ESTADÍSTICA DESCRIPTIVA

DISTRIBUCIÓN DE FRECUENCIAS, HISTOGRAMA, POLIGONO Y ESTADÍSITICOS DE TENDENCIA CENTRAL, DISPERSIÓN, ASIMETRÍA Y CURTOSIS.

Prof.: MSc. Julio R. Vargas

I. Las calificaciones finales de un curso de estadística descriptiva se registran en la tabla siguiente.

68	65	89	57	90	95	79	53
73	78	67	81	93	69	65	85
61	78	73	68	62	60	76	93
66	62	73	60	77	76	75	75
96	80	82	74	95	62	76	72
79	67	73	94	85	76	85	60
65	75	87	75	78	88	63	71
86	88	75	78	63	59	68	75
84	75	61	88	62	78	83	74
79	82	97	72	71	74	71	77

Con relación a esta tabla, encontrar:

a. La puntuación más alta
b. La puntuación más baja
c. El rango
d. Las puntuaciones de los cinco estudiantes de mayor calificación.
e. Las puntuaciones de los cinco estudiantes de menor calificación.
f. La puntuación del décimo estudiante de mayor calificación.
g. ¿Cuántos estudiantes obtuvieron puntuación de 75 o mayor?
h. ¿Cuántos estudiantes obtuvieron puntuación menor que 85?
i. ¿Qué porcentaje de estudiantes obtuvo una puntuación mayor o igual a 65 pero menor que 85.
j. ¿Qué puntuaciones no tiene ningún estudiante?

II. Con los datos construya una distribución de frecuencia de amplitud o tamaño 5 y que la primera clase sea 50 – 54.
III. Construya el histograma y polígono de frecuencia.
IV. Calcular la media aritmética, mediana y moda de la distribución.
V. Calcular el coeficiente de variación, varianza y desviación estándar.
VI. Calcular el valor de asimetría y el curtosis.

Solución:

Los datos de la tabla los ordenaremos en forma ascendente:

Puede ser útil a través del gráfico llamado tallo y hoja. Como los números son de dos dígitos, el tallo serán las decena y las hojas las unidades: las decena son; 5,6,7,8,9.

5 | 5, 9, 3

6 | 8, 1, 6, 5, 5, 2, 7, 7, 1, 8, 0, 2, 3, 2,9, 0, 2, 5, 3, 8, 0

7 | 3, 9, 9, 8, 8, 5, 5, 3, 3, 3, 5, 4, 5, 8, 2, 7, 8, 1, 6, 6, 8, 4, 9, 6, 5, 6, 1, 5, 2, 1, 5, 4, 7

8 | 6, 4, 0, 8, 2, 9, 2, 7, 1, 8, 5, 8, 5, 3, 5

9 | 6, 7, 4, 0, 3, 5, 5, 3

Para construir el gráfico recorrimos el gráfico por columna de izquierda a derecha, aun se puede hacer por fila.

Ahora ubicamos los datos ordenados en la tabla siguiente.

53	57	59	60	60	60	61	61
62	62	62	62	63	63	65	65
65	66	67	67	68	68	68	69
71	71	71	72	72	73	73	73
73	74	74	74	75	75	75	75
75	75	75	76	76	76	76	77
77	78	78	78	78	78	79	79
79	80	81	82	82	83	84	85
85	85	86	87	88	88	88	89
90	93	93	94	95	95	96	97

Con los datos ordenados podemos responder las preguntas del epígrafe (I).

a. La puntuación más alta: **97**
b. La puntuación más baja: **53**
c. El rango: 97-53=**44**
d. Las puntuaciones de los cinco estudiantes de mayor calificación: **97,96,95,95,94**
e. Las puntuaciones de los cinco estudiantes de menor calificación: **60, 60, 59, 57, 53**
f. La puntuación del décimo estudiante de mayor calificación: **88**
g. ¿Cuántos estudiantes obtuvieron puntuación de 75 o mayor? **44**
h. ¿Cuántos estudiantes obtuvieron puntuación menor que 85? **63**
i. ¿Qué porcentaje de estudiantes obtuvo una puntuación mayor o igual a 65 pero menor que 85? 49/80 = **61.25%**
j. ¿Qué puntuaciones no tiene ningún estudiante? **De 0 a 52, 54, 55, 56, 58, 64, 70, 91, 92, 98, 99.**

II. Con los datos construya una distribución de frecuencia de amplitud o tamaño 5 y que la primera clase sea 50 – 54.

Tabla 1: Distribución de Frecuencias

Intervalos	X	Frecuencia Absoluta (f)	Frecuencia abs. acumulada	%frecuencia	% frecuencia acumulado
50-54	52	1	1	1.3	1.3
55-59	57	2	3	2.5	3.8
60-64	62	11	14	13.8	17.5
65-69	67	10	24	12.5	30.0
70-74	72	12	36	15.0	45.0
75-79	77	21	57	26.3	71.3
80-84	82	6	63	7.5	78.8
85-89	87	9	72	11.3	90.0
90-94	92	4	76	5.0	95.0
95-99	97	4	80	5.0	100.0
Total		80		100.0	

X: es la marca de clase o punto medio de cada intervalo, se obtiene dividiendo la suma de los límites inferior y superior de cada intervalo por 2.

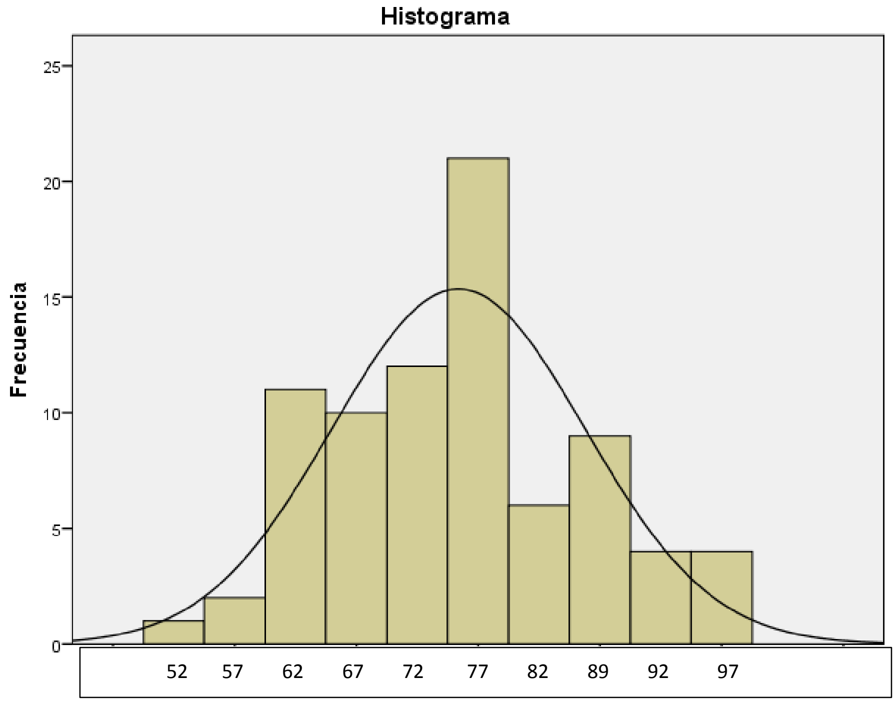
Frecuencia absoluta: se obtiene contando la cantidad de datos que se encuentran en cada intervalo, recuerde que un dato solo puede pertenecer a un solo intervalo. La suma de todas las frecuencias absolutas debe ser n (en nuestro caso 80)

Frecuencia acumulada: Se obtiene sumando las frecuencias absolutas de los intervalos hasta la frecuencia del intervalo actual. El último intervalo debe tener como frecuencia absoluta el total de datos que se están estudiando.

%frecuencia: Es la frecuencia relativa a cada intervalo y se obtiene dividiendo la frecuencia absoluta de cada intervalo por el número total de datos (n=80) multiplicado por 100. La suma de todas las frecuencias relativas debe es 100%

%frecuencia acumulada: Es la frecuencia relativa acumulada. Se obtiene sumando las frecuencias relativas de los intervalos hasta la frecuencia del intervalo actual. El último intervalo debe tener como frecuencia relativa acumulada del total de datos que se están estudiando.

III. Construya el histograma y polígono de frecuencia.



IV. Calcular la media aritmética, mediana y moda de la distribución.

X	Frecuencia Abs. (f)	Xf
52	1	52
57	2	114
62	11	682
67	10	670
72	12	864
77	21	1617
82	6	492
87	9	783
92	4	368
97	4	388
	∑	6030

Tabla 2

Media aritmética:

$$\bar{X} = \frac{\sum Xf}{n} = \frac{6030}{80} = 75.375$$

Mediana:

$M_e = L_i + \left[\frac{n/_2 - (\sum f)_1}{f_{mediana}}\right]$*C

L_i: límite inferior de la clase donde está la mediana.

$(\sum f)_1$: Sumatoria de las frecuencias absolutas anteriores a la clase mediana.

$f_{mediana}$: frecuencia absoluta de la clase donde se encuentra la mediana

n: número total de datos.

C: amplitud de la clase o intervalo.

La mediana divide la distribución exactamente a la mitad, por lo que dividimos el total de datos n=80 por 2. n/2 = 80/2= 40. Buscamos en frecuencia absoluta acumulada (tabla 1) quien contiene a 40 y vemos que es 57, es decir que la median está en la clase 75 – 79.

Por lo que los valores de la ecuación serán:

L_i=75 n/2= 40 C=5 $f_{mediana}$=21 $(\sum f)_1$= 36

$$M_e = L_i + \left[\frac{n/_2 - (\sum f)_1}{f_{mediana}}\right] * C = 75 + \left[\frac{\frac{80}{2} - 36}{21}\right] * 5 = 75 + 0.95 = 75.952$$

Moda:

$$M_o = L_i + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] * C$$

L_i: límite inferior de la clase donde está la mayor frecuencia absoluta.

Δ_1: diferencia entre la frecuencia absoluta modal y la frecuencia absoluta de la clase anterior.

Δ_2: diferencia entre la frecuencia absoluta modal y la frecuencia absoluta de la clase superior

C: amplitud de la clase o intervalo.

Si observamos la tabla 1, en la columna de frecuencia absoluta encontramos que 21 es el valor más alto, por tanto la clase modal será 75 – 79. Ahora calcularemos Δ_1 y Δ_2

Δ_1= 21 – 12 = 9 Δ_2=21 – 6= 19 C=5

$$M_o = L_i + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] * C = 75 + \left[\frac{9}{9 + 19}\right] * 5 = 75 + 1.61 = 76.607$$

Conclusión en relación a los valores de las medidas de tendencia central: Las diferencia entre cada uno de ellos es mínima, el orden es: $\bar{X} < M_e < M_o$. En este caso cualquiera de esas medias es válido para representar el dato central de la distribución.

V. Calcular el coeficiente de variación, varianza y desviación estándar.

Varianza de la distribución: Es la media de los cuadrados de las desviaciones o separaciones de cada una de las observaciones, respecto a la media aritmética.. Se presenta por s^2.

$$s^2 = \frac{\sum fx^2}{n} - \bar{x}^2$$

Desviación Estándar: Es la raíz cuadrada de la varianza. Con ello corregimos el haber tomado cuadrados de separaciones en el cálculo de la varianza.

$$s = \sqrt{\frac{\sum fx^2}{n} - \bar{x}^2}$$

Tabla 3

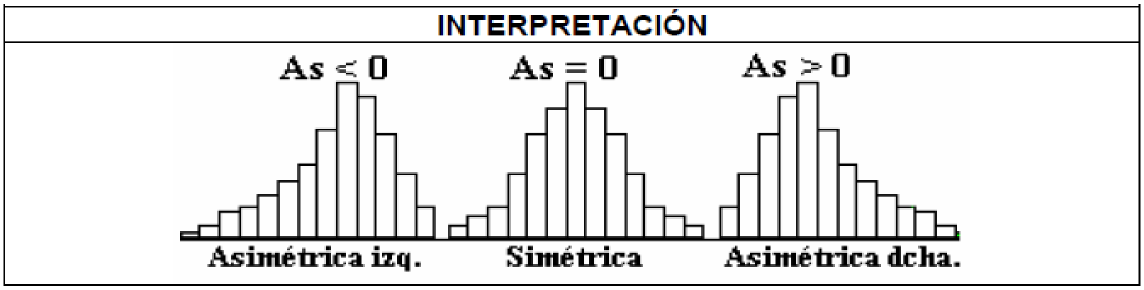
X	Frecuencia Abs. (f)	fX^2	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
52	1	2704	-12771.88	298542.72
57	2	6498	-12408.29	228002.38
62	11	42284	-26319.28	352020.41
67	10	44890	-5874.28	49197.07
72	12	62208	-461.32	1556.96
77	21	124509	90.11	146.43
82	6	40344	1744.65	11558.32
87	9	68121	14139.09	164366.90
92	4	33856	18379.98	305567.11
97	4	37636	40450.91	874751.02
	∑	463050	16969.69	2285709.32

$$s^2 = \frac{\sum fx^2}{n} - \bar{x}^2 = \frac{463050}{80} - (75.375)^2 = 5788.125 - 5681.391 = 106.734$$

$s = \sqrt{\frac{\sum fx^2}{n} - \bar{x}^2} = \sqrt{106.734} = 10.33$ (La desviación estándar nos indica que en hay una distancia promedio de aproximada 10 puntos entre la media y las demás puntuaciones.

VI. Calcular el valor de asimetría y el curtosis.

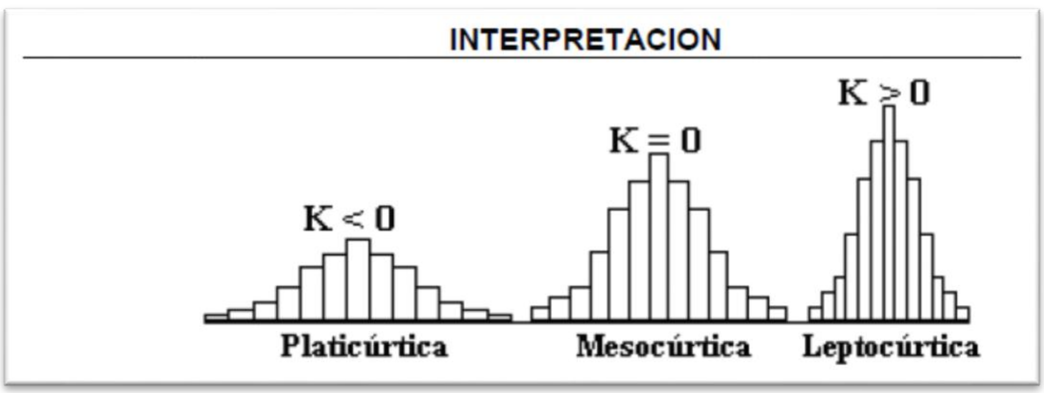
Coeficiente de Asimetría de FISHER: Permite interpretar la forma de la distribución, respecto a ser o no simétrica.



$$As = \frac{\frac{\sum f(x-\bar{x})^3}{n}}{s^3} = \frac{16969.69/80}{10.33^3} = \frac{212.12}{1102.30} = 0.19$$ **Es ligeramente asimétrica por la derecha.**

Cálculos de $\sum f(x - \bar{x})^3$ en la tabla 3.

Coeficiente de Curtosis: Recibe también el nombre de coeficiente de ***concentración central***, midiendo el grado de aplastamiento o apuntamiento de la gráfica de la distribución de la variable estadística. Una mayor concentración de datos en torno al promedio harán que la forma sea alargada, siendo tanto más plana (o aplastada) cuanto mayor sea la dispersión de los mismos.



$$K = \frac{\frac{\sum f(x-\bar{x})^4}{n}}{s^4} - 3 = \frac{2285709.32/80}{(10.33)^4} - 3 = \frac{28571.37}{11386.79} - 3 = 2.50 - 3 = -0.50$$

Es Platicúrtica, un poco aplanada.

Coeficiente de Variación: Si la media es representativa de las observaciones (no existen valores extremos exageradamente distanciados de la mayoría), el coeficiente de variación permite **comparar la dispersión de dos series estadísticas:** mayor coeficiente indica menor homogeneidad, o lo que es lo mismo, mayor dispersión o variabilidad.

$$CV = \frac{s}{\bar{X}} * 100$$

Ejemplo: Importancia del uso del CV: Establezca, con base estadística, en cuál de las siguientes empresas el salario está repartido de forma más equitativa.

Empresa A		Empresa B	
nº personas	salario percibido (€)	nº personas	salario percibido (€)
15	800	10	800
20	1000	30	1000
30	1200	35	1200
20	1500	24	1500
15	7500	1	7500

Empresa A			
X	f	fx	fx^2
800	15	12000	9600000
1000	20	20000	20000000
1200	30	36000	43200000
1500	20	30000	45000000
7500	15	112500	843750000
Suma	100	210500	961550000

Empresa B			
X	f	fx	fx^2
800	10	8000	6400000
1000	30	30000	30000000
1200	35	42000	50400000
1500	24	36000	54000000
7500	1	7500	56250000
Suma	100	123500	197050000

$$\bar{X}_A = \frac{\sum Xf}{n} = \frac{210500}{100} = 2150$$

$$\bar{X}_B = \frac{\sum Xf}{n} = \frac{123500}{100} = 1235$$

$$s_A = \sqrt{\frac{\sum fx^2}{n} - \bar{x}^2} = \sqrt{\frac{961550000}{100} - (2150)^2} = 2234.5$$

$$s_B = \sqrt{\frac{\sum fx^2}{n} - \bar{x}^2} = \sqrt{\frac{197050000}{100} - (1235)^2} = 667.3$$

$$CV_A = \frac{s}{\bar{X}} * 100 = \frac{2234.5}{2150} * 100 = 103.9\%$$

$$CV_B = \frac{s}{\bar{X}} * 100 = \frac{667.3}{1235} * 100 = 54.\%$$

Interpretación: Puede apreciar que el coeficiente de variación de la empresa ha es casi el doble del CV de la empresa B. Esto indica que en la empresa B hay mayor equidad en la distribución de los salarios en comparación con la empresa A.