CBS

Colegio Bautista Shalom

Estadística I
Cuarto PCOC PFS
Cuarto Bimestre

Contenidos

MEDIDAS DE DISPERSIÓN

- ✓ COEFICIENTE DE VARIACIÓN.
- ✓ MÉTODOS DE CÁLCULO.
- ✓ PUNTUACIONES.
  - PUNTUACIONES DIFERENCIALES.
  - PUNTUACIONES TÍPICAS.
- ✓ GRADO DE CONCENTRACIÓN - INDICE DE GINI.
- ✓ COEFICIENTE DE ASIMETRÍA.
- ✓ COEFICIENTE DE FISHER DE SIMETRÍA.
- ✓ COEFICIENTE DE CURTOSIS.
- ✓ DISTRIBUCIONES BIDIMENSIONALES.
- ✓ DISTRIBUCIONES MARGINALES.
- ✓ COEFICIENTE DE CORRELACIÓN LINEAL.
- ✓ REGRESIÓN LINEAL.

INFORMACIÓN (INCLUÍDA EN ESTE DOCUMENTO EDUCATIVO) TOMADA DE:

Sitios web:

1. *http://www.aulafacil.com/cursos/t675/ciencia/estadisticas/estadisticas*
2. *https://es.wikipedia.org/wiki/Medidas_de_dispersi%C3%B3n*
3. *https://economipedia.com/definiciones/coeficiente-de-variacion.html*
4. *https://www.stadcenterecuador.com/estadisticas/ejercicios/14-basicos/22-ejercicios-resueltos-coeficiente-de-variacion-coeficiente-de-asimetria-pearson-y-deciles*
5. *https://www.monografias.com/trabajos88/dispersion-relativa/dispersion-relativa.shtml*
6. *https://www.superprof.es/apuntes/escolar/matematicas/estadistica/descriptiva/coeficiente-de-variacion-y-puntuaciones-tipicas.html*

NOTA: conforme avances en el aprendizaje tu catedrático(a) te indicará la actividad o ejercicio a realizar. Sigue sus instrucciones.

MEDIDAS DE DISPERSIÓN

Estas son *parámetros estadísticos* Parámetros estadísticos que indican como se alejan los datos respecto de la media aritmética. Sirven como indicador de la variabilidad de los datos.

Las medidas de dispersión más utilizadas son el rango, la desviación estándar y la varianza. Estudiadas anteriormente. Otras medidas de dispersión son adimensionales. En otras palabras, no tienen unidades, incluso si la variable en sí tiene unidades.

COEFICIENTE DE VARIACIÓN

El coeficiente de variación, también denominado coeficiente de variación de Pearson es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos.

Es decir, nos informa al igual que otras medidas de dispersión, de si una variable se mueve mucho, poco, más o menos que otra.

Fórmula del coeficiente de variación:

Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la media del conjunto y por lo general se expresa en porcentaje para su mejor comprensión.

$$CV = \frac{\sigma_x}{|\bar{X}|}$$

X: variable sobre la que se pretenden calcular la varianza
σ_x: Desviación típica de la variable X.
$|\bar{x}|$: Es la media de la variable X en valor absoluto con $\bar{x} \neq 0$

El coeficiente de variación se puede ver expresado con las letras CV o r, dependiendo del manual o la fuente utilizada. Su fórmula es la siguiente:

El coeficiente de variación se utiliza para comparar conjuntos de datos pertenecientes a poblaciones distintas. Si atendemos a su fórmula, vemos que este tiene en cuenta el valor de la media. Por lo tanto, el coeficiente de variación nos permite tener una medida de dispersión que elimine las posibles distorsiones de las medias de dos o más poblaciones.

Ejemplos de uso del coeficiente de variación en lugar de la desviación típica.

A continuación, mostramos algunos ejemplos sobre esta medida de dispersión:

1. Comparación de conjuntos de datos de diferente dimensión:

   Se quiere comprar la dispersión entre la altura de 50 alumnos de una clase y su peso. Para comparar la altura podríamos utilizar como unidad de medida metros y centímetros y para el peso el kilogramo. Comparar estas dos distribuciones mediante la desviación estándar, no tendría sentido dado que se pretenden medir dos variables cualitativas distintas (una medida de longitud y una de masa).

2. Comparar conjuntos con gran diferencia entre medias:

   Imaginemos por ejemplo que queremos medir el peso de los escarabajos y el de los hipopótamos. El peso de los escarabajos se mide en gramos o miligramos y el peso de los hipopótamos por lo general se mide en toneladas. Si para nuestra medición convertimos el peso de los escarabajos a toneladas para que ambas poblaciones estén en la misma escala, utilizar la desviación estándar como medida de dispersión no sería lo adecuado. El peso medio de los escarabajos medido en toneladas sería tan pequeño que, si utilizamos la desviación estándar, apenas habría dispersión en los datos. Esto sería un error dado que el peso entre las diferentes especies de escarabajos puede variar de manera considerable.

Ejemplo de cálculo del coeficiente de variación:

Pensemos en una población de elefantes y otra de ratones. La población de elefantes tiene un peso medio de 5.000 kilogramos y una desviación típica de 400 kilogramos. La población de ratones tiene un peso medio de 15 gramos y una desviación típica de 5 gramos. Si comparáramos la dispersión de ambas poblaciones mediante la desviación típica podríamos pensar que hay mayor dispersión para la población de elefantes que para la de los ratones.

Sin embargo, al calcular el coeficiente de variación para ambas poblaciones, nos daríamos cuenta que es justo al contrario.

Elefantes: 400/5000=0,08

Hormigas: 5/15=0,33

Si multiplicamos ambos datos por 100, tenemos que el coeficiente de variación para los elefantes es de apenas un 8%, mientras que el de los ratones es de un 33%. Como consecuencia de la diferencia entre las poblaciones y su peso medio, vemos que la población con mayor dispersión no es la que tiene una mayor desviación típica.

Ejemplo:

Una población de alumnos tiene una estatura media de 160 cm con una desviación estándar de 16 cm. Estos mismos alumnos, tienen un peso medio de 70 kg con una desviación estándar de 14 kg. ¿Cuál de las 2 variables presenta mayor variabilidad relativa?

Solución:

Vamos a comparar la dispersión de 2 variables, la estatura y el peso, usando el coeficiente de variación.

Estatura (E)	Peso (P)
$\mu_E = 160\ cm \quad \wedge \quad \sigma_E = 16\ cm$	$\bar{x}_P = 70\ kg \quad \wedge \quad s_P = 14\ kg$
$CV_E = \frac{\sigma_E}{\mu_E} = \frac{16\ cm}{160\ cm} = \frac{1}{10} = 0{,}1 = 10\%$	$CV_P = \frac{s_P}{\bar{x}_P} = \frac{14\ kg}{70\ kg} = \frac{1}{5} = 0{,}2 = 20\%$

Podemos que ver que $CV_P > CV_E$, por eso, el peso de esta población de alumnos tiene mayor variabilidad relativa que la estatura.

El Coeficiente de variación (*CV*) es una medida de la dispersión relativa de un conjunto de datos, que se obtiene dividiendo la desviación estándar del conjunto entre su media aritmética y se expresa generalmente en términos porcentuales.

Propiedades:

- ✓ Puesto que tanto la desviación estándar como la media se miden en las unidades originales, el CV es una medida independiente de las unidades de medición.
- ✓ Debido a la propiedad anterior el CV es la cantidad más adecuada para comparar la variabilidad de dos conjuntos de datos.

MÉTODOS DE CÁLCULO

Para una población se emplea la siguiente fórmula:

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Para una muestra se emplea la siguiente fórmula:

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Donde:
CV = Coeficiente de variación.
s = desviación estándar de la muestra.
$\bar{x}$ = media aritmética de la muestra.

Ejemplo ilustrativo:

Mathías, un estudiante universitario, tiene las siguientes calificaciones en las 10 asignaturas que recibe en su carrera: 8, 7, 10, 9, 8, 7, 8, 10, 9 y 10. Josué, un compañero de Mathías, tiene las siguientes calificaciones: 8, 9, 8, 7, 8, 9, 10, 7, 8 y 10.

¿Cuál estudiante tiene menor variabilidad en sus calificaciones?

Solución:

Como se está tomando en cuenta todas las asignaturas, se debe calcular el coeficiente de variación poblacional.

Sin agrupar los datos empleando Excel se calcula el coeficiente de variación tal como se muestra en la siguiente figura:

	A	B	C	D	E	F	G	H
1	Mathias				Josué			
2	8				7			
3	7				7			
4	10				8			
5	9				8			
6	8				8			
7	7				8			
8	8				9			
9	10				9			
10	9				10			
11	10				10			
12	μ	8,6	=PROMEDIO(A2:A11)		μ	8,4	=PROMEDIO(E2:E11)	
13	σ	1,1135529	=DESVEST.P(A2:A11)		σ	1,0198039	=DESVEST.P(E2:E11)	
14	CV	12,948289	=(B13/B12)*100		CV	12,140523	=(F13/F12)*100	

Agrupando los datos en tablas de frecuencias se calcula así:

a) Se agrupa las calificaciones y se realiza el cálculo la media aritmética:

Para Mathías se obtiene:

Calificaciones (x_i)	f	fx_i
7	2	14
8	3	24
9	2	18
10	3	30
Total	10	86

$$\mu = \frac{\sum x_i}{N} = \frac{86}{10} = 8{,}6$$

Para Josué se obtiene:

Calificaciones (x_i)	f	fx_i
7	2	14
8	4	32
9	2	18
10	2	20
Total	10	84

$$\mu = \frac{\sum x_i}{N} = \frac{84}{10} = 8{,}4$$

b) Se calcula la desviación estándar:

Para Mathías se obtiene:

Calificaciones (x_i)	f	fx_i	$(x_i - \mu)^2$	$f(x_i - \mu)^2$
7	2	14	2,56	5,12
8	3	24	0,36	1,08
9	2	18	0,16	0,32
10	3	30	1,96	5,88
Total	10	86		12,4

$$\sigma = \sqrt{\frac{\sum f(x_i - \mu)^2}{N}} = \sqrt{\frac{12{,}4}{10}} = 1{,}1136$$

Para Josué se obtiene:

Calificaciones (x_i)	f	fx_i	$(x_i - \mu)^2$	$f(x_i - \mu)^2$
7	2	14	1,96	3,92
8	4	32	0,16	0,64
9	2	18	0,36	0,72
10	2	20	2,56	5,12
Total	10	84		10,4

$$\sigma = \sqrt{\frac{\sum f(x_i - \mu)^2}{N}} = \sqrt{\frac{10{,}4}{10}} = 1{,}0198$$

c) Se calcula el coeficiente de variación:

Para Mathías se obtiene:

$$CV = \frac{\sigma}{\mu} = \frac{1{,}1136}{8{,}6} = 0{,}129 = 12{,}9\%$$

Empleando Excel es como muestra la siguiente figura:

	A	B	C	D	E	F
1	x_i	f	$f(x_i-\mu)^2$			
2	7	2	5,12	=B2*(A2-B7)^2		
3	8	3	1,08	=B3*(A3-B7)^2		
4	9	2	0,32	=B4*(A4-B7)^2		
5	10	3	5,88	=B5*(A5-B7)^2		
6	Total	10	12,4	=SUMA(C2:C5)		
7	μ	8,6	=SUMAPRODUCTO(A2:A5;B2:B5)/SUMA(B2:B5)			
8						
9	σ	1,1135529	=RCUAD(C6/SUMA(B2:B5))			
10						
11	CV	12,948289	=(B9/B7)*100			

Para Josué se obtiene:

$$CV = \frac{\sigma}{\mu} = \frac{1{,}0198}{8{,}4} = 0{,}121 = 12{,}1\%$$

Empleando Excel es como muestra la siguiente figura:

	A	B	C	D	E	F
1	x_i	f	$f(x_i-\mu)^2$			
2	7	2	3,92	=B2*(A2-B7)^2		
3	8	4	0,64	=B3*(A3-B7)^2		
4	9	2	0,72	=B4*(A4-B7)^2		
5	10	2	5,12	=B5*(A5-B7)^2		
6	Total	10	10,4	=SUMA(C2:C5)		
7	μ	8,4	=SUMAPRODUCTO(A2:A5;B2:B5)/SUMA(B2:B5)			
8						
9	σ	1,0198039	=RCUAD(C6/SUMA(B2:B5))			
10						
11	CV	12,140523	=(B9/B7)*100			

Interpretación: *Por lo tanto, el estudiante que tiene menor variabilidad en sus calificaciones es Josué.*

Ejemplo ilustrativo:

Se saca una muestra de un curso de la Universidad UTN sobre las calificaciones en las asignaturas de Matemática y Estadística, resultados que se presentan en las siguientes tablas. ¿En qué asignatura existe mayor variabilidad? Realice los cálculos empleando Excel:

Matemática	
Intervalos	f
2 - 4	8
5 - 7	12
8 - 10	20
Total	40

Estadística	
Intervalos	f
2 - 4	8
5 - 7	14
8 - 10	18
Total	40

Solución:

Los cálculos para la asignatura de Matemática empleando Excel se muestran en la siguiente figura:

	A	B	C	D	E	F	G	H	I
1	Intervalos		f	xm			$f(xm_i - \bar{x})^2$		
2	2	4	10	3	=PROMEDIO(A2:B2)		152,1	=C2*(D2-B7)^2	
3	5	7	8	6	=PROMEDIO(A3:B3)		6,48	=C3*(D3-B7)^2	
4	8	10	22	9	=PROMEDIO(A4:B4)		97,02	=C4*(D4-B7)^2	
5	Total		40	=SUMA(C2:C4)			255,6	=SUMA(G2:G4)	
6									
7	$\bar{x}$	6,9	=SUMAPRODUCTO(D2:D4;C2:C4)/C5						
8									
9	s	2,56	=RCUAD(G5/(C5-1))						
10									
11	CV	37,1	=(B9/B7)*100						

Los cálculos para la asignatura de Estadística empleando Excel se muestran en la siguiente figura:

	A	B	C	D	E	F	G	H	I
1	Intervalos		f	xm			$f(xm_i - \bar{x})^2$		
2	2	4	8	3	=PROMEDIO(A2:B2)		112,5	=C2*(D2-B7)^2	
3	5	7	14	6	=PROMEDIO(A3:B3)		7,875	=C3*(D3-B7)^2	
4	8	10	18	9	=PROMEDIO(A4:B4)		91,125	=C4*(D4-B7)^2	
5	Total		40	=SUMA(C2:C4)			211,5	=SUMA(G2:G4)	
6									
7	$\bar{x}$	6,75	=SUMAPRODUCTO(D2:D4;C2:C4)/C5						
8									
9	s	2,33	=RCUAD(G5/(C5-1))						
10									
11	CV	34,5	=(B9/B7)*100						

TAREA PROPUESTA POR TU CATEDRÁTICO(A).

PUNTUACIONES

PUNTUACIONES DIFERENCIALES

Las **puntuaciones diferenciales** resultan de restarles a las puntuaciones directas la media aritmética.

$$x_i = X_i - \bar{X}$$

PUNTUACIONES TÍPICAS

Las **puntuaciones típicas** son el resultado de dividir las puntuaciones diferenciales entre la desviación típica. Este proceso se llama tipificación.

Las **puntuaciones típicas** se representan por z.

$$z = \frac{X_i - \bar{X}}{\sigma}$$

Observaciones sobre puntuaciones típicas:

- ✓ La media aritmética de las puntuaciones típicas es: 0.
- ✓ La desviación típica de las puntuaciones típicas es: 1.
- ✓ Las puntuaciones típicas son adimensionales, es decir, son independientes de las unidades utilizadas
- ✓ Las puntuaciones típicas se utilizan para comparar las puntuaciones obtenidas en distintas distribuciones

Ejemplo:

En una clase hay 15 alumnos y 20 alumnas. El peso medio de los alumnos es 58,2 kg y el de las alumnas 52,4 kg. Las desviaciones típicas de los dos grupos son, respectivamente, 3,1 kg y 5,1 kg. El peso de José es de 70 kg y el de Ana es 65 kg. ¿Cuál de ellos puede, dentro del grupo de alumnos de su sexo, considerarse más grueso?

$$z_1 = \frac{70 - 58,2}{3,1} = 3,81 \qquad z_2 = \frac{65 - 52,4}{5,1} = 2,47$$

José es más grueso respecto de su grupo que Ana respecto al suyo.

TAREA PROPUESTA POR TU CATEDRÁTICO(A).

GRADO DE CONCENTRACIÓN - ÍNDICE DE GINI

Las **medidas de forma** permiten conocer que forma tiene la curva que representa la serie de datos de la muestra. En concreto, podemos estudiar las siguientes características de la curva:

a) Concentración: mide si los valores de la variable están más o menos uniformemente repartidos a lo largo de la muestra.
b) Asimetría: mide si la curva tiene una forma simétrica, es decir, si respecto al centro de la misma (centro de simetría) los segmentos de curva que quedan a derecha e izquierda son similares.
c) Curtosis: mide si los valores de la distribución están más o menos concentrados alrededor de los valores medios de la muestra.

Concentración:

Para medir el nivel de concentración de una distribución de frecuencia se pueden utilizar distintos indicadores, entre ellos el **índice de Gini**.

Este índice se calcula aplicando la siguiente fórmula:

$$IG = \frac{\sum (pi \ - \ qi)}{\sum pi} \qquad (i \ toma \ valores \ entre \ 1 \ y \ n-1)$$

En donde pi mide el porcentaje de individuos de la muestra que presentan un valor igual o inferior al dexi.

$$pi = \frac{\mathrm{n1} \ + \ \mathrm{n2} \ + \ \mathrm{n3} + \ldots + \mathrm{ni}}{n} \times 100$$

Mientras que qi se calcula aplicando la siguiente fórmula:

$$qi = \frac{(X1 * n1) \ + \ (X2 * n2) + \ldots + (Xi * ni)}{(X1 * n1) \ + \ (X2 * n2) + \ldots + (Xn * nn)} \times 100$$

El **índice Gini** (IG) puede tomar valores entre 0 y 1:

IG = 0: concentración mínima. La muestra está uniformemente repartida a lo largo de todo su rango.

IG = 1: concentración máxima. Un sólo valor de la muestra acumula el 100% de los resultados.

Ejemplo: vamos a calcular el índice Gini de una serie de datos con los sueldos de los empleados de una empresa (millones pesetas).

SUELDOS	EMPLEADOS (FRECUENCIAS ABSOLUTAS)		FRECUENCIAS RELATIVAS	
(Millones)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
x	x	x	x	x
3,5	10	10	25,0%	25,0%
4,5	12	22	30,0%	55,0%
6,0	8	30	20,0%	75,0%
8,0	5	35	12,5%	87,5%
10,0	3	38	7,5%	95,0%
15,0	1	39	2,5%	97,5%
20,0	1	40	2,5%	100,0%

Calculamos los valores que necesitamos para aplicar la fórmula del índice de Gini:

Xi	ni	Σ ni	pi	Xi * ni	Σ Xi * ni	qi	pi - qi
x	x	x	x	x	x	x	x
3,5	10	10	25,0	35,0	35,0	13,6	10,83
4,5	12	22	55,0	54,0	89,0	34,6	18,97
6,0	8	30	75,0	48,0	147,0	57,2	19,53
8,0	5	35	87,5	40,0	187,0	72,8	15,84
10,0	3	38	95,0	30,0	217,0	84,4	11,19
15,0	1	39	97,5	15,0	232,0	90,3	7,62
25,0	1	40	100,0	25,0	257,0	100,0	0
x	x	x	x	x	x	x	x
Σ pi (entre 1 y n-1) =			435,0	x	Σ (pi - qi) (entre 1 y n-1) =		83,99

Por lo tanto:

$$G = \frac{83{,}99}{435{,}0} = 0{,}19$$

Un índice Gini de 0,19 indica que la muestra está bastante uniformemente repartida, es decir, su nivel de concentración no es excesivamente alto.

Ejemplo: Ahora vamos a analizar nuevamente la muestra anterior, pero considerando que hay más personal de la empresa que cobra el sueldo máximo, lo que conlleva mayor concentración de renta en unas pocas personas.

Sueldos	Empleados (Frecuencias absolutas)		Frecuencias relativas	
(Millones)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
x	x	x	x	x
3,5	10	10	25,0%	25,0%
4,5	10	20	25,0%	50,0%
6,0	8	28	20,0%	70,0%
8,0	5	33	12,5%	82,5%
10,0	3	36	7,5%	90,0%
15,0	0	36	0,0%	90,0%
20,0	4	40	10,0%	100,0%

En este caso obtendríamos los siguientes datos:

Xi	ni	Σ ni	pi	Xi * ni	Σ Xi * ni	qi	pi - qi
x	x	x	x	x	x	x	x
3,5	10	10	25,0	35	35	11,7	13,26
4,5	10	20	50,0	45	80	26,8	23,15
6,0	8	28	70,0	48	128	43,0	27,05
8,0	5	33	82,5	40	168	56,4	26,12
10,0	3	36	90,0	30	198	66,4	23,56
15,0	0	36	90,0	0	198	66,4	23,56
25,0	4	40	100,0	100	298	100,0	0,00
x	x	x	x	x	x	x	x
Σ pi (entre 1 y n-1) =			407,5	x	Σ (pi - qi) (entre 1 y n-1) =		136,69

El **índice Gini** sería:

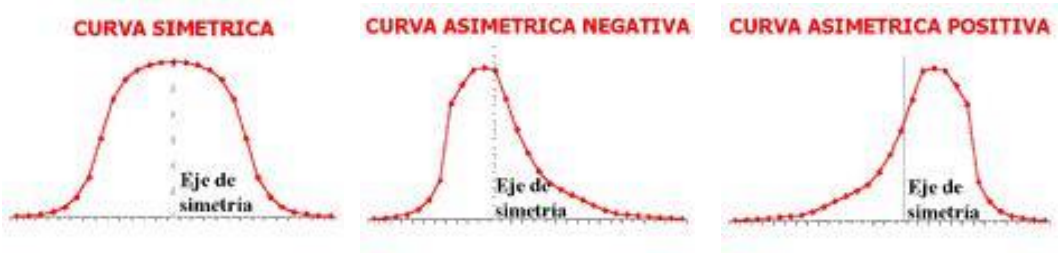
$$IG = \frac{136{,}69}{407{,}5} = 0{,}34$$

El Índice Gini se ha elevado considerablemente, reflejando la mayor concentración de rentas que hemos comentado

COEFICIENTE DE ASIMETRÍA

Asimetría

Hemos comentado que el concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritemética)



Para medir el nivel de asimetría se utiliza el llamado **Coeficiente de Asimetría de Fisher**, que viene definido:

$$g_1 = \frac{(1/n) * \Sigma (x_i - \bar{x}m)^3 * n_i}{((1/n) * \Sigma (x_i - \bar{x}m)^2 * n_i))^{3/2}}$$

Los resultados pueden ser los siguientes:

g1 = 0 *(distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media).*
g1 > 0 *(distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda).*
g1 < 0 *(distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha).*

Ejemplo: Vamos a calcular el Coeficiente de Asimetría de Fisher de la serie de datos referidos a la estatura de un grupo de alumnos (lección 2ª):

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
x	x	x	x	x
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

Recordemos que la media de esta muestra es 1,253

$\sum(xi - x)^3 * ni$	$\sum(xi - x)^2 * ni$
x	**x**
0,000110	0,030467

Luego:

$$g1 = \frac{\left(\frac{1}{30}\right) * (0{,}000110)}{\left(\frac{1}{30}\right) * (0{,}030467)^{3/2}} = -0{,}1586$$

Por lo tanto, el **COEFICIENTE DE FISHER DE SIMETRÍA** de esta muestra es -0,1586, lo que quiere decir que presenta una distribución asimétrica negativa (se concentran más valores a la izquierda de la media que a su derecha).

COEFICIENTE DE CURTOSIS

Curtosis

El **Coeficiente de Curtosis** analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución.

Se definen 3 tipos de distribuciones según su grado de curtosis:

1. **Distribución mesocúrtica:** presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).
2. **Distribución leptocúrtica**: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.
3. **Distribución platicúrtica:** presenta un reducido grado de concentración alrededor de los valores centrales de la variable.

El **Coeficiente de Curtosis** viene definido por la siguiente fórmula:

$$g_2 = \frac{(1/n) * \Sigma (x_i - xm)^4 * n_i}{((1/n) * \Sigma (x_i - xm)^2 * n_i))^2} - 3$$

Los resultados pueden ser los siguientes:

g2 = 0 (distribución mesocúrtica).
g2 > 0 (distribución leptocúrtica).
g2 < 0 (distribución platicúrtica).

Ejemplo: Vamos a calcular el Coeficiente de Curtosis de la serie de datos referidos a la estatura de un grupo de alumnos (lección 2ª):

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
x	x	x	x	x
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

Recordemos que la media de esta muestra es 1,253

$\sum (xi - xm)^4 * ni$	$\sum (xi - xm)^2 * ni$
x	**x**
0,00004967	0,03046667

Luego:

$$g2 = \frac{\left(\frac{1}{30}\right) * (0{,}00004967)}{\left(\frac{1}{30}\right) * (0{,}03046667)^2} - 3 = -1{,}39$$

Por lo tanto, el **Coeficiente de Curtosis** de esta muestra es -1,39, lo que quiere decir que se trata de una distribución platicúrtica, es decir, con una reducida concentración alrededor de los valores centrales de la distribución.

DISTRIBUCIONES BIDIMENSIONALES

Las distribuciones bidimensionales son aquellas en las que se estudian al mismo tiempo dos variables de cada elemento de la población: por ejemplo: peso y altura de un grupo de estudiantes; superficie y precio de las viviendas de una ciudad; potencia y velocidad de una gama de coches deportivos.

Para representar los datos obtenidos se utiliza una **tabla de correlación:**

X / Y	y1	y2	.....	ym-1	ym
x1	n1,1	n1,2	x	n1,m-1	n1,m
x2	n2,1	n2,2	x	n2,m-1	n2,m
.....	x	x	x	x	x
xn-1	nn-1,1	nn-1,2	x	nn-1,m-1	nn-1,m
xn	nn,1	nn,2	x	nn,m-1	nn,m

Las "x" representan una de las variables y las "y" la otra variable. En cada intersección de un valor de "x" y un valor de "y" se recoge el número de veces que dicho par de valores se ha presentado conjuntamente.

Ejemplo: Medimos el peso y la estatura de los alumnos de una clase y obtenemos los siguientes resultados:

Alumno	Estatura	Peso	Alumno	Estatura	Peso	Alumno	Estatura	Peso
x	**x**	**x**	**x**	**X**	**x**	**x**	**x**	**x**
Alumno 1	1,25	32	Alumno 11	1,25	31	Alumno 21	1,25	33
Alumno 2	1,28	33	Alumno 12	1,28	35	Alumno 22	1,28	32
Alumno 3	1,27	31	Alumno 13	1,27	34	Alumno 23	1,27	34
Alumno 4	1,21	34	Alumno 14	1,21	33	Alumno 24	1,21	34
Alumno 5	1,22	32	Alumno 15	1,22	33	Alumno 25	1,22	35
Alumno 6	1,29	31	Alumno 16	1,29	31	Alumno 26	1,29	31
Alumno 7	1,30	34	Alumno 17	1,30	35	Alumno 27	1,30	34
Alumno 8	1,24	32	Alumno 18	1,24	32	Alumno 28	1,24	33
Alumno 9	1,27	32	Alumno 19	1,27	31	Alumno 29	1,27	35
Alumno 10	1,29	35	Alumno 20	1,29	33	Alumno 30	1,29	34

Esta información se puede representar de un modo más organizado en la siguiente tabla de correlación:

Estatura / Peso	31 kg	32 kg	33 kg	34 kg	35 kg
1,21 cm	0	0	1	2	0
1,22 cm	0	1	1	0	1
1,23 cm	0	0	0	0	0
1,24 cm	0	2	1	0	0
1,25 cm	1	1	1	0	0
1,26 cm	0	0	0	0	0
1,27 cm	2	1	0	2	1
1,28 cm	0	1	1	0	1
1,29 cm	3	0	1	1	1
1,30 cm	0	0	0	2	1

Tal como se puede ver, en cada casilla se recoge el número de veces que se presenta conjuntamente cada par de valores (x, y).

Tal como vimos en las distribuciones unidimensionales si una de las variables (o las dos) presentan gran número de valores diferentes, y cada uno de ellos se repite en muy pocas ocasiones, puede convenir agrupar los valores de dicha variable (o de las dos) en tramos.

TAREA PROPUESTA POR TU CATEDRÁTICO(A).

DISTRIBUCIONES MARGINALES

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de cómo se comporta la otra. Estaríamos así en el análisis de una **distribución marginal**.

De cada distribución bidimensional se pueden deducir dos distribuciones marginales: una correspondiente a la variable x, y otra correspondiente a la variable y.

Distribución marginal de X:

X	ni.
x	x
x1	n1.
x2	n2.
.....	...
xn-1	nn-1.
xn	nn.

Distribución marginal de Y:

Y	n.j
x	x
y1	n.1
y2	n.2
.....	...
ym-1	n.m-1
ym	n.m

Ejemplo: a partir del ejemplo que vimos en la lección anterior (serie con los pesos y medidas de los alumnos de una clase) vamos a estudiar sus distribuciones marginales.

Estatura / Peso	31 kg	32 kg	33 kg	34 kg	35 kg
1,21 cm	0	0	1	2	0
1,22 cm	0	1	1	0	1
1,23 cm	0	0	0	0	0
1,24 cm	0	2	1	0	0
1,25 cm	1	1	1	0	0
1,26 cm	0	0	0	0	0
1,27 cm	2	1	0	2	1
1,28 cm	0	1	1	0	1
1,29 cm	3	0	1	1	1
1,30 cm	0	0	0	2	1

Las variables marginales se comportan como variables unidimensionales, por lo que pueden ser representadas en tablas de frecuencias.

a) Distribución marginal de la variable X (estatura):

Obtenemos la siguiente tabla de frecuencia:

Variable	Frecuencias absolutas		Frecuencias relativas	
(Estatura)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
xx	xx	xx	xx	xx
1,21	3	3	10,0%	10,0%
1,22	3	6	10,0%	20,0%
1,23	0	6	0,0%	20,0%
1,24	3	9	10,0%	30,0%
1,25	3	12	10,0%	40,0%
1,26	0	12	0,0%	40,0%
1,27	6	18	20,0%	60,0%
1,28	3	21	10,0%	70,0%
1,29	6	27	20,0%	90,0%
1,30	3	30	10,0%	100,0%

b) Distribución marginal de la variable Y (peso):

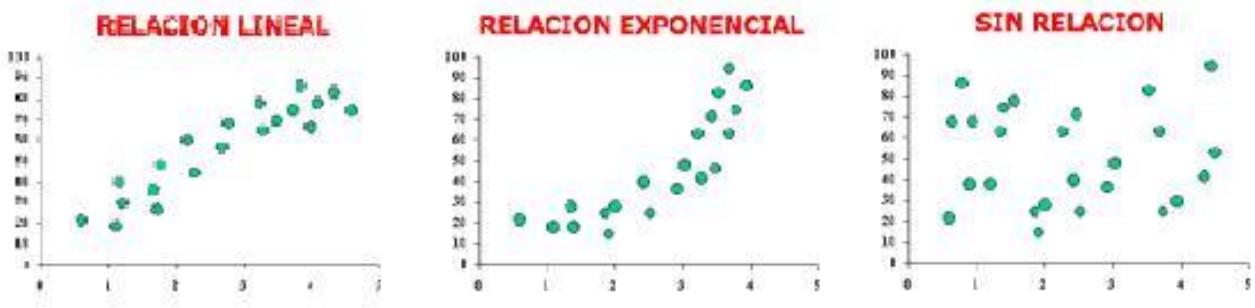
Obtenemos la siguiente tabla de frecuencia:

Variable	Frecuencias absolutas		Frecuencias relativas	
(Peso)	**Simple**	**Acumulada**	**Simple**	**Acumulada**
xx	xx	xx	xx	xx
31	6	6	20,0%	20,0%
32	6	12	20,0%	40,0%
33	6	18	20,0%	60,0%
34	7	25	23,3%	83,3%
35	5	30	16,6%	100,0%

COEFICIENTE DE CORRELACIÓN LINEAL

En una distribución bidimensional puede ocurrir que las dos variables guarden algún tipo de relación entre si.
Por ejemplo, si se analiza la estatura y el peso de los alumnos de una clase es muy posible que exista relación entre ambas variables: mientras más alto sea el alumno, mayor será su peso.

El coeficiente de correlación lineal mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).



No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado. Para ver, por tanto, si se puede utilizar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y ver qué forma describe.

El **coeficiente de correlación lineal** se calcula aplicando la siguiente fórmula:

$$r = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{\left((1/n * \Sigma (x_i - xm)^2) * (1/n * \Sigma (y_i - ym)^2) \right)^{1/2}}$$

Es decir:

Numerador: se denomina **covarianza** y se calcula de la siguiente manera: en cada par de valores (x,y) se multiplica la "x" menos su media, por la "y" menos su media. Se suma el resultado obtenido de todos los pares de valores y este resultado se divide por el tamaño de la muestra.

Denominador se calcula el producto de las varianzas de "x" y de "y", y a este producto se le calcula la raíz cuadrada.
Los valores que puede tomar el **coeficiente de correlación "r"** son: -1 < r < 1

Si "r" > 0, la correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1.

Por ejemplo: altura y peso: los alumnos más altos suelen pesar más.

Si "r" < 0, la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.

Por ejemplo: peso y velocidad: los alumnos más gordos suelen correr menos.

Si "r" = 0, no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

De todos modos, aunque el valor de "r" fuera próximo a 1 o -1, tampoco esto quiere decir obligatoriamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

Ejemplo: vamos a calcular el coeficiente de correlación de la siguiente serie de datos de altura y peso de los alumnos de una clase:

Alumno	Estatura	Peso	Alumno	Estatura	Peso	Alumno	Estatura	Peso
x	x	x	x	x	x	x	x	x
Alumno 1	1,25	32	Alumno 11	1,25	33	Alumno 21	1,25	33
Alumno 2	1,28	33	Alumno 12	1,28	35	Alumno 22	1,28	34
Alumno 3	1,27	34	Alumno 13	1,27	34	Alumno 23	1,27	34
Alumno 4	1,21	30	Alumno 14	1,21	30	Alumno 24	1,21	31
Alumno 5	1,22	32	Alumno 15	1,22	33	Alumno 25	1,22	32
Alumno 6	1,29	35	Alumno 16	1,29	34	Alumno 26	1,29	34
Alumno 7	1,30	34	Alumno 17	1,30	35	Alumno 27	1,30	34
Alumno 8	1,24	32	Alumno 18	1,24	32	Alumno 28	1,24	31
Alumno 9	1,27	32	Alumno 19	1,27	33	Alumno 29	1,27	35
Alumno 10	1,29	35	Alumno 20	1,29	33	Alumno 30	1,29	34

Aplicamos la fórmula:

$$r = \frac{\left(\frac{1}{30}\right) * 0{,}826}{\left(\left(\frac{1}{30}\right) * 0{,}02568\right) * \left(\left(\frac{1}{30}\right) * 51{,}366\right)^{1/2}}$$

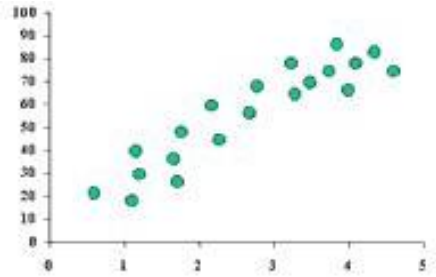
Luego,

r = 0,719

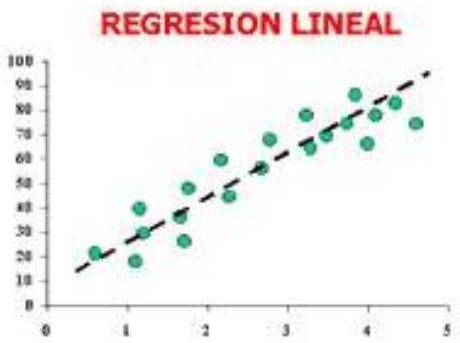
Por lo tanto, la correlación existente entre estas dos variables es elevada (0,7) y de signo positivo.

REGRESIÓN LINEAL

Representamos en un gráfico los pares de valores de una distribución bidimensional: la variable "x" en el eje horizontal o eje de abscisa, y la variable "y" en el eje vertical, o eje de ordenada. Vemos que la nube de puntos sigue una tendencia lineal:



El **coeficiente de correlación lineal** nos permite determinar si, efectivamente, existe relación entre las dos variables. Una vez que se concluye que sí existe relación, la **regresión** nos permite definir la recta que mejor se ajusta a esta nube de puntos.



Una recta viene definida por la siguiente fórmula:

$$y = a + bx$$

Donde "y" sería la variable dependiente, es decir, aquella que viene definida a partir de la otra variable "x" (variable independiente). Para definir la recta hay que determinar los valores de los parámetros "a" y "b":

El **parámetro "a"** es el valor que toma la variable dependiente "y", cuando la variable independiente "x" vale 0, y es el punto donde la recta cruza el eje vertical.

El **parámetro "b"** determina la pendiente de la recta, su grado de inclinación.

La **regresión lineal** nos permite calcular el valor de estos dos parámetros, definiendo la recta que mejor se ajusta a esta nube de puntos.

El **parámetro "b"** viene determinado por la siguiente fórmula:

$$b = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{1/n * \Sigma (x_i - xm)^2}$$

Es la covarianza de las dos variables, dividida por la varianza de la variable "x".

El **parámetro "a"** viene determinado por:

$$a = ym - (b * xm)$$

Es la media de la variable "y", menos la media de la variable "x" multiplicada por el parámetro "b" que hemos calculado.

Ejemplo: vamos a calcular la recta de regresión de la siguiente serie de datos de altura y peso de los alumnos de una clase. Vamos a considerar que la altura es la variable independiente "x" y que el peso es la variable dependiente "y" (podíamos hacerlo también, al contrario):

Alumno	Estatura	Peso	Alumno	Estatura	Peso	Alumno	Estatura	Peso
x	x	x	x	x	x	x	x	x
Alumno 1	1,25	32	Alumno 11	1,25	33	Alumno 21	1,25	33
Alumno 2	1,28	33	Alumno 12	1,28	35	Alumno 22	1,28	34
Alumno 3	1,27	34	Alumno 13	1,27	34	Alumno 23	1,27	34
Alumno 4	1,21	30	Alumno 14	1,21	30	Alumno 24	1,21	31
Alumno 5	1,22	32	Alumno 15	1,22	33	Alumno 25	1,22	32
Alumno 6	1,29	35	Alumno 16	1,29	34	Alumno 26	1,29	34
Alumno 7	1,30	34	Alumno 17	1,30	35	Alumno 27	1,30	34
Alumno 8	1,24	32	Alumno 18	1,24	32	Alumno 28	1,24	31
Alumno 9	1,27	32	Alumno 19	1,27	33	Alumno 29	1,27	35
Alumno 10	1,29	35	Alumno 20	1,29	33	Alumno 30	1,29	34

El **parámetro "b"** viene determinado por:

$$b = \frac{\left(\frac{1}{30}\right) * 1{,}034}{\left(\frac{1}{30}\right) * 0{,}00856} = 40{,}265$$

Y el **parámetro "a"** por:

$$a = 33{,}1 - (40{,}265 * 1{,}262) = -17{,}714$$

Por lo tanto, la **recta** que mejor se ajusta a esta serie de datos es:

$$y = -17{,}714 + (40{,}265 * x)$$

Esta recta define un valor de la variable dependiente (peso), para cada valor de la variable independiente (estatura):

Estatura	Peso
x	x
1,20	30,6
1,21	31,0
1,22	31,4
1,23	31,8
1,24	32,2
1,25	32,6
1,26	33,0
1,27	33,4
1,28	33,8
1,29	34,2
1,30	34,6

TAREA PROPUESTA POR TU CATEDRÁTICO(A).