ESTADÍSTICA DESCRIPTIVA

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 alumnos 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 fenómeno 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.

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	x	x	x	x
X1	n1	n1	f1 = n1 / n	f1
X2	n2	n1 + n2	f2 = n2 / n	f1 + f2
...	...	...	...	...
Xn-1	nn-1	n1 + n2 +..+ nn-1	fn-1 = nn-1 / n	f1 + f2 +..+fn-1
Xn	nn	S n	fn = nn / n	S f
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
x	x	x	x	x	x
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
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%

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).

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
x	x	x	x	x	x
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 información.

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
x	x	x	x	x
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.

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:

$$Xm = \frac{(X1 * n1) + (X2 * n2) + (X3 * n3) + ...+ (Xn-1 * nn-1) + (Xn * nn)}{n}$$

b) Media geométrica: se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final 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 antes.

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%

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:

$$X = ((1,20^{1}) * (1,21^{4}) * (1,22^{4}) * \ldots\ldots * (1,29^{3}) * (1,30^{3}))^{(1/30)}$$

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.

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. Los deciles y centiles se calculan de igual manera, aunque harían 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 sitúa 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 sitúa 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.

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 sumatoria de las diferencias al cuadrado entre cada

valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. La sumatoria obtenida 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 variación 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**
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%

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:

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

Por lo tanto, la varianza es 0,0010

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

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

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 desviación 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.