones es (son) verdadera(s)?
 - i. Al 30% de los alumnos, lo que más les gusta es chatear
 - ii. A la mitad de los alumnos, lo que más les gusta es ver TV o jugar
 - iii. Al 30% de los alumnos, lo que más les gusta es leer o jugar



5. La tabla adjunta muestra la distribución de los puntajes obtenidos por los alumnos de un curso en una prueba de matemática. ¿Cuál(es) de las siguientes afirmaciones es (son) verdadera(s)?

Intervalos de puntaje	Frecuencia
10 - 19	6
20 - 29	8
30 - 39	12
40 - 49	5
50 - 59	9

- i. El total de alumnos que rindió la prueba es 40
- ii. La mediana se encuentra en el intervalo: 20 – 29
- iii. El intervalo modal es el intervalo: 30 – 39

CRÉDITOS

- *El presente documento denominado ESTADÍSTICA DESDE CERO, es un material híbrido en cuanto a su contenido. Algunos conceptos y problemas se extrajeron de los materiales elaborados por DANNY PERICH CAMPANA – PSU Matemática y RICHARD LEVIN & DAVID RUBIN – Estadística para Administradores.*