LECCION 1ª
Introducción a la Estadística Descriptiva

La **estadística descriptiva** es una ciencia que analiza series de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc) y trata de extraer conclusiones sobre el comportamiento de estas variables.

Las **variables** pueden ser de dos tipos:

Variables cualitativas o atributos: no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).

Variables cuantitativas: tienen valor numérico (edad, precio de un producto, ingresos anuales).

Las **variables** también se pueden clasificar en:

Variables unidimensionales: sólo recogen información sobre una característica (por ejemplo: edad de los alunmos de una clase).

Variables bidimensionales: recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).

Variables pluridimensionales: recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

Por su parte, las **variables cuantitativas** se pueden clasificar en discretas y continuas:

Discretas: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3....,etc, pero, por ejemplo, nunca podrá ser 3,45).

Continuas: pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h...etc.

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase,

cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo.

Población: conjunto de todos los individuos (personas, objetos, animales, etc.) que porten información sobre el fenómeo que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

Muestra: subconjunto que seleccionamos de la población. Así, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

LECCION 2ª
Distribución de frecuencia

La **distribución de frecuencia** es la representación estructurada, en forma de tabla, de toda la información que se ha recogido sobre la variable que se estudia.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
X_1	n_1	n_1	$f_1 = n_1 / n$	f_1
X_2	n_2	$n_1 + n_2$	$f_2 = n_2 / n$	$f_1 + f_2$
...	...	...	...	...
X_{n-1}	n_{n-1}	$n_1 + n_2 + .. + n_{n-1}$	$f_{n-1} = n_{n-1} / n$	$f_1 + f_2 + .. + f_{n-1}$
X_n	n_n	Σn	$f_n = n_n / n$	Σf
Siendo **X** los distintos valores que puede tomar la variable.				
Siendo **n** el número de veces que se repite cada valor.				
Siendo **f** el porcentaje que la repetición de cada valor supone sobre el total				

Veamos **un ejemplo**:

Medimos la altura de los niños de una clase y obtenemos los siguientes resultados (cm):

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

Si presentamos esta información estructurada obtendríamos la siguiente **tabla de frecuencia**:

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
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%

Si los valores que toma la variable son muy diversos y cada uno de ellos se repite muy pocas veces, entonces conviene agruparlos por intervalos, ya que de otra manera obtendríamos una tabla de frecuencia muy extensa que aportaría muy poco valor a efectos de síntesis. (tal como se verá en la siguiente lección).

LECCION 3ª
Distribuciones de frecuencia agrupada

Supongamos que medimos la estatura de los habitantes de una vivienda y obtenemos los siguientes resultados (cm):

Habitante	Estatura	Habitante	Estatura	Habitante	Estatura
Habitante 1	1,15	Habitante 11	1,53	Habitante 21	1,21
Habitante 2	1,48	Habitante 12	1,16	Habitante 22	1,59
Habitante 3	1,57	Habitante 13	1,60	Habitante 23	1,86
Habitante 4	1,71	Habitante 14	1,81	Habitante 24	1,52
Habitante 5	1,92	Habitante 15	1,98	Habitante 25	1,48
Habitante 6	1,39	Habitante 16	1,20	Habitante 26	1,37
Habitante 7	1,40	Habitante 17	1,42	Habitante 27	1,16
Habitante 8	1,64	Habitante 18	1,45	Habitante 28	1,73
Habitante 9	1,77	Habitante 19	1,20	Habitante 29	1,62
Habitante 10	1,49	Habitante 20	1,98	Habitante 30	1,01

Si presentáramos esta información en una tabla de frecuencia obtendriamos una tabla de 30 líneas (una para cada valor), cada uno de ellos con una frecuencia absoluta de 1 y con una frecuencia relativa del 3,3%. Esta tabla nos aportaría escasa imformación

En lugar de ello, preferimos agrupar los datos por intervalos, con lo que la información queda más resumida (se pierde, por tanto, algo de información), pero es más manejable e informativa:

Estatura	Frecuencias absolutas		Frecuencias relativas	
Cm	Simple	Acumulada	Simple	Acumulada
1,01 - 1,10	1	1	3,3%	3,3%
1,11 - 1,20	3	4	10,0%	13,3%
1,21 - 1,30	3	7	10,0%	23,3%
1,31 - 1,40	2	9	6,6%	30,0%
1,41 - 1,50	6	15	20,0%	50,0%
1,51 - 1,60	4	19	13,3%	63,3%
1,61 - 1,70	3	22	10,0%	73,3%
1,71 - 1,80	3	25	10,0%	83,3%
1,81 - 1,90	2	27	6,6%	90,0%
1,91 - 2,00	3	30	10,0%	100,0%

El número de tramos en los que se agrupa la información es una decisión que debe tomar el analista: la regla es que mientras más tramos se

utilicen menos información se pierde, pero puede que menos representativa e informativa sea la tabla.

LECCION 4ª
Medidas de posición central

Las medidas de posición nos facilitan información sobre la serie de datos que estamos analizando. Estas medidas permiten conocer diversas características de esta serie de datos.

Las **medidas de posición** son de dos tipos:

a) Medidas de posición central: informan sobre los valores medios de la serie de datos.

b) Medidas de posición no centrales: informan de como se distribuye el resto de los valores de la serie.

a) Medidas de posición central

Las principales medidas de posición central son las siguientes:

1.- Media: es el valor medio ponderado de la serie de datos. Se pueden calcular diversos tipos de media, siendo las más utilizadas:

a) Media aritmética: se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos de la muestra:

$$Xm = \frac{(X_1 * n_1) + (X_2 * n_2) + (X_3 * n_3) + ...+ (X_{n-1} * n_{n-1}) + (X_n * n_n)}{n}$$

b) Media geométrica: se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto fiinal se le calcula la raíz "n" (siendo "n" el total de datos de la muestra).

$$X = (X_1^{n1} * X_2^{n2} * X_3^{n3} * .... * X_n^{nn})^{(1/n)}$$

Según el tipo de datos que se analice será más apropiado utilizar la media aritmética o la media geométrica.

La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.

Lo más positivo de la media es que en su cálculo se utilizan todos los valores de la serie, por lo que no se pierde ninguna información.

Sin embargo, presenta el problema de que su valor (tanto en el caso de la media aritmética como geométrica) se puede ver muy influido por valores extremos, que se aparten en exceso del resto de la serie. Estos valores anómalos podrían condicionar en gran medida el valor de la media, perdiendo ésta representatividad.

2.- Mediana: es el valor de la serie de datos que se sitúa justamente en el centro de la muestra (un 50% de valores son inferiores y otro 50% son superiores).

No presentan el problema de estar influido por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos (no pondera cada valor por el número de veces que se ha repetido).

3.- Moda: es el valor que más se repite en la muestra.

Ejemplo: vamos a utilizar la tabla de distribución de frecuencias con los datos de la estatura de los alumnos que vimos en la lección 2ª.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
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%

Vamos a calcular los valores de las distintas posiciones centrales:

1.- Media aritmética:

$$Xm = \frac{(1{,}20*1) + (1{,}21*4) + (1{,}22*4) + (1{,}23*2) + \ldots\ldots\ldots + (1{,}29*3) + (1{,}30*3)}{30}$$

Luego:

Xm = 1,253

Por lo tanto, la estatura media de este grupo de alumnos es de 1,253 cm.

2.- Media geométrica:

X = ((1,20^ 1) * (1,21^4) * (1,22^ 4) * .....* (1,29^3)* (1,30^3)) ^ (1/30)

Luego:

Xm = 1,253

En este ejemplo la media aritmética y la media geométrica coinciden, pero no tiene siempre por qué ser así.

3.- Mediana:

La mediana de esta muestra es 1,26 cm, ya que por debajo está el 50% de los valores y por arriba el otro 50%. Esto se puede ver al analizar la columna de frecuencias relativas acumuladas.

En este ejemplo, como el valor 1,26 se repite en 3 ocasiones, la media se situaría exactamente entre el primer y el segundo valor de este grupo, ya que entre estos dos valores se encuentra la división entre el 50% inferior y el 50% superior.

4.- Moda:

Hay 3 valores que se repiten en 4 ocasiones: el 1,21, el 1,22 y el 1,28, por lo tanto esta seria cuenta con 3 modas.

LECCION 5ª
Medidas de posición no central

Medidas de posición no centrales

Las medidas de posición no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suelen utilizar una serie de valores que dividen la muestra en tramos iguales:

Cuartiles: son 3 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.

Deciles: son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.

Percentiles: son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

Ejemplo: Vamos a calcular los cuartiles de la serie de datos referidos a la estatura de un grupo de alumnos (lección 2ª). Los deciles y centiles se calculan de igual manera, aunque haría falta distribuciones con mayor número de datos.

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%

1º cuartil: es el valor 1,22 cm, ya que por debajo suya se situa el 25% de la frecuencia (tal como se puede ver en la columna de la frecuencia relativa acumulada).

2º cuartil: es el valor 1,26 cm, ya que entre este valor y el 1º cuartil se situa otro 25% de la frecuencia.

3º cuartil: es el valor 1,28 cm, ya que entre este valor y el 2º cuartil se sitúa otro 25% de la frecuencia. Además, por encima suya queda el restante 25% de la frecuencia.

Atención: cuando un cuartil recae en un valor que se ha repetido más de una vez (como ocurre en el ejemplo en los tres cuartiles) la medida de posición no central sería realmente una de las repeticiones.

LECCION 6ª
Medidas de dispersión

Estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Existen diversas **medidas de dispersión**, entre las más utilizadas podemos destacar las siguientes:

1.- Rango: mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo.

2.- Varianza: Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las difrencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatorio obtenido se divide por el tamaño de la muestra.

$$S^2_x = \frac{\Sigma (x_i - xm)^2 * n_i}{n}$$

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

3.- Desviación típica: Se calcula como raíz cuadrada de la varianza.

4.- Coeficiente de varización de Pearson: se calcula como cociente entre la desviación típica y la media.

Ejemplo: vamos a utilizar la serie de datos de la estatura de los alumnos de una clase (lección 2ª) y vamos a calcular sus medidas de dispersión.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
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%

1.- Rango: Diferencia entre el mayor valor de la muestra (1,30) y el menor valor (1,20). Luego el rango de esta muestra es 10 cm.

2.- Varianza: recordemos que la media de esta muestra es 1,253. Luego, aplicamos la fórmula:

$$S^2_x = \frac{((1,20-1,253)^2 * 1) + ((1,21-1,253)^2 * 4) + ((1,22-1,253)^2 * 4) + ... + ((1,30-1,253)^2 * 3)}{30}$$

Por lo tanto, la varianza es 0,0010

3.- Desviación típica: es la raíz cuadrada de la varianza.

$$\sigma = (S^2_x)^{(1/2)}$$

Luego:

$$\sigma = (0,010)^{(1/2)} = 0,0320$$

4.- Coeficiente de variación de Pearson: se calcula como cociente entre la desviación típica y la media de la muestra.

Cv = 0,0320 / 1,253

Luego,

Cv = 0,0255

El interés del coeficiente de variación es que al ser un porcentaje permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desvación típica, ya que viene expresada en las mismas unidas que los datos de la serie.

Por ejemplo, para comparar el nivel de dispersión de una serie de datos de la altura de los alumnos de una clase y otra serie con el peso de dichos alumnos, no se puede utilizar las desviaciones típicas (una viene vienes expresada en cm y la otra en kg). En cambio, sus coeficientes de variación son ambos porcentajes, por lo que sí se pueden comparar.

LECCION 7ª
Medidas de forma: Grado de concentración

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.

a) Concentración

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

Este índice se calcula aplicando la siguiente fórmula:

$$IG = \frac{\Sigma (p_i - q_i)}{\Sigma p_i}$$

(i toma valores entre 1 y n-1)

En donde p_i mide el pocentaje de individuos de la muestra que presentan un valor igual o inferior al de x_i.

$$p_i = \frac{n_1 + n_2 + n_3 + ... + n_i}{n} \times 100$$

Mientras que q_i se calcula aplicando la siguiente fórmula:

$$q_i = \frac{(X_1 * n_1) + (X_2 * n_2) + ... + (X_i * n_i)}{(X_1 * n_1) + (X_2 * n_2) + ... + (X_n * n_n)} \times 100$$

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

IG = 0 : concentración mínima. La muestra está unifomemente 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 Indice 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
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 Indice de Gini:

X_i	n_i	Σn_i	p_i	$X_i * n_i$	$\Sigma X_i * n_i$	q_i	$p_i - q_i$
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
		Σp_i (entre 1 y n-1) =	435,0		$\Sigma (p_i - q_i)$ (entre 1 y n-1) =		83,99

Por lo tanto:

$$IG = 83,99 / 435,0 = 0,19$$

Un Indice 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
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:

X_i	n_i	Σn_i	p_i	$X_i * n_i$	$\Sigma X_i * n_i$	q_i	$p_i - q_i$
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
		Σp_i (entre 1 y n-1) =	407,5		$\Sigma (p_i - q_i)$ (entre 1 y n-1) =		136,69

El **Indice Gini** sería:

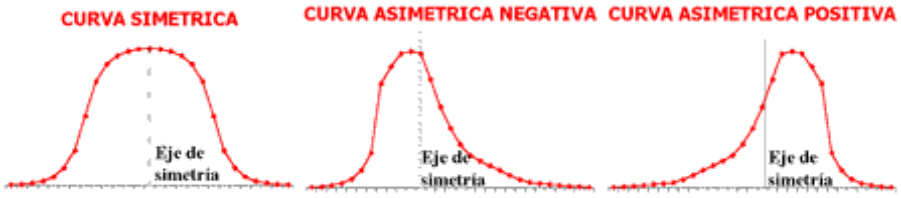
$$IG = 136,69 / 407,5 = 0,34$$

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

LECCION 8ª
Medidas de forma: Coeficiente de Asimetría

b) 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:

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

Los resultados pueden ser los siguientes:

$g_1 = 0$ (distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media)

$g_1 > 0$ (distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda)

$g_1 < 0$ (distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha)

Ejemplo: Vamos a calcular el Coefiente 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

Σ((xi - x)^3)*ni	Σ((xi - x)^2)*ni
0,000110	0,030467

Luego:

$$g_1 = \frac{(1/30) * 0{,}000110}{(1/30) * (0{,}030467)^{(3/2)}} = -0{,}1586$$

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).

LECCION 9ª
Medidas de forma: Coeficiente de Curtosis

c) Curtosis

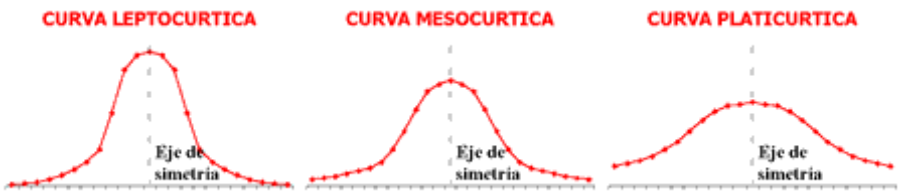
El **Coeficiente de Curtosis** analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución.

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

Distribución mesocúrtica: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).

Distribución leptocúrtica: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

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:

$g_2 = 0$ (distribución mesocúrtica).

$g_2 > 0$ (distribución leptocúrtica).

$g_2 < 0$ (distribución platicúrtica).

Ejemplo: Vamos a calcular el Coefiente 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

$\Sigma((x_i - x_m)^4)*n_i$	$\Sigma((x_i - x_m)^2)*n_i$
0,00004967	0,03046667

Luego:

$$g_2 = \frac{(1/30) * 0{,}00004967}{((1/30) * (0{,}03046667))^2} - 3 = -1{,}39$$

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.

LECCION 10ª
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	y_1	y_2	.....	y_{m-1}	y_m
x_1	$n_{1,1}$	$n_{1,2}$		$n_{1,m-1}$	$n_{1,m}$
x_2	$n_{2,1}$	$n_{2,2}$		$n_{2,m-1}$	$n_{2,m}$
.....					
x_{n-1}	$n_{n-1,1}$	$n_{n-1,2}$		$n_{n-1,m-1}$	$n_{n-1,m}$
x_n	$n_{n,1}$	$n_{n,2}$		$n_{n,m-1}$	$n_{n,m}$

Las "x" representan una de las variables y las "y" la otra variable. En cada intersección de una 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
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.

LECCION 11ª
Distribuciones marginales

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de como 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	$n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
.....	...
x_{n-1}	$n_{n-1.}$
x_n	$n_{n.}$

Distribución marginal de Y

Y	$n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
.....	...
y_{m-1}	$n_{.m-1}$
y_m	$n_{.m}$

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
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:

x

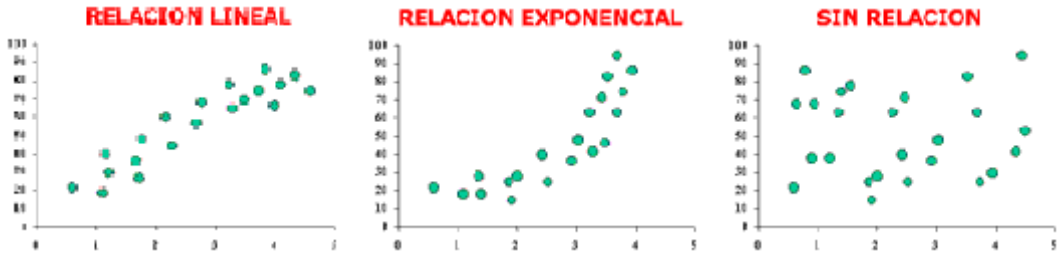
Variable	Frecuencias absolutas		Frecuencias relativas	
(Peso)	Simple	Acumulada	Simple	Acumulada
31	6	6	20,0%	20,0%
32	6	12	20,0%	40,0%
33	6	18	20,0%	60,0%
34	7	25	23,3%	83,3%
35	5	30	16,6%	100,0%

LECCION 12ª
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 varables es lineal (es decir, si representaramos en un gáfico 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 que forma describen.

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 produto de las varianzas de "x" y de "y", y a este produto 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
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:

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

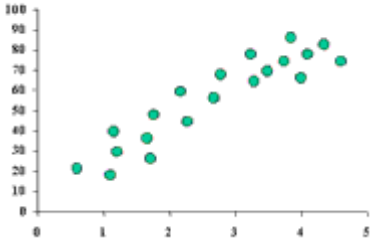
Luego,

$$r = 0{,}719$$

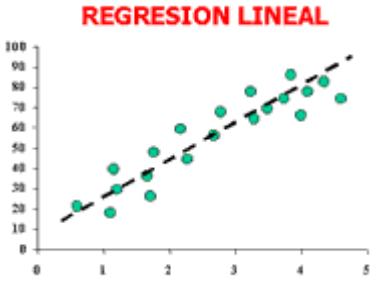
Por lo tanto, la correlación existente entre estas dos variables es elevada (0,7) y de signo posítivo

LECCION 13ª
Regresión lineal

Representamos en un gráfico los pares de valores de una distribución bidimensional: la variable "x" en el eje horizontal o eje de abcisa, 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:

$$a = y_m - (b * x_m)$$

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
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:

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

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:

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

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

Estatura	Peso
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