ENAC Caritas
Centro de Formación Técnica

ESTADÍSTICAS

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.

VARIABLES

Variables cualitativas o atributos
Variables cuantitativas

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).

ESTADÍSTICAS

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

unidimensionales
bidimensionales
pluridimensionales

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).

ESTADÍSTICAS

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

Discretas

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.

ESTADÍSTICAS

CONCEPTOS ESTADISTICOS

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

Individuo

Población

Muestra

Parámetro

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.

ESTADÍSTICAS

- **Parámetro:** Se conoce como parámetro al dato que se considera como imprescindible y orientativo para lograr evaluar o valorar una determinada situación. A partir de un parámetro, una cierta circunstancia puede comprenderse o ubicarse en perspectiva. Por dar algunos ejemplos concretos: “Si nos basamos en los parámetros habituales, resultará imposible comprender esta situación”, “El paciente está evolucionando de acuerdo a los parámetros esperados”, “Estamos investigando pero no hay parámetros que nos permitan establecer una relación con el caso anterior”, “La actuación del equipo en el torneo local es el mejor parámetro para realizar un pronóstico sobre su participación en el campeonato mundial”.

Distribuciones 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 (Valor)	Frecuencias absolutas Simple	Frecuencias absolutas Acumulada	Frecuencias relativas Simple	Frecuencias relativas 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
Alumno10	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%

ESTADISTICAS

- 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 un condominio 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
Habitante10	1,49	Habitante 20	1,98	Habitante 30	1,01

ESTADISTICAS

- Si presentáramos esta información en una tabla de frecuencia obtendríamos 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	5	6	16,6%	20,0%
1,21 - 1,30	1	7	3,3%	23,3%
1,31 - 1,40	3	10	10,0%	33,3%
1,41 - 1,50	5	15	16,6%	50,0%
1,51 - 1,60	5	20	16,6%	66,6%
1,61 - 1,70	2	22	6,6%	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 %

ESTADISTICAS

- 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.

ESTADISTICAS

Medidas de tendencia central: media, mediana y la moda

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 cómo se distribuye el resto de los valores de la serie.

ESTADISTICAS

a) Medidas de posición central

Las principales medidas de posición central son las siguientes:

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

Características de la media:

1.- Es una medida totalmente numérica o sea sólo puede calcularse en datos de características cuantitativas.

2.- En su cálculo se toman en cuenta todos los valores de la variable.

3.- Es lógica desde el punto de vista algebraico.

4.- La media aritmética es altamente afectada por valores extremos.

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:

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

ENAC caritas
Centro de Formación Técnica

ESTADISTICAS

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).

-Para el cálculo de la mediana interesa que los valores estén ordenados de menor a mayor.

Características de la mediana:

A.- no pondera cada valor por el número de veces que se ha repetido.

B.- Tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas.

C.- Es de significado estadístico menos intuitivo que la media aritmética.

4.- Su aplicación se ve limitada, ya que solo considera el orden jerárquico de los datos y no alguna propiedad propia de los datos, como en el caso de la media aritmética.

Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales.

7, 8, 9, 10, 11, 12

Me = 9.5

ESTADISTICAS

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 :

Vamos a calcular los valores de las distintas posiciones centrales:

Variable	Frecuencias absolutas		
(Valor)	Simple	Acumulada	x . f
x	fi	F	f
1,20	1	1	1,20
1,21	4	5	4,84
1,22	4	9	4,88
1,23	2	11	2,46
1,24	1	12	1,24
1,25	2	14	2,50
1,26	3	17	3,78
1,27	3	20	3,81
1,28	4	24	5,12
1,29	3	27	3,87
1,30	3	30	3,90
	30		37.60

1.- MEDIA ARITMÉTICA:

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

- **Xm = 1,2533**

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

También tenemos que la media se puede calcular de esta forma:

$$\text{MEDIA} = \sum X\, f / n$$

$$= 37{,}6/30$$

$$= 1{,}2533 \text{ cm.}$$

2.- 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%.
- También se puede calcular con n/2= 15. Buscamos en la frecuencia acumulada y si no se encuentra ese mismo numero el que le sigue en valor. En este caso 17. Y buscamos el valor de x = 1,26
- 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.

3.- MODA:

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

X	f	F	x f
13	3	3	39
14	14	17	196
15	23	40	345
16	10	50	160
17	5	55	85
18	4	59	72
19	1	60	19
	60		916

ENAC Caritas
Centro de Formación Técnica

ESTADISTICAS

MEDIA ARITMETICA

MEDIA = ∑ X f / n

= 916/30

= 15,26 años

MEDIANA

Posición : n/2 = 60/2 = 30 , en Frecuencia absoluta, significa 15 años, en valor x

MODA: La que mas se repite es la frecuencia 23, es decir en x = 15 años.

2	4	7	2	2	5	10	9	9	7

x	f	F	x f
2	3	3	6
4	1	4	4
5	1	5	5
7	2	7	14
9	2	9	18
10	1	10	10
	10		57

ESTADISTICAS

MEDIA ARITMETICA

MEDIA = $\sum x f / n$

= $57/10$

= 5,7 respuestas.

MEDIANA

Posición : $n/2 = 10/2 = 5$, en Frecuencia absoluta, significa 5 años, en valor x

MODA: La que mas se repite es la frecuencia 3, es decir en x = 2.

Datos agrupados en tabla de frecuencia

Escala	fi
(0;3)	75
(3;6)	458
(6;9)	1932
(9;10)	521

ESTADISTICAS

Escala	**x**	**fi**	**F**	**x f**
(0;3)	1.5	75	75	112,5
(3;6)	4.5	458	533	2.061
(6;9)	**7.5**	**1932**	**2465**	**14.490**
(9;10)	9.5	521	2986	4.949,5
	Promedio de los limites	2986		21.613

ESTADISTICAS

- MEDIA ARITMETICA

= 21.613/2.986

=7,23 puntos

- MEDIANA. Posición n/2 = 2.986/2 = 1.493

$Me = Li + A \left[(n/2 - F_{i-1}) / f_i \right]$

- $Me = 6 + 3 \left[(2986/2 - 533) / 1932 \right]$
- $Me = 6 + 3 \times (0.49689)$
- $Me = 7,49$

ESTADISTICAS

- MODA

$Mo = Li + A \times [(f_i - f_{i-1}) / (f_i - f_{i-1}) + (f_i - f_{i+1})]$

$Mo = 6 + 3 [(1932 - 458) / (1932 - 458) + (1932 - 521]$

$Mo = 6 + 3 [(1474) / (1474) + (1411)]$

$Mo = 6 + 3 [(1474) / (2885)]$

$Mo = 6 + 3 [0{,}5109185]$

$Mo = 6 + 1{,}532755$

$Mo = 7{,}53$

ESTADISTICAS

- Medidas de posición no central
- 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.
- **Quintiles:**son 4 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cinco tramos iguales, en los que cada uno de ellos concentra el 20% 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 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%

ENAC Caritas
Centro de Formación Técnica

ESTADISTICAS

- **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.

ESTADISTICAS

- Medidas de dispersión - rango, varianza, desviación típica y coeficiente de variación
- 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.

ESTADISTICAS

- **2.- VARIANZA:** Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las diferencias 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.
- 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.

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

ESTADISTICAS

- **3.- DESVIACIÓN ESTÁNDAR O TÍPICA**: Se calcula como raíz cuadrada de la varianza. Mide cuanto se alejan los valores de un conjunto de datos respecto de su media. En general se estima que la desviación estándar es mejor indicador de dispersión de datos que 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 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%

ENAC Centro de Formación Técnica

ESTADISTICAS

- **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:

Por lo tanto, la varianza es 0,0010

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

3.- DESVIACIÓN TÍPICA: es la raíz cuadrada de la varianza.

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

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