1. Estadística Descriptiva

ÍNDICE

MOTIVACIÓN

Cuando se está analizando una variable y se dispone de un conjunto de datos, es muy importante conocer en profundidad dicha variable, y la forma en la que se han generado los datos de la misma. Además, resulta fundamental sintetizar la información para poder describir los datos.

La estadística descriptiva, que se presenta en esta Unidad Didáctica, trata de resumir la información que proporciona un conjunto de datos mediante una serie de procedimientos. Se distinguen dos tipos de procedimientos: la elaboración de tablas y el cálculo de una serie de medidas numéricas que permiten resumir la información disponible.

PROPÓSITOS

Los principales propósitos de esta Unidad Didáctica son:

- Aprender a elaborar tablas que resuman la información que proporcionan un conjunto de datos y saber interpretar correctamente dichas tablas.
- Aprender a calcular medidas que sinteticen la información sobre un conjunto de datos y saber interpretar correctamente dichas medidas.
- Aprender a realizar un análisis de la dependencia lineal entre dos variables y saber interpretar correctamente dicho análisis.

PREPARACIÓN PARA LA UNIDAD

Esta Unidad Didáctica comienza con una descripción de los distintos tipos de variables. A continuación, se presentan los diferentes tipos de frecuencias y las distribuciones de frecuencias. Finalmente, se explican las diferentes medidas que permiten resumir la información que proporciona un conjunto de datos, haciendo especial énfasis en la interpretación de las mismas.

En la segunda parte que corresponde al análisis descriptivo bidimensional se analiza el comportamiento conjunto de dos variables. El análisis de la relación existente entre dos variables se puede realizar siguiendo dos enfoques diferentes.

- Regresión: consiste en analizar la forma de la dependencia existente entre las variables.
- Correlación: consiste en analizar el grado de dependencia existente entre las variables.

En esta Unidad Didáctica aprenderás a calcular medidas que permiten resumir la información que proporciona un conjunto de datos, rectas de regresión y coeficientes de correlación. No se trata solo de que sepas calcularlos, sino también de que aprendas a interpretar correctamente los resultados obtenidos.

1. VARIABLES CUANTITATIVAS Y VARIABLES CUALITATIVAS

Los fenómenos pueden dar lugar a observaciones de naturaleza cuantitativa o cualitativa. Los fenómenos de naturaleza cuantitativa son aquellos cuyas observaciones tienen carácter numérico. Los fenómenos de naturaleza cualitativa son aquellos cuyas observaciones no tienen carácter numérico.

Se distinguen, por tanto, **variables cuantitativas** que tienen carácter numérico (por ejemplo, la estatura) y **variables cualitativas** o atributos que no tienen carácter numérico (por ejemplo, la profesión).

Las variables cuantitativas se representan por letras mayúsculas (las últimas del abecedario), $X, Y...$ y sus concreciones (datos) con minúsculas: $x_1, x_2, ..., x_n$.

Se distinguen dos tipos de variables cuantitativas.

- **Variables discretas:** son aquellas que toman valores aislados y no pueden tomar ningún valor entre dos consecutivos dados. Por ejemplo, el número de hijos de una familia.
- **Variables continuas:** son aquellas que pueden tomar cualquier valor. Por ejemplo, la estatura de una persona.

2. TIPOS DE FRECUENCIA

En esta sección se definen los diferentes tipos de frecuencias, que como se verá en la sección siguiente, se colocan en una tabla.

Frecuencia absoluta de un valor es el número de veces que el valor se repite. Se representa por n_i.

La suma de todas las frecuencias absolutas es igual al número total de datos de la distribución que se representará por N.

$$\sum_{i=1}^{n} n_i = N$$

El operador $\sum$ nos permite expresar una suma de forma abreviada.

Frecuencia relativa de un valor es el cociente entre la frecuencia absoluta y el número total de datos N. Se representa por f_i.

$$f_i = \frac{n_i}{N}$$

La suma de todas las frecuencias relativas es igual a 1.

$$\sum_{i=1}^{n} f_i = 1$$

La frecuencia relativa de un valor multiplicada por 100 nos indica el porcentaje de datos iguales a ese valor.

Frecuencia absoluta acumulada de un valor es el número de datos menores o iguales que el valor considerado. Se representa por N_i.

Se calcula sumando la frecuencia absoluta de ese valor y de todos los anteriores a él.

$N_1 = n_1$

$N_2 = n_1 + n_2$

$N_n = n_1 + n_2 + ... + n_n = N$

Frecuencia relativa acumulada de un valor es el cociente entre la frecuencia absoluta acumulada y el número total de datos. Se representa por F_i.

$$F_i = \frac{N_i}{N}$$

La frecuencia relativa acumulada del último valor es igual a 1.

$$F_n = 1$$

La frecuencia relativa acumulada multiplicada por 100 nos indica el porcentaje de datos iguales al valor considerado e inferiores a el.

3. DISTRIBUCIONES DE FRECUENCIAS

Para expresar de forma resumida la información se utilizará la distribución de frecuencias que es el conjunto formado por los valores que toma la variable junto con sus correspondientes frecuencias.

Se distinguen dos tipos de distribuciones de frecuencias:

Distribuciones no agrupadas

Cada valor de la variable aparece con su frecuencia asociada. La distribución se representa mediante una tabla de frecuencias. En la primera columna se presentan los valores de la variable ordenados de menor a mayor, y en las siguientes las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

Distribuciones agrupadas en intervalos

Cuando el número de valores distintos que ha tomado la variable es muy grande se agrupan en intervalos. Esto ocasiona pérdida de información pero se gana en simplicidad y operatividad.

Consideramos intervalos abiertos por la izquierda y cerrados por la derecha. El límite inferior de un intervalo se representará por L_i y el límite superior por L_{i+1}. Por tanto, los intervalos son de la forma $(L_i, L_{i+1}\ ]$.

La **amplitud** de un intervalo, que se representa por c_i, es la diferencia entre el límite superior y el límite inferior del mismo.

$$c_i = L_{i+1} - L_i$$

Los intervalos pueden ser de amplitud constante o de amplitud variable.

Como representante de cada intervalo elegimos su punto medio que se denomina **marca de clase**. La marca de clase se representa por x_i.

$$x_i = \frac{L_i + L_{i+1}}{2}$$

La distribución se representa mediante una tabla de frecuencias. En la primera columna se presentan los intervalos, en la segunda la marca de clase, y en las siguientes las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

Ejemplo 1. En una empresa con 20 empleados, 3 trabajadores perciben un salario mensual de 1021 euros, 5 de 841 euros, 2 de 1232 euros, 7 de 901 euros y 3 de 720 euros. Obtener la tabla de frecuencias completa.

x_i	n_i	f_i	N_i	F_i
720	3	3/20	3	3/20
841	5	5/20	8	8/20
901	7	7/20	15	15/20
1021	3	3/20	18	18/20
1232	2	2/20	20	1

Siendo:

n_i la frecuencia absoluta, que nos indica el número de datos iguales al valor considerado.

f_i la frecuencia relativa, que es el cociente entre la frecuencia absoluta y el número total de datos.

N_i la frecuencia absoluta acumulada, que se calcula sumando la frecuencia absoluta de ese valor y de todos los anteriores a el.

F_i la frecuencia relativa, acumulada que es el cociente entre la frecuencia absoluta acumulada y el número total de datos.

Obsérvese que:

1. Los datos se han ordenado de menor a mayor.
2. La suma de todas las frecuencias absolutas es igual a 20, que es el número total de datos de la distribución.
3. La suma de todas las frecuencias relativas es igual a 1.
4. La frecuencia absoluta acumulada del último valor es igual a 20, que coincide con el número total de datos de la distribución.
5. La frecuencia relativa acumulada del último valor es igual a 1.

4. MEDIDAS DE POSICIÓN

Las medidas son valores que nos permiten resumir la información que proporcionan un conjunto de datos. En esta sección, se presentan las principales medidas de posición: media aritmética, mediana y moda.

MEDIA ARITMÉTICA

La media aritmética, que se representará por $\bar{x}$, es la suma de todos los valores de la variable dividida entre el número total de datos.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

El operador $\sum$ nos permite expresar una suma de forma abreviada. $\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$.

Ejemplo 2. Se han observado los siguientes datos de una variable:

12; 10; 11; 14; 12; 11; 10; 12; 12; 14

Determinar la media aritmética.

La media es:

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{12+10+...+14}{10} = 11{,}80$$

Si la distribución está tabulada, calcularemos la media utilizando la siguiente expresión:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i n_i}{N}$$

Siendo x_i los valores de la variable, si la distribución no está agrupada en intervalos o la marca de clase de cada intervalo, si la distribución está agrupada en intervalos.

MEDIANA

Es el valor de la variable que deja a su derecha y a su izquierda el mismo número de observaciones, supuesto que los datos están ordenados de menor a mayor. Se representará por M_e. Si el número de datos es par, se calcula como la media aritmética de los dos centrales.

Cálculo

Distribuciones no agrupadas

- **Frecuencias unitarias**

Si no hay datos repetidos, se ordenan de menor a mayor. Se distinguen dos situaciones:

Si el número de datos es impar, la mediana es el valor central.

Ejemplo 3. Considere los siguientes datos:

2, 7, 13, 5, 14, 10, 1.

Determinar la mediana.

En primer lugar ordenamos los datos de menor a mayor. 1, 2, 5, 7, 10, 13, 14. La mediana es 7, puesto que es el valor que deja a su izquierda y a su derecha el mismo número de observaciones.

Si el número de datos es par, la mediana es la media aritmética de los dos centrales.

Ejemplo 4. Considere los siguientes datos:

2, 7, 13, 5, 14, 9, 10, 1.

Determinar la mediana.

En primer lugar ordenamos los datos de menor a mayor. 1, 2, 5, 7, 9, 10, 13, 14. En este caso el número de datos es par. La mediana es la media aritmética de los dos centrales $M_e = \frac{7+9}{2} = 8$.

- **Frecuencias diferentes**

Para calcular la mediana de la distribución se procede de la siguiente forma:

Se calculan las frecuencias absolutas acumuladas N_i.

Se calcula $\frac{N}{2}$ y se busca ese valor en la columna de las frecuencias absolutas acumuladas. Se pueden presentar dos posibles situaciones:

- $\frac{N}{2}$ coincide con la frecuencia absoluta acumulada de un valor. En ese caso, la mediana es la media aritmética de ese valor y el valor siguiente.

Ejemplo 5. Considere la siguiente distribución.

x_i	n_i
0	2
1	3
2	4
3	1

Determinar la mediana.

1. Calculamos las frecuencias absolutas acumuladas.

x_i	n_i	N_i
0	2	2
1	3	5
2	4	9
3	1	10

2. Calculamos $\frac{N}{2} = \frac{10}{2} = 5$.

$\frac{N}{2}$ coincide exactamente con una frecuencia absoluta acumulada. La mediana es la media aritmética del valor cuya frecuencia absoluta acumulada coincide con $\frac{N}{2}$ y el valor siguiente. $M_e = \frac{1+2}{2} = 1{,}5$.

- $\frac{N}{2}$ no coincide con la frecuencia absoluta acumulada de un valor. $\frac{N}{2} \neq N_i \quad \forall i$. En ese caso, la mediana es el valor cuya frecuencia absoluta acumulada es la inmediatamente superior a $\frac{N}{2}$.

Ejemplo 6. Considere la siguiente distribución.

x_i	n_i
0	2
1	4
2	3
3	1

Determinar la mediana.

1. Calculamos las frecuencias absolutas acumuladas.

x_i	n_i	Ni
0	2	2
1	4	6
2	3	9
3	1	10

2. Calculamos $\frac{N}{2} = \frac{10}{2} = 5$.

$\frac{N}{2}$ no coincide con ninguna frecuencia absoluta acumulada. La mediana es el valor cuya frecuencia absoluta acumulada es inmediatamente superior a $\frac{N}{2}$. Por tanto, $M_e = 1$.

Distribuciones agrupadas

Para calcular la mediana de la distribución se procede de la siguiente forma:

- Se calculan las frecuencias absolutas acumuladas N_i.
- Se calcula $\frac{N}{2}$ y se busca ese valor en la columna de las frecuencias absolutas acumuladas.

  Se pueden presentar dos posibles situaciones:

  - $\frac{N}{2}$ coincide con la frecuencia absoluta acumulada de un valor. En ese caso la mediana es el limite superior del intervalo cuya frecuencia absoluta acumulada coincide con $\frac{N}{2}$.

Ejemplo 7. Considere la siguiente distribución.

INTERVALOS	n_i
2 - 4	2
4 - 6	3
6 - 8	5

Determinar la mediana.

1. Calculamos las frecuencias absolutas acumuladas.

INTERVALOS	n_i	N_i
2 - 4	2	2
4 - 6	3	5
6 - 8	5	10

2. Calculamos $\frac{N}{2} = \frac{10}{2} = 5$.

$\frac{N}{2}$ coincide exactamente con la frecuencia absoluta acumulada del intervalo 4 – 6. La mediana es el límite superior de ese intervalo. Por tanto, $M_e = 6$.

- $\frac{N}{2}$ no coincide con la frecuencia absoluta acumulada de un valor. $\frac{N}{2} \neq N_i \quad \forall i$. El intervalo cuya frecuencia absoluta acumulada es la inmediatamente superior a $\frac{N}{2}$ es el que contiene a la mediana. La mediana se calcula a partir de la siguiente expresión:

$$M_e = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i$$

Siendo:

L_i el límite inferior del intervalo mediano.

N el número total de datos de la distribución.

N_{i-1} la frecuencia absoluta acumulada del intervalo anterior al mediano.

n_i la frecuencia absoluta del intervalo mediano.

c_i la amplitud del intervalo mediano.

Ejemplo 8. Considere la siguiente distribución.

INTERVALOS	n_i
2 - 4	4
4 - 6	10
6 - 8	40
8 - 10	20
10 - 12	1

Determinar la mediana.

1. Calculamos las frecuencias absolutas acumuladas.

INTERVALOS	n_i	Ni
2 - 4	4	4
4 - 6	10	14
6 - 8	40	54
8 - 10	20	74
10- 12	1	75

2. Calculamos $\frac{N}{2} = \frac{75}{2} = 37{,}5$.

$\frac{N}{2}$ no coincide con ninguna frecuencia absoluta acumulada. El intervalo cuya frecuencia absoluta acumulada es la inmediatamente superior a 37,5 es el intervalo 6 - 8 que es el que contiene a la mediana.

Para calcular la mediana utilizamos la siguiente expresión:

$$M_e = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i = 6 + \frac{37{,}5 - 14}{40} \times 2 = 7{,}175$$

MODA

Es el valor de la variable que más se repite, es decir, el que tiene la mayor frecuencia absoluta asociada. Se representará por M_0.

Cálculo

- Distribuciones no agrupadas

Si la distribución no está agrupada, para determinar la moda se busca en la tabla de frecuencias el valor que tiene la mayor frecuencia absoluta.

Ejemplo 9. Considere la siguiente distribución.

x_i	n_i
0	2
1	4
2	3
3	1

Determinar la moda.

$M_0 = 1$ puesto que es el valor que tiene la mayor frecuencia absoluta.

- Distribuciones agrupadas

El cálculo es diferente para intervalos de amplitud constante y de amplitud variable.

Intervalos de amplitud constante

- Buscamos el intervalo modal que es el que tiene la mayor frecuencia absoluta.
- Determinamos la moda a partir de la siguiente expresión:

$$M_0 = L_i + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} c_i$$

Siendo:

L_i el límite inferior del intervalo modal.

n_i la frecuencia absoluta del intervalo modal.

n_{i-1} la frecuencia absoluta del intervalo anterior al modal.

n_{i+1} la frecuencia absoluta del intervalo posterior al modal.

c_i la amplitud del intervalo modal

Ejemplo 10. Considere la siguiente distribución.

INTERVALOS	n_i
2 - 4	4
4 - 6	10
6 - 8	40
8 - 10	20
10 - 12	1

Determinar la moda.

Obsérvese que los intervalos son de amplitud constante.

En primer lugar, buscamos el intervalo modal que es el que tiene mayor frecuencia absoluta. El intervalo modal es el 6 – 8.

A continuación, calcularemos la moda.

$$M_0 = L_i + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} c_i = 6 + \frac{(40-10)}{(40-10) + (40-20)} \times 2 = 7{,}2$$

Intervalos de amplitud variable

- Buscamos el intervalo modal que es el que tiene la mayor densidad de frecuencia. La densidad de frecuencia, que se representa por d_i, se define como el cociente entre la frecuencia absoluta de un intervalo y la amplitud del mismo.

$$d_i = \frac{n_i}{c_i}$$

- Determinamos la moda a partir de la siguiente expresión:

$$M_0 = L_i + \frac{d_i - d_{i-1}}{(d_i - d_{i-1}) + (d_i - d_{i+1})} c_i$$

Siendo:

L_i el límite inferior del intervalo modal.

d_i la densidad de frecuencia del intervalo modal.

d_{i-1} la densidad de frecuencia del intervalo anterior al modal.

d_{i+1} la densidad de frecuencia del intervalo posterior al modal.

c_i la amplitud del intervalo modal.

Ejemplo 11. Considere la siguiente distribución.

INTERVALOS	n_i
0 - 30	32
30 - 40	8
40 - 50	10

Obsérvese que los intervalos son de amplitud variable.

En primer lugar, buscamos el intervalo modal que es el que tiene mayor densidad de frecuencia.

INTERVALOS	n_i	d i
0 - 30	32	32/30
30 - 40	8	8/10
40 - 50	10	1

El intervalo modal es el 0-30, ya que es el que tiene mayor densidad de frecuencia.

A continuación, calcularemos la moda.

$$M_0 = L_i + \frac{d_i - d_{i-1}}{(d_i - d_{i-1}) + (d_i - d_{i+1})} c_i = 0 + \frac{((32/30) - 0)}{((32/30) - 0)) + ((32/30) - 0{,}8)} \times 30 =$$
$$= 24{,}09$$

5. MEDIDAS DE DISPERSIÓN

MEDIDAS DE DISPERSIÓN ABSOLUTA

Sirven para medir la proximidad o alejamiento que existe entre los valores de la variable y/o alguna medida de posición. A continuación, se presentarán las principales medidas de dispersión absoluta: el rango, la varianza y la desviación típica.

RANGO O RECORRIDO

El rango, que se representa por R, es la diferencia entre el mayor y el menor valor de la variable.

$$R = \max x_i - \min x_i$$

VARIANZA

La varianza, que se representa por s^2, es la suma de los cuadrados de las desviaciones de los valores de la variable respecto a la media dividida entre el número total de datos.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_N - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

Desarrollando el cuadrado de la última expresión se puede obtener otra expresión para la varianza que facilita el cálculo de la misma.

$$s^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2$$

Ejemplo 12. A partir de los datos del ejemplo 2, determinar la varianza.

La varianza es:

$$s^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \left(\frac{\sum_{i=1}^{N} x_i}{N} \right)^2 = \frac{12^2 + 10^2 + ... + 14^2}{10} - (11{,}8)^2 = 1{,}76$$

La varianza mide la dispersión, es decir, la proximidad o alejamiento de los valores de la variable con respecto a la media. Cuanto mayor sea la dispersión mayor será la varianza y por tanto, menos representativa será la media.

Si la distribución está tabulada, calcularemos la varianza utilizando la siguiente expresión:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2$$

siendo x_i los valores de la variable, si la distribución no está agrupada en intervalos o la marca de clase de cada intervalo, si la distribución está agrupada en intervalos.

La varianza no puede ser negativa, puesto que es una suma de cuadrados.

DESVIACIÓN TÍPICA

La varianza presenta el inconveniente de que aparece expresada en las unidades de la variable al cuadrado, lo que hace difícil su interpretación. Para resolver este problema se define la desviación típica.

La desviación típica, que se representa por s , es la raíz cuadrada positiva de la varianza.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2}$$

Ejemplo 13. A partir de la varianza calculada en el ejemplo 12, obtener la desviación típica.

La desviación típica es la raíz cuadrada positiva de la varianza. Por tanto, se tiene que:

$$s = \sqrt{1{,}76} = 1{,}33$$

MEDIDAS DE DISPERSIÓN RELATIVA

Las medidas de dispersión relativa son medidas adimensionales, es decir, no tienen unidades de medida o monetarias. Resultan útiles para comparar distintas distribuciones o datos. A continuación, se presentarán las principales medidas de dispersión relativa: el coeficiente de variación de Pearson y la variable tipificada.

COEFICIENTE DE VARIACIÓN DE PEARSON

El coeficiente de variación de Pearson se utiliza para comparar distribuciones en términos de dispersión. Supóngase que se quiere comparar la dispersión de dos distribuciones cuya media es distinta o que están expresadas en distintas unidades. En estos casos no se puede utilizar la desviación típica. Para ello se necesita construir medidas adimensionales, es decir, que no tengan unidades.

El coeficiente de variación, que se representa por CV, se define como el cociente entre la desviación típica y la media aritmética, supuesto que esta sea distinta de cero.

$$CV = \frac{s}{\bar{x}}$$

Es adimensional, es decir, no tiene unidades. Presenta el inconveniente de que no se puede utilizar cuando la media aritmética es nula. En ese caso, el valor de coeficiente no nos mediría la dispersión de la distribución.

Ejemplo 14. A partir de los datos del ejemplo 1, calcular el coeficiente de variación de Pearson.

$$CV = \frac{s}{\bar{x}} = \frac{1{,}326}{11{,}80} = 0{,}1124$$

VARIABLE TIPIFICADA

La variable tipificada se utiliza para comparar datos de distribuciones. Supóngase que se dispone de un conjunto de datos $x_1, x_2, \ldots, x_N$ cuya media y cuya desviación típica se representarán por $\bar{x}$ y s respectivamente. Los valores tipificados se obtienen realizando una transformación lineal en los datos originales que consiste en restar a cada dato la media y el resultado obtenido dividirlo entre la desviación típica.

Por tanto, los valores tipificados serán:

$$z_1 = \frac{x_1 - \bar{x}}{s},\ z_2 = \frac{x_2 - \bar{x}}{s}, \ldots,\ z_N = \frac{x_N - \bar{x}}{s}.$$

Obsérvese que:

1. El signo del valor tipificado depende del signo del numerador, puesto que la desviación típica es positiva. Se pueden presentar tres situaciones:

   - Si el dato está por encima de la media, el valor tipificado será positivo.
   - Si está por debajo de la media, el valor tipificado será negativo.
   - Si coincide con la media, el valor tipificado será nulo.

2. Los valores tipificados son adimensionales, es decir, no tienen unidades de medida.

6. VARIABLE ESTADÍSTICA BIDIMENSIONAL

Una variable estadística bidimensional es un vector de dimensión 2, que se representará por (X,Y), cuyas componentes son variables estadísticas. Por ejemplo, el peso y la estatura de una persona.

La **covarianza**, que se representa por S_{xy}, es una medida del sentido de la dependencia lineal existente entre dos variables.

La covarianza viene dada por la siguiente expresión:

$$S_{xy} = \sum_i \frac{(x_i - \bar{x})(y_j - \bar{y})}{N}$$

A partir de dicha expresión, se deduce otra que se empleará para simplificar el cálculo:

$$S_{xy} = \sum_i \frac{x_i y_j}{N} - \bar{x} \times \bar{y}$$

Interpretación del signo de la covarianza.

- Si la covarianza es positiva $S_{xy} > 0$, existe dependencia lineal directa. Las variables varían en el mismo sentido.
- Si la covarianza es negativa $S_{xy} < 0$, existe dependencia lineal inversa. Las variables varían en sentido contrario.

- Si la covarianza es igual a 0 $S_{xy} = 0$, no existe dependencia lineal entre las variables.

La covarianza puede ser positiva, negativa o igual a cero.

7. REGRESIÓN LINEAL MÍNIMO-CUADRÁTICA

Si tenemos una variable estadística bidimensional y queremos analizar la relación existente entre las variables, podemos realizar dicho análisis siguiendo dos enfoques diferentes:

- **Regresión**: consiste en analizar la forma de la dependencia existente entre las variables.
- **Correlación**: consiste en analizar el grado de dependencia existente entre las variables.

Para llevar a cabo un análisis de regresión tenemos que:

1. Elegir la función a ajustar (lineal, parabólica, potencial etc,).
2. Determinar los valores de los parámetros de la función. El criterio que vamos a seguir para determinar dichos parámetros es el de mínimos cuadrados. Consiste en determinar los valores que hagan mínima la suma de los cuadrados de las diferencias entre los valores observados y los valores teóricos determinados por la función.

RECTA DE REGRESIÓN DE Y SOBRE X

Se decide ajustar una recta de la forma $y = a + bx$.

Obsérvese que la X es la variable independiente y la Y es la variable dependiente.

Sean $y_1, y_2, ..., y_N$ los valores observados de la variable Y, y sean $x_1, x_2, ..., x_N$ los valores observados de la variable X.

Los valores teóricos determinados por la recta se representarán por $\hat{y}_1, \hat{y}_2, ..., \hat{y}_N$. Por tanto, $\hat{y}_1 = a + bx_1$, $\hat{y}_2 = a + bx_2, ...,$ $\hat{y}_n = a + bx_N$.

Se denomina residuo a la diferencia entre el valor observado y el valor teórico determinado por la función.

Los residuos se representarán por $e_1, e_2, ..., e_n$ siendo: $e_1 = y_1 - \hat{y}_1$, $e_2 = y_2 - \hat{y}_2, ..., e_N = y_N - \hat{y}_N$.

Para determinar los valores de los parámetros a y b, minimizaremos la suma de los cuadrados de los residuos.

Se trata, por tanto, de resolver el siguiente problema:

$$\underset{a,b}{Min} \quad S = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (a + bx_i)]^2$$

Para encontrar el mínimo se igualan las derivadas parciales a 0.

$$\frac{\partial S}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = -2\sum_i x_i (y_i - a - bx_i) = 0$$

Se genera el siguiente sistema de ecuaciones:

$$\sum_i y_i = aN + b\sum_i x_i$$

$$\sum_i x_i y_i = a\sum_i x_i + b\sum_i x_i^2$$

Resolviendo el sistema se obtienen los valores de a y b.

$$a = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

Por tanto, la recta de regresión de Y sobre X es:

$$y = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x$$

RECTA DE REGRESIÓN DE X SOBRE Y

Se decide ajustar una recta de la forma $x = a' + b'y$.

Obsérvese que la Y es la variable independiente y la X es la variable dependiente.

Procediendo de forma análoga a la que acabamos de exponer, se deduce la recta de regresión de X sobre Y.

$$x = \bar{x} - \frac{s_{xy}}{s_y^2}\bar{y} + \frac{s_{xy}}{s_y^2}y$$

Ejemplo 15. A partir de una variable bidimensional se han obtenido las siguientes medidas:

$$\bar{x} = 3 \quad \bar{y} = 7 \quad s_x^2 = 2 \quad s_y^2 = 8 \quad S_{xy} = 4$$

Determinar la recta de regresión de Y sobre X.

La recta de regresión de Y sobre X viene dada por la siguiente expresión:

$$y = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x$$

Sustituyendo se tiene que:

$$a = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} = 7 - \frac{4}{2} \times 3 = 1$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{4}{2} = 2$$

Por tanto, la recta de Y sobre X es:

$$y = 1 + 2x$$

8. CORRELACIÓN LINEAL

La correlación es el grado de dependencia existente entre las variables. Para medir el grado de correlación se emplean una serie de coeficientes. A continuación, se presentan dos coeficientes que miden el grado de dependencia lineal: el coeficiente de correlación lineal y el coeficiente de determinación lineal

COEFICIENTE DE CORRELACIÓN LINEAL

El coeficiente de correlación lineal, que se representa por R, mide el grado y el sentido de la dependencia lineal existente entre las variables.

$$R = \frac{S_{xy}}{S_x S_y}$$

El coeficiente de correlación lineal:

1. Es adimensional, es decir, no tiene unidades de medida.
2. Toma valores en el intervalo $[-1, 1]$.

Interpretación

- Si $R = 1$ las variables varían en el mismo sentido y en la misma proporción. Se dice que la correlación es directa y perfecta.
- Si $R = -1$ las variables varían en sentido contrario y en la misma proporción. Se dice que la correlación es inversa y perfecta.

- Si $R=0$ no existe dependencia lineal entre las variables. La correlación es nula.
- Si $0<R<1$ las variables varían en el mismo sentido. La correlación es directa.
- Si $-1<R<0$ las variables varían en sentido contrario. La correlación es inversa.

COEFICIENTE DE DETERMINACIÓN LINEAL

El coeficiente de determinación lineal, que se representa por R^2, es el cuadrado del coeficiente de correlación lineal. Mide el grado de dependencia lineal, pero no el sentido de la dependencia.

$$R^2=\frac{S_{xy}^2}{S_x^2\,S_y^2}$$

El coeficiente de determinación lineal:

1. Es adimensional, es decir, no tiene unidades de medida.
2. Toma valores en el intervalo $[0,1]$. Cuanto más próximo esté a 1 más fuerte será la dependencia.

Ejemplo 16. A partir de una variable bidimensional se han obtenido las siguientes medidas:

$$\bar{x}=3 \quad \bar{y}=7 \quad s_x^2=2 \quad s_y^2=8 \quad S_{xy}=4$$

Determinar el coeficiente de correlación lineal e interpretar el resultado obtenido.

$$R=\frac{S_{xy}}{S_x\,S_y}=\frac{4}{\sqrt{2}\times\sqrt{8}}=1$$

La correlación es directa y perfecta. Las variables varían en el mismo sentido y en la misma proporción.

CUADROS

CUADRO 1. TIPOS DE FRECUENCIAS

Frecuencia absoluta	es el número de veces que el valor se repite. Se representa por n_i.
Frecuencia relativa	es el cociente entre la frecuencia absoluta y el número total de datos N. Se representa por f_i.
Frecuencia absoluta acumulada	es el número de datos menores o iguales que el valor considerado. Se representa por N_i.
Frecuencia relativa acumulada	es el cociente entre la frecuencia absoluta acumulada y el número total de datos. Se representa por F_i.

CUADRO 2. MEDIDAS QUE CARACTERIZAN A LA VARIABLE

Media aritmética	La media aritmética, que se representará por $\bar{x}$, es la suma de todos los valores de la variable dividida entre el número total de datos.
Mediana	Es valor de la variable que deja a su derecha y a su izquierda el mismo número de observaciones, supuesto que los datos están ordenados de menor a mayor. Se representará por M_e.
Moda	Es el valor de la variable que más se repite, es decir, el que tiene la mayor frecuencia absoluta asociada. Se representará por M_0.
Varianza	La varianza es la suma de los cuadrados de las desviaciones de los valores de la variable respecto a la media dividida entre el número total de datos.
Desviación típica	La desviación típica es la raíz cuadrada positiva de la varianza.
Coeficiente de Variación	El coeficiente de variación de Pearson se define como el cociente entre la desviación típica y la media aritmética, supuesto que esta sea distinta de cero.

CUADRO 3. REGRESION Y CORRELACIÓN

REGRESION LINEAL	**Recta de Regresión de Y sobre X**	$y = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x$
	Recta de Regresión de X sobre Y	$x = \bar{x} - \frac{s_{xy}}{s_y^2}\bar{y} + \frac{s_{xy}}{s_y^2}y$
CORRELACIÓN LINEAL	**Coeficiente de correlación lineal**	$R = \frac{S_{xy}}{S_x S_y}$
	Coeficiente de determinación lineal	$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$

CONCLUSIONES

Cuando se está analizando una variable y se dispone de un conjunto de datos de la misma, es fundamental realizar un análisis riguroso de los datos. En este análisis se distinguen tres etapas.

En primer lugar, es necesario conocer en profundidad la variable que se está analizando y la forma en la que se han generado cada uno de los datos de los que se dispone. Si los datos no se han generado de la misma forma, será necesario hacerlos homogéneos para poder analizarlos de forma conjunta.

A continuación, se pasa a resumir la información mediante tablas y medidas que caracterizan a la variable.

Finalmente, se describen los datos mediante la interpretación de las tablas y las medidas calculadas.

RECAPITULACIÓN

Una distribución de frecuencias esta formada por los valores que toma la variable acompañados de sus correspondientes frecuencias. Se distinguen dos tipos de distribuciones de frecuencias:

- **Distribuciones no agrupadas**: cada valor de la variable aparece con su frecuencia asociada.
- **Distribuciones agrupadas intervalos**: cuando el número de valores distintos que ha tomado la variable es muy grande se agrupan número de valores distintos que ha tomado la variable es muy grande se agrupan en intervalos. Esto ocasiona perdida de información pero se gana en simplicidad y operatividad.

La información que proporcionan los datos se puede resumir mediante una serie de medidas. Se distinguen:

- **Medidas de centralización**. Las principales son la media aritmética, la mediana y la moda.
- **Medidas de dispersión**. Las principales son la varianza, la desviación típica y el coeficiente de variación de Pearson.

Cuando se dispone de datos de dos variables, y se quiere analizar la relación que existe entre ellas, se pueden seguir dos enfoques distintos:

- **Regresión**: consiste en analizar la forma de la dependencia existente entre las variables.
- **Correlación**: consiste en analizar el grado de dependencia existente entre las variables.

AUTOCOMPROBACIÓN

1. La frecuencia absoluta de un valor es:

   a) El número de veces que el valor se repite.

   b) El cociente entre el número de veces que el valor se repite y el número total de datos.

   c) El número de datos menores o iguales que el valor considerado.

   d) El porcentaje de datos menores o iguales que el valor considerado.

2. La frecuencia relativa de un valor es:

   a) El número de veces que el valor se repite.

   b) El cociente entre el número de veces que el valor se repite y el número total de datos.

   c) El número de datos menores o iguales que el valor considerado.

   d) El porcentaje de datos menores o iguales que el valor considerado.

3. La frecuencia absoluta acumulada de un valor es:

   a) El número de veces que el valor se repite.

   b) El cociente entre el número de veces que el valor se repite y el número total de datos.

   c) El número de datos menores o iguales que el valor considerado.

   d) El porcentaje de datos menores o iguales que el valor considerado.

4. La frecuencia relativa acumulada de un valor es:

   a) El número de veces que el valor se repite.

   b) El cociente entre el número de veces que el valor se repite y el número total de datos.

   c) El número de datos menores o iguales que el valor considerado.

   d) El porcentaje de datos menores o iguales que el valor considerado.

5. La marca de clase de un intervalo es:

   a) El punto medio del intervalo.

   b) El limite superior del intervalo.

   c) El limite inferior del intervalo.

   d) La diferencia entre el limite superior y el limite inferior del intervalo.

6. La mediana es:

   a) El valor que deja a su derecha y a su izquierda el mismo número de observaciones, supuesto que los datos están ordenados de mayor a menor.

   b) El valor que deja a su derecha y a su izquierda el mismo número de observaciones, supuesto que los datos están ordenados de menor a mayor.

   c) El valor que más se repite.

   d) Ninguna respuesta es correcta.

7. La moda es:

   a) El valor que deja a su derecha y a su izquierda el mismo número de observaciones, supuesto que los datos están ordenados de menor a mayor.

   b) La frecuencia absoluta del valor que más se repite.

   c) El valor que más se repite.

   d) Ninguna respuesta es correcta.

8. La varianza:

   a) Nunca puede ser nula.

   b) Puede ser positiva o nula.

   c) Tomará valores negativos sólo si la variable estadística toma valores negativos.

   d) Puede ser positiva, negativa o nula.

9. Si la desviación típica es igual a 9:

   a) La varianza es igual a 3.

   b) La varianza es igual a 81.

   c) No es posible calcular la varianza.

   d) Ninguna respuesta es correcta.

10. Si la media de una variable es igual a 2 y la desviación típica es igual a 4, el coeficiente de variación de Pearson es:

    a) 2.

    b) 4.

    c) 0,5.

    d) Ninguna respuesta es correcta.

SOLUCIONARIO

1.	a	2.	b	3.	c	4.	d	5.	a
6.	b	7.	c	8.	b	9.	b	10.	a

PROPUESTAS DE AMPLIACIÓN

Si estás interesado en ampliar los conocimientos sobre Estadística Descriptiva, puedes consultar la página web del Instituto Nacional de Estadística www.ine.es.

BIBLIOGRAFÍA

Tomeo Perucha V. y Uña Juárez I. (2003). Lecciones de Estadística Descriptiva. Curso teórico - práctico. Editorial Thomson.

Peralta Astudillo, M.J., Rúa Vieyes, A., Redondo Palomo, R., y del Campo Campos, C. (2007). Estadística. Problemas resueltos. Editorial Pirámide.