Documentación de apoyo a la preparación de la oposición al Cuerpo General Administrativo de la Administración del Estado, especialidad Estadística

Módulo 2: Estadística Básica

ÍNDICE

Unidad 4. Estadística. Conceptos básicos. Variables estadísticas. Distribuciones de frecuencias. Tipos de frecuencias. Tablas estadísticas. Representación gráfica. 4

4.1 Estadística.....4

4.2 Conceptos básicos 5

4.3 Variables estadísticas.....6

4.4 Distribuciones de frecuencias.....7

4.4.1 Tipos de frecuencias 8

4.4.2 Tablas estadísticas.....9

4.5 Representación gráfica.....13

Unidad 5. Medidas descriptivas de una variable estadística. Medidas de posición y de dispersión. Medidas de forma y concentración. 19

5.1 Medidas descriptivas de una variable estadística 19

5.2 Medidas de posición.....19

5.3 Medidas de dispersión.....26

5.4 Medidas de forma.....28

5.5 Medidas de concentración.....30

Unidad 6. Distribuciones bidimensionales. Tablas de correlación y contingencia. Distribuciones marginales y condicionadas. Independencia estadística. 33

6.1 Distribuciones bidimensionales.....33

6.2 Tablas de correlación y contingencia.....34

18.2.1 Tablas de correlación 34

6.2.2 Tablas de contingencia.....36

6.3 Distribuciones marginales y condicionadas 36

18.3.1 Distribuciones marginales.....36

6.3.2 Distribuciones condicionadas 37

6.4 Independencia estadística 38

Unidad 7. Medidas descriptivas de variables bidimensionales: Covarianza y coeficiente de correlación lineal. La recta de regresión. 40

7.1 Medidas descriptivas de variables bidimensionales 40

7.1.1 Covarianza 40

19.1.2 Coeficiente de correlación lineal 41

7.2 Recta de regresión 42

7.2.1 Ajuste por mínimos cuadrados 43

Unidad 8. Tasas de variación. Números índices simples. Índices complejos. Índices de precios y de cantidad. 44

8.1 Tasas de variación 44

8.2 Números índices 48

8.2.1 Números índices simples 48

8.2.2 Números índices complejos 49

8.3 Índices de precios y de cantidad 50

Anexo ejercicios 53

Unidad 4. Estadística. Conceptos básicos. Variables estadísticas. Distribuciones de frecuencias. Tipos de frecuencias. Tablas estadísticas. Representación gráfica.

4.1 Estadística

La disponibilidad de información se hace necesaria para las empresas, las administraciones públicas y los individuos de una población a la hora de tomar decisiones. El proceso para la obtención de información abarca desde la recogida de los datos de una población o parte de ella, hasta la organización y procesamiento de los mismos para convertirlos en información útil para el usuario.

La **Estadística** es la ciencia que se ocupa de desarrollar y estudiar métodos para recoger, analizar, describir e interpretar los datos empíricos. Asimismo, la Estadística, resulta ser una herramienta multidisciplinar, ya que encuentra aplicación en prácticamente todos los campos científicos.

La ciencia Estadística se divide en dos ramas principalmente:

- **La Estadística Descriptiva** trata del recuento, ordenación, clasificación y representación de una serie de características del conjunto de datos involucrados en el estudio. El objetivo principal es proporcionar un resumen de los datos obtenidos con objeto de caracterizarlos y clasificarlos, pero no nos permite extraer conclusiones más allá de los datos que hemos analizado, ni llegar a soluciones sobre ninguna hipótesis que pudiéramos haber formulado. Son simplemente un conjunto de herramientas para describir nuestros datos.
- **Inferencia estadística** comprende las distintas técnicas que permiten a partir de un subconjunto de la población a estudio, hacer generalizaciones sobre las características de la población total con un riesgo de error, el cual podemos medir en términos de probabilidad. Los métodos de la inferencia estadística se pueden dividir en dos grupos principalmente: la estimación de parámetros y los contrastes de hipótesis.

Vinculado a la Estadística se encuentra el Cálculo de Probabilidades

- **Cálculo de Probabilidades:** tiene como objetivo central el estudio de métodos de análisis del comportamiento de aquellos fenómenos que aunque se realizan en

condiciones similares su resultado se atribuye al azar y no es posible predecir cuál va a ser el resultado con exactitud . Se conocen con el nombre de fenómenos aleatorios y su estudio se incluye en el quinto módulo.

Las Estadística Descriptiva resulta de gran utilidad puesto que es difícil concebir lo que muestran los datos, especialmente si estos son cuantiosos. Por ejemplo, si obtuviéramos los resultados de los ingresos de 150.000 hogares del territorio nacional, la lectura de dichos datos uno a uno no proporcionaría información de interés. Por otro lado, si se resume la información exhaustiva de cada hogar puede resultar útil, el ingreso medio por hogar o el salario medio por persona, harían más fácil entender la gran cantidad de datos a la que nos enfrentamos. En las unidades que integran este módulo vamos a introducir conceptos y procedimientos de esta rama de la estadística.

4.2 Conceptos básicos

Se define **población** como el conjunto de unidades o elementos, pueden ser individuos u objetos, sobre los que se pretende observar o medir una o varias características.

A cada uno de los componentes de la población se le denomina elemento o **unidad estadística** de la población. En un estudio estadístico, la información se obtiene de los elementos, por lo que también se denominan unidades de análisis.

Por distintas razones, tales como el coste de la recogida, el tamaño excesivamente grande de la población a estudio o la destrucción de las unidades en el proceso de observación, será imposible investigar en algunas ocasiones, a todos y cada uno de los elementos que conforman la población. Por ello, tendremos que seleccionar una parte de dicha población, que denominaremos **muestra** y realizar nuestro estudio mediante esta porción de la población. Parece lógico pedir que la muestra que se seleccione sea un buen espejo de la población a estudio, es decir, pediremos que la muestra sea representativa de la población.

4.3 Variables estadísticas

Dada una población, habitualmente lo que nos interesa de ésta son algunas de sus características. Por ejemplo la nacionalidad, el color de pelo, la edad, el sexo, la estatura o la profesión, son características que se pueden observar de los residentes en el territorio nacional.

Definimos como **variable estadística** unidimensional, X, a una característica de la población que estamos estudiando y que posee varios grados, modalidades o valores posibles. Para cada elemento de la población, dicha variable tomará uno y sólo uno de los valores posibles o de las modalidades existentes. A dicho valor concreto, que denotaremos por x, lo denominamos dato.

Las variables estadísticas pueden ser clasificadas en dos grupos: variables cualitativas o variables cuantitativas.

a) **Variable cualitativa:** es cualquier cualidad o atributo de los elementos de una población que no puede expresarse numéricamente. Las variables cualitativas permiten clasificar los elementos de la población determinando si el elemento pertenece o no a una modalidad, y dichas modalidades no son medibles, aunque estas modalidades sean identificadas con un código numérico para su tratamiento.

   Ejemplo: La provincia de residencia de una persona es una variable cualitativa aunque podría identificarse cada provincia con su código correspondiente: 8 Barcelona, 15 A Coruña, 28 Madrid.

   En el caso que el atributo objeto de estudio, presente dos modalidades se le denomina dicotómico y cuando presenta más de dos modalidades, múltiple.

   Ejemplo:

   - Variable cualitativa dicotómica: el sexo de una persona presenta dos modalidades, hombre o mujer.
   - Variable cualitativa o atributo de modalidad múltiple: Color favorito de una persona, (rojo, azul, amarillo, verde...), nacionalidad de una persona (española, italiana, portuguesa,...).

A su vez pueden clasificarse atendiendo a las propiedades de su escala de medida como:

- **Escala Nominal:** cuando en las distintas categorías o modalidades de la variable no resulta posible establecer una ordenación, las variables se denominan nominales. Generalmente, para su tratamiento, se suele asignar un número a

cada posible valor, proceso que recibe el nombre de codificación. Como ejemplos podemos mencionar el estado civil o la religión de un individuo, como casos de variables cualitativas nominales.

- **Escala Ordinal**: resulta cuando entre las modalidades de la variable cualitativa puede establecerse algún orden, a diferencia de lo que ocurre con las variables nominales. Por ejemplo: el nivel de estudios de un individuo, el grado de satisfacción de un producto o servicio recibido por un cliente en un supermercado,…

b) **Variable cuantitativa**: es cualquier característica de los elementos a la que se le puede asignar un número. Es decir, son medibles. Las variables cuantitativas se clasifican en continuas o discretas:

- Discretas: cuando toman un número finito de valores o un número infinito numerable de valores. Por ejemplo, el número de alumnos en una clase o número de hijos (será 0, 1, 2,3,...pero nunca 1,86).
- Continuas: son aquellas variables que pueden tomar infinitos valores dentro de un cierto intervalo de la recta real. Por ejemplo, el peso de los alumnos de una clase, la altura de los jugadores del equipo nacional de baloncesto,…

Definimos como **dominio** de una variable al conjunto de los posibles valores o modalidades diferentes que puede tomar la variable.

Notación: En general, dada una variable, X ,a los k posibles valores o modalidades que puede tomar los representamos por $x_1, x_2, \ldots, x_k$, siendo este conjunto el dominio de la variable. En el caso de la población "Provincias de España", el dominio sería la enumeración de cada una de las 52 provincias españolas.

4.4 Distribuciones de frecuencias

Cuando se toma la información de una población, ya sea en su totalidad o de una muestra de la misma, se obtiene en la mayoría de ocasiones un gran volumen de datos que resulta de difícil manejo. Por ello, se hace preciso disponer de una batería de técnicas que permitan la ordenación, presentación y reducción de los datos y así facilitar el análisis de los resultados obtenidos.

Denominamos tabulación al procedimiento que implica realizar una ordenación y agrupación de los datos o valores comunes de una variable. Por otro lado, hablamos

de **distribución de frecuencias** para aludir a las tablas con la disposición final obtenida para los datos.

Tanto con la tabulación de los datos, así como con la representación gráfica de los mismos, que veremos en esta misma unidad, se pretende que el usuario pueda hacerse una idea de las características principales de la distribución de esos datos sin necesidad de recurrir a la observación de la enumeración completa de los mismos.

Supongamos que tenemos una población o muestra de una población, compuesta por N elementos. Sea X la variable de estudio y supongamos que X puede tomar un número k de valores distintos. Representamos dichos valores por $x_1, x_2, \dots x_k$. De manera general, se representará cada valor por x_i, donde i tomará valores entre 1 y k.

4.4.1 Tipos de frecuencias

- **Frecuencia absoluta** de un valor x_i ,es el número de veces que aparece cada valor o modalidad de la variable en la muestra tomada .Se suele representar por n_i.
    - Evidentemente, n_i es un número entero mayor o igual que cero.
    - La suma total de todas las frecuencias absolutas de los k posibles valores de la variable es el tamaño de la muestra de elementos observados. Se denomina frecuencia total y se representa por N.

$$\sum_{i=1}^{k} n_i = n_1+n_2+n_3+\dots\dots+n_k = N$$

- **Frecuencia relativa** de un valor x_i, se obtiene al dividir la frecuencia absoluta entre el número total de elementos observados en la población (frecuencia total). Se representa por

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

Se verifica que:

a) El valor de la frecuencia relativa estará comprendido entre 0 y 1.

b) La suma de las frecuencias relativas de todos los valores de la variable estadística es siempre la unidad. Veamos a continuación que efectivamente esto es cierto:

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{n_1}{N} + \frac{n_2}{N} + \ldots\ldots + \frac{n_k}{N} = \frac{(n_1 + n_2 + \cdots + n_k)}{N} = \frac{N}{N} = 1$$

- **Frecuencia absoluta acumulada** de un valor x_i , es la suma de las frecuencias absolutas de todos los valores que son menores o iguales que él. Se denota por N_i. Suponiendo que los valores de X están ordenados crecientemente, tal que $x_1 < x_2 < \cdots < x_k$.La frecuencia absoluta acumulada de x_i será:

$$N_i = n_1 + n_2 + \ldots\ldots + n_{i-1} + n_i = \sum_{j=1}^{i} n_j$$

  La frecuencia absoluta acumulada del último valor de la variable tiene que coincidir con la frecuencia total N, por ser la suma de todas las frecuencias absolutas.

  En el caso de variables cualitativas de escala nominal no tiene sentido calcularla puesto que las modalidades de nuestra variable no son susceptibles de ordenación.

- **Frecuencia relativa acumulada** de un valor x_i, es igual a la suma de las frecuencias relativas de todos los valores menores o iguales que x_i. Se representa por F_i .

  Es decir, suponiendo que los valores de X están ordenados crecientemente, tal que $x_1 < x_2 < \cdots < x_k$, la frecuencia absoluta acumulada de xi será:

  - $F_i = \sum_{j=1}^{i} f_j = f_1 + f_2 + \ldots + f_{i-1} + f_i$
  - También se puede obtener como $F_i = {}^{N_i}/_{N}$, i=1,..., k

4.4.2 Tablas estadísticas

El proceso que hemos visto hasta ahora, una vez son obtenidos los datos tras la observación, comprende la ordenación de los mismos y la agrupación de aquellos valores que se repiten (frecuencias).

Las distribuciones de frecuencias se presentan mediante **tablas estadísticas**. A grandes rasgos el formato de las tablas sería el siguiente: en una columna se disponen los valores o modalidades de la variable en estudio y en otra/s la/s frecuencia/s correspondiente/ a ese determinado valor. Si se observa un solo carácter de la población, sus datos se presentarán mediante una tabla denominada distribución de frecuencias unidimensional.

Es conveniente recordar que a la hora de tabular, es necesario tener en cuenta el número de elementos observados así como el número de valores o modalidades distintos que toma la variable.

Ejemplo: Se conoce que la distribución de alumnos de enseñanza secundaria de un centro de la Comunidad de Madrid por sexo, es de 295 mujeres y 215 hombres. El tamaño de la población sería 110 alumnos.

Sexo	**Número de alumnos**	**Frecuencias relativas**
Mujeres	295	295/510
Hombres	215	215/510

Vamos a realizar la exposición de la construcción de tablas, tratando por separado los casos en los que los datos están agrupados, de aquellos en los que los datos no se encuentran agrupados por existir algunas diferencias.

- **Datos no agrupados en intervalos**

  Supongamos que el número de posibles valores o modalidades distintas que toma la variable no es muy grande, a pesar de que se disponga de un gran volumen de observaciones. En este caso, se presentan los datos en una distribución de frecuencias, de tal forma que en la tabla se reflejan en la primera columna los pocos valores o modalidades del carácter o variable y, en las siguientes columnas, cada uno de los diferentes tipos de frecuencias de cada dato.

  El primer paso para realizar la tabulación de los datos es contar los elementos con igual valor o modalidad (frecuencia absoluta). Una vez obtenida la frecuencia absoluta es fácil determinar el resto de frecuencias.

  La tabla tendría las siguientes cabeceras

Valores de la variable ordenados de menor a mayor(x_i)	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa f_i (o en %)	Frecuencia relativa acumulada F_i (o en %)

- **Datos agrupados en intervalos**

Si el número de posibles valores distintos o modalidades de la variable es elevado, la tabla construida como se ha expuesto anteriormente, daría lugar a unas columnas larguísimas, que carecerían de utilidad. Para ello se agrupan valores de la variable que sean próximos en intervalos o clases, con la idea de simplificar el tamaño de la tabla, sin perder demasiada información, obteniendo **intervalos o clases** homogéneos. Para la agrupación no existe criterio estándar.

- **Clases:** Cada una de las agrupaciones que surgen al unir los valores consecutivos o próximos de una variable (cuantitativa o cualitativa), teniendo en cuenta que las clases deben ser excluyentes y exhaustivas.

  La frecuencia de la clase se obtendrá como suma de las frecuencias de los valores que se han agrupado en una misma clase. No existe criterio estándar para determinar el número de clases pero lo ideal es trabajar con distribuciones de entre 5 y 15 clases.

- **Límites de clase**: corresponden a los valores inferior y superior de cada clase. Denotaremos con L_{i-1} al límite inferior de la i-ésima clase y L_i al valor mayor de la i-ésima clase.

- **Intervalo**: manera de identificar una clase por los valores llamados límites de clase o límites del intervalo. Los intervalos normalmente acaban en un número (límite superior) que es el mismo con el que empieza el intervalo siguiente (límite inferior), y suelen ser de la forma $(L_{i-1}, L_i]$ es decir abiertos inferiormente y cerrados superiormente. Para establecer el número aproximado de clases no existe un criterio estándar, algunos autores como Sturges proponen tomar como número de intervalos k: k= 1 + (3,3 log N)

- **Amplitud del intervalo**: se define como la diferencia entre el límite superior y el inferior de una clase. Se suele representar por $a_i = L_i - L_{i-1}$.

  La amplitud entre las distintas clases de un mismo conjunto de datos, puede ser constante o no, en función de las necesidades o requerimientos del estudio, aunque es conveniente que todas las clases tengan la misma amplitud. Existe una regla aproximada para determinar la anchura de cada intervalo:

$$anchura = (valor\ máximo - valor\ minimo)/nº\ de\ clases$$

- **Marca de clase:** es el valor concreto representativo de cada clase, necesario para el tratamiento numérico. Es habitual tomar el valor central del intervalo

como marca de clase, para su cálculo se realiza la media de los dos límites del intervalo y se denota por $c_i = (L_i + L_{i-1})/2$.

Ya hemos comentado con anterioridad que la agrupación de valores lleva consigo asociado un coste, en lo que a pérdida de información se refiere. Al tomar en nuestro estudio como valor representativo la marca de clase, la información finalmente obtenida será distinta de lo que se obtendría si no se realizase la agrupación de los datos.

La tabla obtenida tendría las siguientes columnas:

Intervalos $(L_{i-1}, L_i]$	Marcas de clase c_i	Frecuencia absoluta n_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa f_i	Frecuencia relativa acumulada F_i

Ejercicio 16.1 La edad de los habitantes de un municipio de la provincia de Soria, viene representada en la siguiente tabla:

Años	0-9	10-19	20-29	30-35	36-49	50-59	60-69	70-100
Personas	404	490	484	330	994	778	529	801

Se pide:

- En una tabla, presentar la población en cuatro grupos: menores de 20, de 20 a 35, de 36 a 59 y mayores de 59.
- Calcular el valor porcentual de cada grupo en forma simple y acumulada.

Nota: No existen individuos en el municipio con edad mayor a 100 años.

Solución 16.1: Nos dan la población clasificada según unos intervalos que no se corresponden a los solicitados en el enunciado, y piden presentar la población en una tabla con cuatro clases. Por ello, los intervalos de clase serían:

- Menores de 20 años: (0,19]: la última edad que se cuenta es la de 19 años puesto que es un intervalo cerrado por la derecha.
- De 20 a 35: (19, 35] (la primera edad que se cuenta es 20, porque el intervalo es abierto por la izquierda y la última 35, porque está cerrado por la derecha)
- De 36 a 59: (35, 59] (la primera edad que se cuenta es 36, porque el intervalo es abierto por la izquierda y la última 59 por la derecha)

- Mayores de 59: (59, 100] (la primera edad que se cuenta es 60, porque el intervalo es abierto por la izquierda y la última 100 por la derecha) –

Calculamos las marcas de clase para cada uno de los intervalos construidos:

- (0 +19)/2=9,5
- (19+35)/2= 27
- (35+59)/2=47
- (59+100)/2=79,5

Calculamos los diferentes tipos de frecuencia y lo presentamos en la siguiente tabla:

Intervalos $(L_{i-1}, L_i]$	Marcas de clase c_i	Frecuencia absoluta n_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa f_i	Frecuencia relativa acumulada F_i
(0,19]	9,5	894	894	0,18	0,18
(19,35]	27	814	1.708	0,17	0,35
(35, 59]	47	1772	3.480	0,37	0,72
(59,100]	79,5	1330	4810	0,28	1

A la vista de los resultados obtenidos podemos decir, por ejemplo, que el 72% de la población objeto de estudio tiene menos de 60 años o que algo más de la cuarta parte de la población es mayor de 59 años.

4.5 Representación gráfica

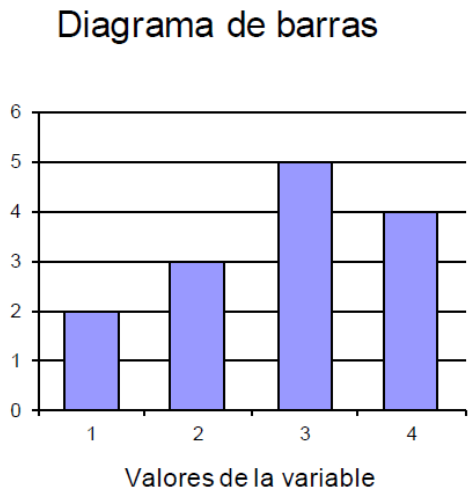
Acabamos de ver que la tabla de frecuencias de un conjunto de datos proporciona una forma organizada de los mismos, que permite que se analicen los resultados obtenidos de manera más sencilla. Ahora bien, la representación gráfica de los datos contenidos en una tabla estadística sería un recurso adicional del que dispondríamos y que puede ofrecer una visión más clara y rápida que la observación de los datos numéricos, así como ayudar detectar pautas en los datos.

No olvidemos que la representación gráfica en cualquier ámbito resulta ser un recurso potente puesto que produce mayor impacto en los usuarios que la tabulación de resultados.

1. Gráficos para representar variables cuantitativas

Presentamos a continuación los gráficos más comúnmente utilizados para representar este tipo de distribuciones. Observemos que para la obtención de los mismos se pueden utilizar las frecuencias relativas o absolutas y las acumuladas.

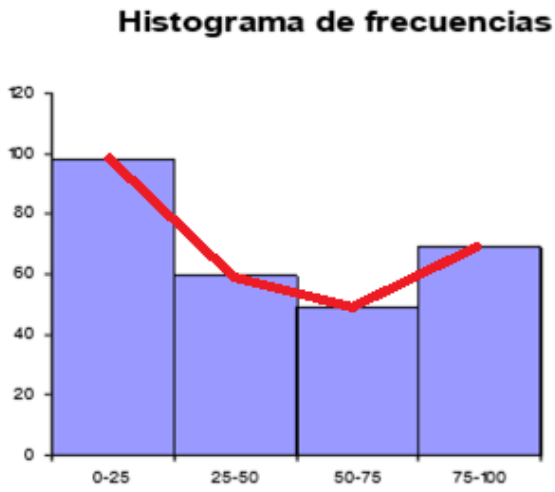
- **Diagrama de barras:** Se usa para variables cuantitativas discretas. Se considera un sistema de ejes cartesianos. El eje horizontal (abscisas) indica los valores de la variable por $x_1, x_2, \ldots x_k$ y el eje vertical (ordenadas) indica las frecuencias. Con la longitud de la barra vertical se representa su frecuencia absoluta o relativa.



- **Histograma:** resulta ser uno de los gráficos más usados para variables cuantitativas continuas. Sobre el eje de abscisas se representan los intervalos en los que han sido agrupados los datos y sobre cada uno de los intervalos se construye un rectángulo cuya base es la anchura del intervalo y de tal forma que las áreas de las barras son proporcionales a las frecuencias correspondientes.

  Los histogramas ayudan a conocer dónde se concentran los datos, cuáles son los extremos y si existen valores atípicos; en definitiva, mostrar la forma de la distribución de los datos.

  Es posible obtener lo que se conoce con el nombre de **polígono de frecuencias** si se unen los puntos medios de las bases superiores de los rectángulos. El polígono de frecuencias tiene como objetivo dar una visión aproximada de la curva que define la distribución de los datos.



2. Gráficos para variables cualitativas

Recordamos que los atributos no se expresan numéricamente sino por sus modalidades. Para su representación gráfica se puede utilizar el mismo método que para las variables cuantitativas discretas de pocos datos diferentes situando en abscisas las modalidades y en ordenadas las frecuencias.

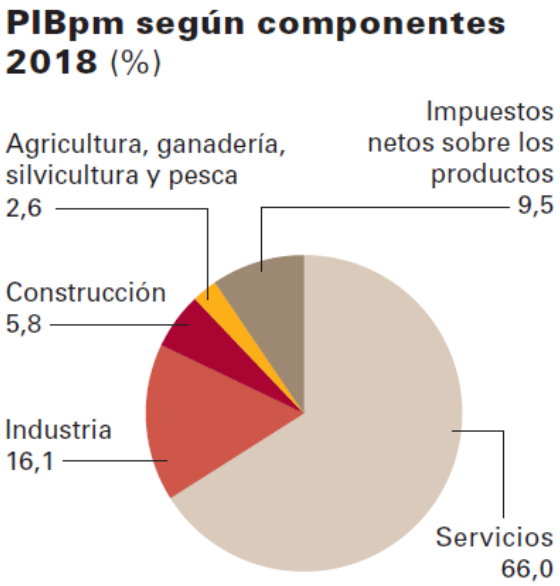
- **Diagrama de rectángulos o barras**: nuestra variable a estudio tiene distintas modalidades, para cada una de ellas se representa una barra vertical de amplitud constante y cuya altura corresponderá con su frecuencia absoluta o relativa. Cada una de las barras no se representan unidas para no dar la sensación de continuidad en las distintos valores que toma nuestra variable cualitativa.

  Ejemplo: En una clase de Análisis Matemático de la Facultad de Matemáticas se han matriculado 102 hombres y 105 mujeres. Representar en un diagrama de rectángulos la población clasificada según la variable sexo.

3. Otras representaciones gráficas

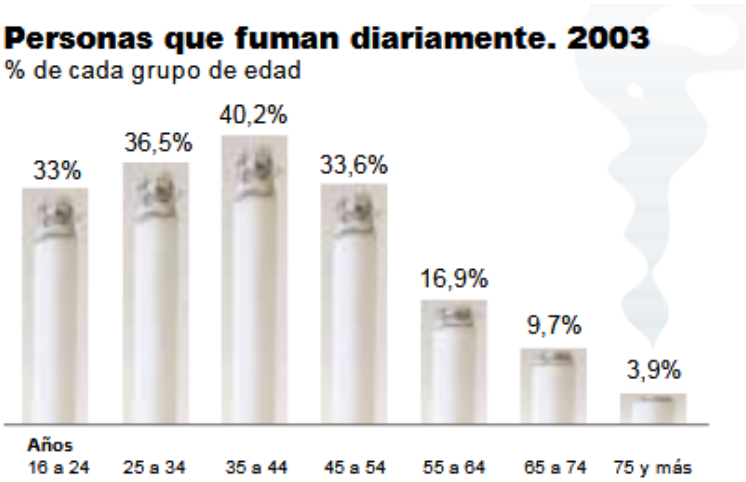
i. **Diagrama de sectores:** sirve para representar variables tanto cuantitativas como cualitativas y resulta ser uno de los gráficos más utilizados. Cada sector o porción del círculo representa un valor de la variable. La amplitud de las porciones o sectores tiene que ser proporcional al valor de la frecuencia correspondiente. En los casos en los que la variable a estudio tome muchos valores o tenga muchas modalidades distintas, no resulta óptimo su uso.

Ejemplo: Imagen obtenida de la publicación "España en cifras 2019" de INE



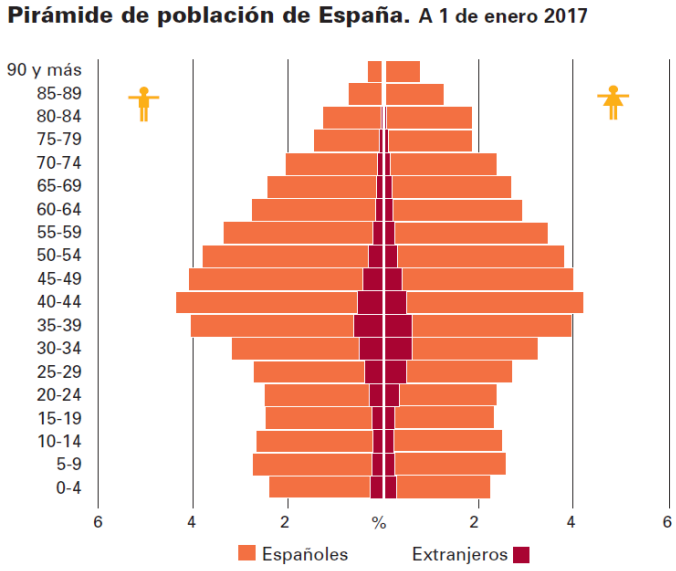
ii. **Pictogramas:** se basan en la repetición de una figura cuyo tamaño es proporcional a las frecuencias absolutas.

Ejemplo: Imagen obtenida del Boletín Cifras INE: La salud de los españoles 2/2005 Fuente INE



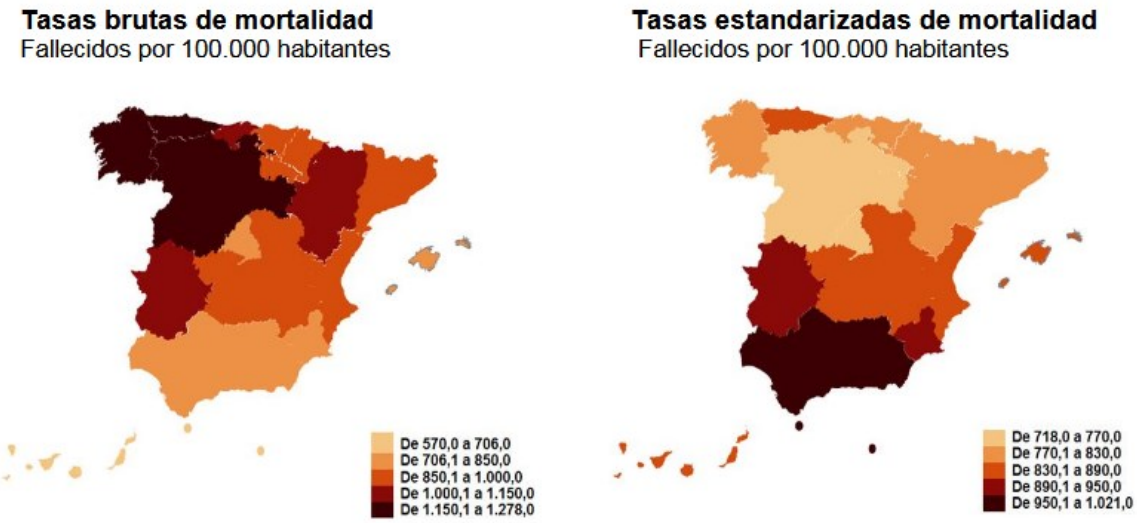
iii. **Pirámide de población:** consiste en dos histogramas, uno para hombres y otro para mujeres, correspondientes a los habitantes de una cierta comunidad, repartidos por edades. Este tipo de gráficos resultan de utilidad para el estudio de la situación demográfica.

Ejemplo: Imagen obtenida de la publicación "España en cifras 2018" de INE (Fuente: INE)



iv. **Cartogramas:** es una representación gráfica que muestra asociado a áreas geográficas datos referidos a variables cuantitativas. Los datos pueden representarse en cada área asignando distintos colores o símbolos con puntos o rayas. En otras ocasiones lo que se hace es modificar la superficie real del área ajustándolo proporcionalmente a la variable que se está representando.

Ejemplo: Imagen obtenida de la Nota de Prensa de Defunciones según causa de muerte del año 2017. (Fuente: INE)



Unidad 5. Medidas descriptivas de una variable estadística. Medidas de posición y de dispersión. Medidas de forma y concentración.

5.1 Medidas descriptivas de una variable estadística

En la Unidad anterior hemos presentado las tablas y las distintas representaciones gráficas, como posibles recursos para ver de una manera organizada y de un vistazo las principales características de la distribución de los datos de una variable a estudio. Pero en ocasiones resulta complicado, o poco factible, hacerse una idea de que valores tiene una variable.

Así pues, una vez que recopilemos un conjunto de datos, no solo debemos tabularlos y obtener su distribución de frecuencias, sino que resultaría útil sustituir todos los valores observados por uno sólo que sintetizara e informara del algún aspecto relevante de la distribución. Estas medidas representarán al conjunto de valores y nos facilitarán la comparabilidad con otros conjuntos de datos, sin tener que recordar las frecuencias absolutas y relativas de cada uno de ellos.

Estas medidas que condensan la información pueden ser de varios tipos: medidas de posición, medidas de dispersión, de forma y de concentración. A continuación pasamos a ver cada una de éstas de manera detallada.

5.2 Medidas de posición

Las medidas de posición son expresiones que resumen la información relativa a la variable de estudio de un conjunto completo de datos, a través de un único valor que refleja el centro de la distribución de los datos.

Las medidas de posición más utilizadas y que vamos a ver son: la media aritmética, la mediana, la moda y los cuantiles.

1. **Media Aritmética**

Cualquier individuo experto o no en estadística piensa en la media como medida resumen de un conjunto de datos. Esta es la primera medida de posición que vamos a ver, la media aritmética o simplemente media.

Definición: se define la media aritmética como la suma de todos los valores de dicha variable dividida por el por el número total de observaciones, y lo denotaremos por $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad donde\ x_i\ es\ el\ valor\ i-ésimo\ de\ la\ variable$$

Si los datos se encuentran ordenados y organizados en una tabla de frecuencias, la media aritmética se calcula de la siguiente forma

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

En el caso de que estemos trabajando con datos agrupados, debemos trabajar con marcas de clase (c_i). Lo que no debemos obviar es que la media variará "ligeramente" si consideramos los datos agrupados frente a no agrupados. La expresión en este caso para la media sería la que sigue:

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{N}$$

Cuando calculamos la media aritmética de una serie de datos, en la fórmula sumamos todos ellos, dando de esta forma la misma importancia a cada valor. Pero en ocasiones queremos dar más "peso" a alguno de los valores sobre el resto porque lo podemos considerar más importante. Es decir, hablamos de otorgar a cada valor de la variable x_i una ponderación o peso distinto de la frecuencia de repetición. En este caso la media se denomina ponderada y se definiría como

$$\bar{x} = \frac{\sum_{i}^{N} x_i w_i}{\sum_{i}^{N} w_i} = \frac{\sum_{i}^{k} x_i w_i n_i}{\sum_{i}^{k} w_i} \quad donde\ w_i\ son\ las\ ponderaciones\ respectivas$$

En caso de distribución de frecuencias

Propiedades de la media aritmética

- El valor de la media aritmética es un número comprendido entre el mínimo y el máximo de los valores que toma la variable.
- La media aritmética representa el centro de gravedad de un conjunto de datos. Esto quiere decir que si calculamos las distancias de cada uno de los valores respecto a la media, la suma de los que están por encima de ella es igual a la suma de los que son menores.

$$\sum_{i}^{n}(x_i - \bar{x})\, n_i = 0$$

- No es robusta porque se ve claramente afectada por valores extremos. Es decir, que si uno de los valores del conjunto de datos es muy elevado con respecto a los demás, hará subir la media comparada con el resto de valores. Y viceversa, si uno de los datos es demasiado pequeño en relación con el resto, la media aritmética se ve afectada hacia abajo. Por este motivo cuando se calculan las medias hay que observar que no haya valoras anormalmente distanciados del resto. Es lo que se denominan valores outliers o valores anómalos. Si se detecta la presencia de éstos, para evitar que alteren la media resultante a veces se suelen eliminar en el cálculo. En ese caso hablamos de la media truncada.

- La media se ve afectada por cambios de origen. Sea X la variable de estudio y supongamos que X puede tomar un número N de valores distintos. Representamos dichos valores por $x_1, x_2, \ldots x_N$. Si aplicamos un cambio de origen (b) los nuevos valores serán $x_1 + b, x_2 + b, \ldots x_N + b$. Entonces la media de los nuevos valores es

$$\bar{x} = \frac{\sum_{i=1}^{N}(x_i + b)}{N} = \frac{\sum_{i=1}^{N} x_i}{N} + \frac{Nb}{N} = \bar{x} + b$$

- La media aritmética se ve afectada por cambios de escala. Si los valores de una variable los multiplicamos a todos por una constante(a), la media de los nuevos datos queda también multiplicada por dicha constante a.

$$\bar{x} = \frac{\sum_{i=1}^{N}(ax_i)}{N} = a\frac{\sum_{i=1}^{N} x_i}{N} = a\bar{x}$$

Ejercicio 5.1: Supongamos que queremos calcular la renta media de cinco países, cada uno de los cuáles tiene una población diferente:

País	Renta per cápita	Población
Suiza	85.160	8.714.514
Francia	43.500	65.626.338
Reino Unido	42.040	66.447.024
Italia	34.780	59.992.606
España	31.910	45.743.867

Se pide calcular la renta media. ¿Sería posible calcular una media ponderada?

Solución 5.1: Si calculamos la renta media como la media aritmética de las rentas, sumaríamos la de los 5 países y lo dividiríamos por cinco:

$$\bar{x} = \frac{85.160 + \cdots + 34.780}{5} = 47.478$$

Es decir que diríamos que la renta media de los 276 millones de personas que viven en estos 5 países está entorno a los 47 mil euros. Pero para calcular la media de las rentas estamos "valorando" la renta de Suiza, que no llega a los diez millones de habitantes, de la misma manera que la renta de Francia que tiene 66 millones de habitantes. Tendríamos que utilizar otro mecanismo que no otorgara la misma relevancia a cada uno de los países. Por ello, parece que la media ponderada puede ajustarse más a nuestras necesidades, donde sumaremos las rentas, pero cada una ponderada por el peso de su población:

$$\bar{\mathrm{x}} = \frac{\sum_{i=1}^{5} w_i . X_i}{\sum_{i=1}^{5} w_i} = \frac{8.714.514x85.160 + \cdots + 45.743.867x31.910}{8.714.514 + \cdots + 45.743.867} = 40.306{,}51$$

Según esta fórmula la renta media de los 246 millones de personas de estos países es menor que la que nos salía con la media aritmética. Y es razonable, puesto que de esos 246 millones, más de 100 millones de individuos(los residentes en España e Italia) no superan los 35 mil euros, con lo parece que la medida de centralización más adecuada en la media ponderada y así situar el dato medio en los 40.306.

2 Media geométrica

La media geométrica se maneja para calcular valores medios de cualquier magnitud que suponga variaciones entre distintos puntos en el tiempo como las tasas, tipos de interés a lo largo de varios periodos, etc.

Definición: Sea X la variable de estudio y supongamos que X puede tomar un número k de valores distintos $x_1, x_2, \ldots x_k$, cuyas frecuencias absolutas son $n_1, n_2, \ldots n_k$ respectivamente y tal que $N = \sum_i^n n_i$ es el número total de observaciones. Se define la media geométrica como:

$$\overline{x_g} = \sqrt[N]{\prod_{i=1}^{k} X_i^{n_i}} = \sqrt[N]{X_1^{n_1} * \ldots * X_k^{n_k}}$$

Al igual que la media aritmética se ve afectada por los cambios de escala.

Como inconvenientes de esta medida de posición podríamos señalar los siguientes:

- Si los valores son negativos no siempre es posible calcular la raíz cuadrada. O si uno de los valores es cero, entonces la media siempre será cero, lo que no tendría sentido.
- El significado estadístico es menos intuitivo que el de la media aritmética.

Por otro lado, la media geométrica de diferentes tasas de varios periodos se puede interpretar como la tasa media constante que debería tener cada periodo para obtener la misma variación global desde el momento inicial al final que nos dan los valores de los que calculamos la media. Veámoslo con un ejemplo.

Ejercicio 5.2 El tipo de interés de depósito a plazo está referenciado al Euribor, por lo que cada año se obtienen unas ganancias diferentes. Los últimos 5 años han sido: 3,4% 4,5%, 2,3%, 2,1%, 1,7%. Calcular la media geométrica

Solución 5.2:

La tasa asociada para cada uno de los cinco últimos años serían 1,034; 1,045; 1,023; 1,021; 1,017. Si calculamos la media geométrica de estos valores obtenemos:

$$\overline{x_g} = \sqrt[5]{1{,}034\,.\,1{,}045\,.\;1{,}023\,.\;1{,}021\,.\,1{,}017} = 1{,}02794$$

La media geométrica de las rentabilidades sería 2,794%. Es decir, que si cada año el tipo de interés hubiera sido constante en este valor, la rentabilidad del depósito hubiera sido la misma que la que nos ha dado estos cinco valores. Esto es fácil de comprobar: Si el depósito inicial es de 100 euros y el tipo de interés primero es de 3,4%, el primer año tendríamos 3,4 euros de ganancia. Supongamos que dejamos invertido en el mismo depósito este beneficio, y aplicamos el tipo de interés del segundo año (4,5%). Entonces la rentabilidad que se obtendría sería ahora la cantidad depositada (103,4 euros) por el tipo de interés: 103,4*4,5%=4,65 euros. Así aplicando los tipos de interés sucesivamente durante los cinco periodos, tendríamos al final de cinco años que nuestro depósito sería de 114,78 euros. Pues bien, si durante los cinco años el tipo de interés hubiera sido en todos ellos la media geométrica (2,794%), los 100 euros también se hubieran convertido en 114,78 euros. Por ese motivo se dice que es un valor medio. Si hubiéramos calculado la media aritmética de los cinco tipos de interés, el valor resultante no hubiera cumplido esta propiedad.

3 Otras medidas de posición

No siempre la media es la mejor de las opciones como medida de posición, veamos una situación que refleja esto. Por ejemplo, consideremos la distribución de las distintas categorías de nivel de estudios de los residentes en una urbanización de un municipio, ¿se podría obtener una medida que fuese una categoría media? Es obvio que no con una media aritmética. Pero sí que podríamos decir cuál es la categoría del nivel de estudios de la mayoría de los residentes en la citada urbanización. O también podríamos ordenar las categorías de menor a mayor, y ver cuál es la categoría que tiene al menos la mitad de los residentes.

Las medidas que vamos a ver a continuación, también reflejan de alguna manera una posición dentro de los posibles valores que toma la variable, por lo tanto son también medidas de posición. Vamos a ver tres de ellas: Moda, Mediana y Cuantiles.

3.1 Moda

La moda de una variable estadística discreta es el valor de la variable que más veces aparece (el que tiene una mayor frecuencia absoluta). Se representa por las letras Mo.

En el caso que tuviéramos una variable estadística continua agrupada en intervalos, hablaríamos del intervalo o clase modal, que sería aquella clase que presente la mayor frecuencia.

En el caso en el que estemos tratando con una variable estadística en escala nominal, la moda resulta ser la medida más representativa de dicha distribución, ya que en estos casos no se pueden calcular otras medidas al no poderse establecer un orden.

3.1.1 Propiedades de la moda

- Al igual que ocurría con la media, la moda queda afectada tanto por cambios de origen como de escala.
    - $z_i = \mathrm{a}x_i + b \;\; donde\ a\ y\ b\ son\ constantes\ Mo_z = aMo_x + \mathrm{b}$
- La moda de una variable estadística no tiene por qué ser única. Hablamos de distribución unimodal cuando es única, multimodal cuando existe más de una.

3.2 Mediana

Si recordamos, uno de los principales problemas que plantea el uso de la media aritmética como media de posición central es que resulta muy sensible a la presencia de valores extremos. Por ello, es necesario establecer otras medidas de localización que no presenten este inconveniente y sean más estables a la presencia de valores atípicos.

Definición: Si se ordenan los datos de una distribución de menor a mayor, la mediana de dichos datos es el valor que ocupa el lugar central de todos ellos, es decir, la mediana sería el que deja a su izquierda y a su derecha el 50% de los datos. Se denota por Me.

La expresión para calcular la mediana depende si el número de datos es par o impar:

- El número de observaciones es impar, la Me="valor que ocupa la posición (N+1)/2 ".
- El número de observaciones es par: la Me="media aritmética de los dos valores centrales, es decir, de los que ocupan las posiciones (N/2) y (N/2)+1"

3.2.1 Propiedades de la mediana

- Al depender la mediana sólo del orden de las observaciones, no se ve afectada por valores extremos.

- La mediana queda afectada por los cambios de origen y de escala, al igual que ocurre con la media aritmética y la moda.

$$z_i = ax_i + b \ \ donde\ a\ y\ b\ son\ constantes\ Me_z = aMe_x + b$$

3.3 Cuantiles

Definición: Sea X la variable de estudio y supongamos que X puede tomar un número k de valores distintos $x_1, x_2, \dots x_k$, cuyas frecuencias absolutas son $n_1, n_2, \dots n_k$ respectivamente y tal que $N = \sum_i^n n_i$. Los cuantiles son en una distribución aquellos valores que la dividen en partes que contienen el mismo número de observaciones.

Dentro de los cuantiles vamos a mencionar aquellos cuyo uso es más habitual:

- **Cuartiles**: Son tres puntos o valores que consiguen dividir en cuatro partes en las que hay un 25% de observaciones en cada una de ellas. Se denotan por Q_i i=1,2,3).
- **Deciles**: dividen a la distribución en 10 partes iguales. Resulta ser nueve valores que denotaremos por D_i i=1,2,…9.
- **Percentiles**: son noventa y nueve valores que dividen a la distribución en partes iguales. La mediana es el Percentil 50 y los cuartiles son los Percentiles 25, 50 y 75.

Para su obtención se procede en primer lugar a obtener las frecuencias acumuladas y a continuación se busca el valor de la distribución que ocupe el lugar $\frac{s}{r}N$. Donde s es el cuantil que se desea calcular y r las partes en que se divide la distribución. Si deseáramos calcular el tercer cuartil, s=3 y r=4.

Cuando los datos se encuentran agrupados, el primer paso es identificar el intervalo de clase al que pertenece el cuantil. De un total de N datos, el cuantil ocupará la posición $\left(\frac{s}{r}\right)$N. Observaremos en la columna de frecuencias acumuladas, en que intervalo se encuentra

El cálculo se realiza como se indica

$$C_s = L_{i-1} + \frac{\frac{s}{r}N - N_{i-1}}{n_i} a_i$$

Donde C_s es el valor del cuantil que se busca, s el número de cuantil que se busca, r el número de partes en que se divide la distribución , a_i la amplitud del intervalo al que pertenece el cuantil y L_{i-1} el límite inferior del intervalo de clase.

5.3 Medidas de dispersión

Hemos visto un conjunto de medidas que nos ofrecen una idea de la localización de los datos muestrales, como por ejemplo su "centro de gravedad" (media aritmética) o su "punto medio" (mediana), entre otros. Pero es posible que dichas medidas no sean suficientes para la descripción de un conjunto de datos.

Por ejemplo, si en una clase de cuarto de primaria todos los alumnos obtienen una calificación de 7 en la asignatura de lengua, la nota media de todos los alumnos también será 7 y este dato sería una medida representativa de las notas de los alumnos. Ahora bien, supongamos otra clase de cuarto de primaria donde hay 20 personas, 10 de ellas obtienen un 4 y otras 10 sacan un 10. Aunque la nota media en esta segunda clase sería la misma, no resultaría representativa puesto que es una clase en la que suspenden la mitad de los individuos y la otra mitad son alumnos de matrícula de honor.

Por lo tanto necesitamos introducir algunas nuevas medidas que nos indiquen hasta qué punto las medidas de posición son valores distintivos del conjunto de los datos. Está claro que si los valores de la variable estuvieran en torno a las medidas de posición, éstas serían lo suficientemente representativas. Ahora bien, en el caso de existir dispersión en los valores del conjunto de datos a estudio, disminuiría la representatividad y tendríamos que buscar otras expresiones que minimizaran esto.

En este epígrafe nos vamos a centrar en el estudio de la dispersión de los datos y su representación mediante alguna medida resumen de la misma.

1. Rango o Recorrido

Se define por rango o recorrido de un conjunto de datos, a la diferencia entre el mayor y el menor de los valores. Resulta de utilidad cuando la variable estudiada está uniformemente distribuida.

2. Recorrido Intercuartílico

Se define por recorrido intercuartílico a la diferencia entre el tercer y el primer cuartil. Es una medida de dispersión cuyo intervalo contiene al 50% de los datos.

3. Varianza

Dado un conjunto de datos $x_1, .., x_N$, se define su varianza como:

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

Si los datos están agrupados en una tabla, la varianza se obtendrá con la siguiente expresión:

$S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2$, donde $N=n_1+n_2+\ldots+n_k$ y siendo $x_1, .., x_k$, los valores que puede tomar la variable X, cuyas frecuencias absolutas son $n_1, n_2, \ldots n_k$ respectivamente.

Propiedades de la varianza

- La varianza toma valores positivos o cero.
- Al igual que la media se ve afectada por la presencia de valores extremos.
- No se ve afectada por los cambios de origen

$$z_i = x_i + b \ \ donde\ b\ es\ una\ constante\ S^2{}_z = S^2{}_x$$

- Sí se ve afectada por los cambios de escala

$$z_i = x_i a \ \ donde\ a\ es\ una\ constante\ S^2{}_z = a^2 S^2{}_x$$

4. Desviación típica

Hemos visto como la varianza es una medida de dispersión que se calcula al agregar las distancias elevadas al cuadrado. Por lo tanto si ponemos unidades a esta dispersión sería "unidades al cuadrado". Si queremos que se empleen las unidades de partida de los datos, podemos obtener una nueva medida de dispersión haciendo uso de la raíz cuadrada positiva de la varianza, a la que se denomina desviación típica.

$$S = +\sqrt{S^2}$$

5. El Coeficiente de variación

Como hemos visto para calcular la varianza y la desviación típica se consideran las magnitudes en que vienen expresados los datos. Esto puede hacer que los datos obtenidos estén "contaminados" por la magnitud de estas unidades y para comparar dos conjuntos de datos distintos perdamos la perspectiva de cuál está más disperso.

Por esa razón se define una medida de dispersión relativa: el coeficiente de variación. Se define éste como el cociente entre la desviación típica y la media aritmética:

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

Tiene la ventaja de que permite comparar distribuciones de datos medidos en unidades distintas. Ahora bien, su inconveniente principal resulta cuando la media aritmética de los datos muestrales vale cero y por ello dicho coeficiente no estaría definido.

5.4 Medidas de forma

A parte de las medidas vistas en los epígrafes anteriores, existen otras que miden y describen la forma de la distribución de los datos, sin necesidad de recurrir a la representación gráfica de los mismos.

1. Medidas de asimetría

Las medidas de asimetría (o simetría) permiten conocer si los datos están dispuestos de manera simétrica o asimétrica con respecto a un valor central (media aritmética).

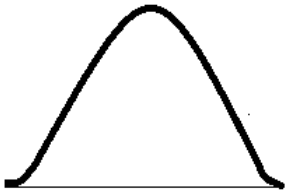
Uno de los posibles índices de simetría es el que se encuentra a continuación, que es el coeficiente de asimetría de Fisher. Este mide el nivel de asimetría de la distribución de los datos respecto a la media aritmética. Su expresión es la que sigue en el caso que los datos se encuentren agrupados:

$$\gamma_1 = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^3}{N} n_i}{S^3}$$

Analizando el coeficiente podemos decir:

- Si toma valores negativos$(\gamma_1 < 0)$ la distribución será asimétrica negativa

- Si es igual a cero($\gamma_1 = 0$) será simétrica



- Si toma valores positivos ($\gamma_1 > 0$) diremos que es asimétrica positiva.

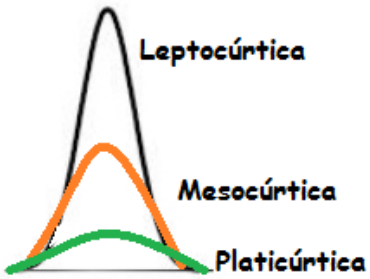
2. Medidas de curtosis

El coeficiente de curtosis o aplastamiento mide el grado de concentración que presentan los datos en torno a la media. Tomamos a la distribución normal como distribución de referencia para medir la curtosis. La expresión del coeficiente es la siguiente:

$$\gamma_2 = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^4}{N} n_i}{S^4} - 3$$

La clasificación de las variables estadísticas según dicho coeficiente sería:

- Si $\gamma_2 > 0$ la distribución es leptocúrtica o más apuntada que la normal.
- Si $\gamma_2 = 0$ tiene el mismo nivel de apuntamiento que la normal
- Si $\gamma_2 < 0$ es platicúrtica o menos apuntada que la normal



Resulta ser un coeficiente adimensional y la constante que se resta (3) se ha elegido para que en la distribución normal el coeficiente valga cero.

5.5 Medidas de concentración

En este epígrafe vamos a introducir lo que se conoce con el nombre de medidas de concentración, con ellas se evalúa el mayor o menor grado de igualdad en el reparto del total de los valores de la variable.

1. Curva de Lorenz

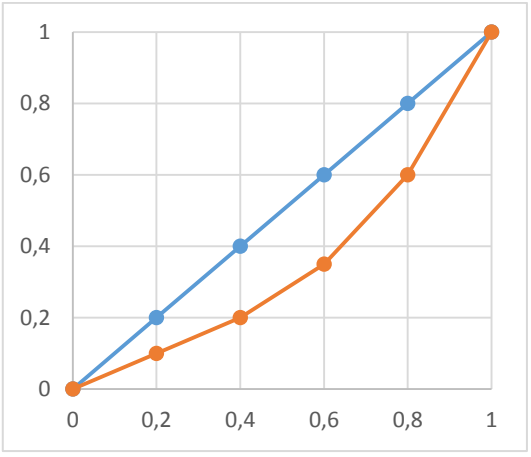
La curva de Lorenz es la forma gráfica de visualizar concentraciones. Como ejemplo resulta útil en casos como el de representar la desigualdad del reparto de la renta existente en un determinado territorio.

Definición: Sea X la variable de estudio y supongamos que X puede tomar un número k de valores distintos $x_1, x_2, \dots x_k$ ordenados de manera creciente, cuyas frecuencias absolutas son $n_1, n_2, \dots n_k$ respectivamente, entonces definimos

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

$$q_i = \frac{\sum_1^i x_j n_j}{\sum_1^k x_j n_j}$$

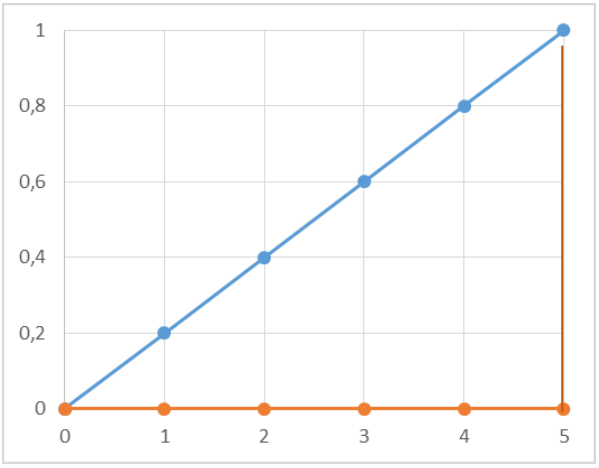
Denominamos Curva de Lorenz a la poligonal L(p) que une los puntos $\{(\boldsymbol{p_i}, \boldsymbol{q_i})\ \ \boldsymbol{i = 1 \dots k}\}$



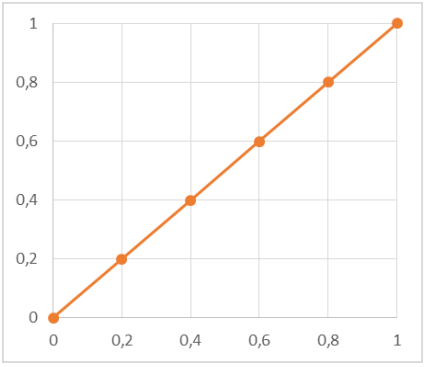
En la figura superior la Curva de Lorenz sería la línea naranja. Hacemos notar que la curva se encuentra siempre por debajo de la diagonal del cuadrante y el área que queda entre la diagonal y la curva de indica el grado de concentración. A mayor área, mayor es la concentración.

Veamos los casos extremos que se presentan:

- Todas las unidades menos una de la población no poseen nada de la variable a estudiar (renta, cuota de mercado, etc), y solo una tiene el 100%. Entonces la gráfica queda pegada a los lados del recuadro inferior. Sería un caso de máxima concentración.



- Todas las unidades acumulan de forma equitativa la variable investigada. Entonces la gráfica quedaría en la diagonal. Estaríamos ante un caso de máxima homogeneidad.



2. Índice de Gini

Esta misma expresión de medida de desigualdad (o concentración) también puede expresarse numéricamente. La medida se define a través del índice de Gini, que compara los valores acumulados de las unidades de la población y de la magnitud objetivo.

Partiendo de la notación anterior, la expresión del índice de Gini es:

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i}$$

Este índice toma valores entre 0 a 1 y es proporcional al área que queda entre la diagonal y la curva de Lorenz.

- Si $I_G = 0$ entonces hay una concentración mínima, es decir un reparto totalmente equitativo de la magnitud observada entre las unidades
- Si $I_G = 1$ entonces hay una concentración máxima, es decir todo queda concentrado en una única unidad o grupo de unidades.

Unidad 6. Distribuciones bidimensionales. Tablas de correlación y contingencia. Distribuciones marginales y condicionadas. Independencia estadística.

En la Unidad 5, hemos estudiado el comportamiento de una única variable X medida sobre cada elemento de una misma población o muestra. En las Unidades 6 y 7, vamos a desarrollar una serie de conceptos necesarios cuando se está interesado en el estudio simultáneo de dos variables para evaluar la posible existencia de relaciones entre ellas.

Un ejemplo básico en el que nos podría interesar el estudio simultáneo de dos características en una población o muestra, sería la relación entre el ingreso y el gasto de las familias en el territorio nacional.

6.1 Distribuciones bidimensionales

Supongamos que sobre una misma población (o muestra de una población) de N individuos se estudian dos variables que denotaremos por $X\, e\, Y$, donde la variable X presenta k valores distintos $x_1, x_2, \dots x_k$ y la variable Y presenta m valores distintos, $y_1, y_2, \dots y_m$.

Denominaremos distribución bidimensional de frecuencias (X, Y) al conjunto de ternas $\{(x_i, y_j, n_{ij})\ \ i = 1 \dots k, j = 1 \dots . m)\}$, donde n_{ij} denota la **frecuencia absoluta conjunta**, que es el número de veces que se presenta en la muestra el valor x_i de la variable X con el valor y_j de la variable Y. La frecuencia absoluta conjunta verifica:

$$0 \le n_{ij} \le N, i \in (1,..k), j \in (1,..m) \quad y \quad \sum_i^k \sum_j^m n_{ij} = N$$

Tal y como ocurría en el caso unidimensional, hablamos de frecuencia relativa, que denotaremos por f_{ij}, l cociente entre la frecuencia absoluta conjunta y el número total de observaciones

$$f_{ij} = \frac{n_{ij}}{N}, 0 \le f_{ij} \le 1 \ \ i \in (1,..k), j \in (1,..m) \ \ y \sum_i^k \sum_j^m f_{ij} = 1$$

6.2 Tablas de correlación y contingencia

Cuando abordamos el caso unidimensional, se habló de la necesidad de presentar la información de la característica a estudio de manera organizada y para ello introdujimos las tablas de distribuciones de frecuencias. Las **tablas de doble entrada** son el recurso disponible para la organización de los datos en el caso bidimensional.

La tabla de doble entrada se construye de la siguiente manera: en la primera columna se sitúan los diferentes valores observados para la variable X y en la primera fila se disponen los diferentes valores para la variable Y que han sido observados. Y en el interior se colocan los valores correspondientes a las frecuencias conjuntas, que pueden ser tanto absolutas como relativas dependiendo de los objetivos y fines del estudio.

En el caso de que las variables a estudio sean cuantitativas hablamos de **tabla de correlación** y si son cualitativas de **tabla de contingencia**.

18.2.1 Tablas de correlación

Supongamos una población con un total de N individuos, en la que se estudian simultáneamente dos variables cuantitativas, X e Y, que nos definen una variable estadística bidimensional. La **tabla de correlación** tendría el siguiente aspecto:

X_i / Y_j	y_1	y_2	.	.	.	y_j	.	.	.	y_m	$n_{i.}$
x_1	n_{11}	n_{12}				n_{1j}				n_{1m}	$\boldsymbol{n_{1.}}$
x_2	n_{21}	n_{22}				n_{2j}				n_{2m}	$\boldsymbol{n_{2.}}$
.											
.											
x_j	n_{j1}	n_{j2}				n_{jj}				n_{jm}	$\boldsymbol{n_{j.}}$
.											
.											
x_k	n_{k1}	n_{k2}				n_{kj}				n_{km}	$\boldsymbol{n_{k.}}$
$n_{.j}$	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	.	.	.	$\mathbf{n_{.j}}$	.	.	.	$\mathbf{n_{.m}}$	N

Donde $N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$ $n\acute{u}mero\ total\ de\ observaciones$

$$\boldsymbol{n_{.j}} = \sum_{i=1}^{k} n_{ij}\ \ frecuencia\ marginal\ de\ y_j\ , j\ \in (1, \dots m\ )$$

$$\boldsymbol{n_{i.}} = \sum_{j=1}^{m} n_{ij}\ \ frecuencia\ marginal\ de\ x_i\ \ \ , i\ \in (1, \dots k\ )$$

La suma de las frecuencias marginales es igual a N

$$N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{m} n_{.j}$$

Una serie de observaciones antes de terminar el epígrafe:

- Si trabajamos con datos agrupados, tanto para una como para las dos variables, sustituimos en la tabla anterior el valor de la variable por la marca de clase.
- Si queremos pasar de tener una tabla con frecuencias absolutas a una con frecuencias relativas, sólo es preciso dividir toda frecuencia absoluta por el

número total de observaciones, obteniendo así la de frecuencias relativas que es la que se precisaba.

6.2.2 Tablas de contingencia

Hablamos de **tabla de contingencia** cuando las variables objeto de estudio son cualitativas.

Resulta ser una tabla como la anterior, salvo que en la que la primera columna y en la primera fila se expresan modalidades de los atributos en vez de los valores numéricos de las variables cuantitativas.

Hacemos notar que es posible elaborar tablas de contingencia combinando variables cualitativas y cuantitativas.

6.3 Distribuciones marginales y condicionadas

18.3.1 Distribuciones marginales

Las distribuciones marginales de una distribución bidimensional aparecen del estudio por separado de cada una de las variables implicadas.

Dada una distribución bidimensional $\{(x_i, y_j, n_{ij})\ \ i = 1 \dots k\ , j = 1 \dots . m)\}\ y\ N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$.

- Llamaremos **distribución marginal de X**, al conjunto de pares $\{(x_i, n_{i.})\ \ i = 1 \dots k\}$ donde $n_{i.} = \sum_{j=1}^{m} n_{ij}$ es la frecuencia marginal de $x_i \ \ para\ todo\ i \in (1, \dots k\ )$ y que verifica $0 \leq n_{i.} \leq N\ para\ todo\ i \in (1, .. k)\ y \ \ \sum_{i=1}^{k} n_{i.} = N$
- Análogamente, llamaremos **distribución marginal de Y**, al conjunto de pares $\{(y_j, n_{.j})\ \ i = 1 \dots m\}$ donde $n_{.j} = \sum_{i=1}^{k} n_{ij}$ es la frecuencia marginal de $y_j\ para\ todo\ \ j \in (1, \dots m\ )$ y que verifica $0 \leq n_{.j} \leq N\ para\ todo\ j \in (1, .. m)\ y \ \ \sum_{j=1}^{m} n_{.j} = N$

Si en vez de frecuencias absolutas consideramos frecuencias relativas, las frecuencias relativas marginales de $X\ e\ Y$ son respectivamente:

$$f_{i.} = \frac{n_{i.}}{N} \quad \forall\, i \in (1 \dots k\,) \qquad f_{.j} = \frac{n_{.j}}{N} \quad \forall\, j \in (1 \dots m\,)$$

$$\sum_{i=1}^{k} f_{i.} = \sum_{j=1}^{m} f_{.j} = 1$$

Resulta de utilidad presentar la distribución de frecuencias marginales en una tabla de manera separada y pueden obtenerse todas las medidas estudiadas para el caso unidimensional.

6.3.2 Distribuciones condicionadas

Las distribuciones condicionadas dicen cómo se distribuye una de las variables sobre un conjunto de elementos de una población que toman un determinado valor de los posibles de la otra variable.

Dada una distribución bidimensional $\{(x_i, y_j, n_{ij})\ \ i = 1 \dots k\,, j = 1 \dots . m)\}$ y $N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$.

- Llamaremos distribución de X condicionada a Y $(X/Y = y_j)$, al conjunto de pares $\{(x_i, n_{ij})\ \ i = 1 \dots k\ y\ \ para\, j\ dado\ \}$ donde $n_{.j} = \sum_{i=1}^{k} n_{ij}$ es la frecuencia marginal X.

- Análogamente, llamaremos distribución de Y condicionada a X $(Y/X = x_i)$, al conjunto de pares al conjunto de pares $\{(y_j, n_{ij})\ \ j = 1 \dots m\ \ y\, para\, i\ dado\}$ donde $n_{i.} = \sum_{j=1}^{m} n_{ij}$ es la frecuencia marginal Y.

  En el caso de las distribuciones condicionadas se suele trabajar con frecuencias relativas en vez de absolutas.

- Frecuencia relativa condicionada para X dado que $Y = y_j$ es

$$f_{i/j} = \frac{n_{ij}}{n_{.j}}\ para\ todo\ i = 1 \dots k\ y\ para\ j\ fijo$$

La tabla correspondiente a la distribución $X/Y = y_j$ es la siguiente:

$X/Y = y_j$	n_{ij}	$f_{i/j} = \frac{n_{ij}}{n_{.j}}$
x_1	n_{1j}	$f_{1/j}$
…		
x_k	n_{kj}	$f_{k/j}$
	$n_{.j}$	1

- Frecuencia relativa condicionada para Y dado que $X = x_i$ es

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} \; para\ todo\ j = 1 \ldots m\ y\ para\ i\ fijo$$

La tabla correspondiente a la distribución $Y/X = x_i$ es la siguiente:

$Y/X = x_i$	n_{ij}	$f_{j/i} = \frac{n_{ij}}{n_{i.}}$
y_1	n_{i1}	$f_{1/i}$
…		
y_m	n_{im}	$f_{m/i}$
	$n_{i.}$	1

Puesto que las distribuciones condicionadas son unidimensionales, es posible obtener las medidas unidimensionales estudiadas para su descripción.

6.4 Independencia estadística

Diremos que dos variables son independientes una de la otra, cuando los valores que tome una de ellas no afectan a la otra y viceversa.

Entonces diremos que la variable Y se distribuye independientemente de la variable X, si y sólo si la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir,

$$f_{ij} = f_{i.}f_{.j} \quad \forall i \forall j \Leftrightarrow \frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N} \; \forall i \forall j \Leftrightarrow n_{ij} = \frac{n_{i.}n_{.j}}{N} \quad \forall i \forall j$$

En este caso, las frecuencias relativas condicionadas serán

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}n_{.j}}{Nn_{.j}} = \frac{n_{i.}}{N} = f_{i.} \forall j$$

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}n_{.j}}{Nn_{i.}} = \frac{n_{.j}}{N} = f_{.j} \, \forall i$$

Es decir, las frecuencias relativas condicionales son iguales a sus correspondientes frecuencias relativas marginales.

Unidad 7. Medidas descriptivas de variables bidimensionales: Covarianza y coeficiente de correlación lineal. La recta de regresión.

7.1 Medidas descriptivas de variables bidimensionales

Tal y como hemos comentado en la Unidad anterior, resulta de interés el estudio de manera simultánea de dos variables para poder ver la posible existencia de relación entre ellas.

Como primera aproximación al estudio de dicha relación entre dos variables, $X\ e\ Y$, podríamos hacer uso de la representación gráfica. En concreto el gráfico que se usa se conoce con el nombre de diagrama de dispersión o nube de puntos. Otra forma es través de medidas que reducen los datos bidimensionales permitiendo tener una idea general de la distribución conjunta sin tener que recordar todos los pares de valores con sus frecuencias. En particular, vamos a estudiar dos medidas: la covarianza y el coeficiente de correlación lineal.

7.1.1 Covarianza

La covarianza mide la asociación lineal de dos variables, cuyo signo nos indica el sentido de la variación conjunta de las dos variables consideradas.

Definición: Partimos de una población de N individuos, en la que se estudian simultáneamente dos variables $X\ e\ Y$ que nos definen una variable estadística bidimensional. Cada individuo vendrá dado por el par de valores $\{(x_i, y_i)\ , i = 1 \dots N\}$. Supongamos que la variable X presenta k valores distintos $x_1, x_2, \dots x_k$ y la variable Y presenta m valores distintos, $y_1, y_2, \dots y_m$. La covarianza presenta la siguiente expresión:

- Si los datos no están agrupados

$$S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i\, y_i \ - \ \bar{x}\,.\,\bar{y}$$

- Si los datos están agrupados, tal que la distribución bidimensional viene dada por $\{(x_i, y_j, n_{ij}), i = 1 \dots k,\ \ j = 1, \dots, m\}$

$$S_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{n_{ij}(x_i - \bar{x})(y_j - \bar{y})}{N} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{m} x_i\, y_j \ - \ \bar{x}\,.\,\bar{y}$$

Si la covarianza presenta un valor positivo se observa una asociación lineal creciente, mientras que si el signo de la covarianza es negativo, la relación será decreciente. El indicativo de que no existe relación entre las variables es que la covarianza sea nula.

Sin embargo, la covarianza presenta un inconveniente, y es que su valor depende de las unidades de medida de las variables. Esto hace necesario un coeficiente adimensional que permita medir la intensidad de la relación entre variables y comparar el grado de asociación lineal entre distintos pares de variables.

Observación: Cuando las variables son independientes la covarianza vale cero, sin embargo el recíproco no es necesariamente cierto. Puede suceder que la covarianza sea cero y que las variables a estudio no sean independientes.

19.1.2 Coeficiente de correlación lineal

El coeficiente de correlación lineal es una medida del grado de dependencia lineal mutua entre variables adimensional. Su expresión es la siguiente:

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

Donde S_{xy} *es la covarianza de las variables X e Y*

S_x *y* S_y *son las desviaciones típicas de X e Y respectivamente*

Algunas consideraciones sobre el coeficiente de correlación lineal:

- Su signo viene determinado por el de la covarianza.
- El valor de este coeficiente está comprendido entre -1 y 1. Si el valor es -1 o próximo, indica una fuerte asociación negativa, y si es 1 o próximo a 1 indica una asociación lineal positiva.

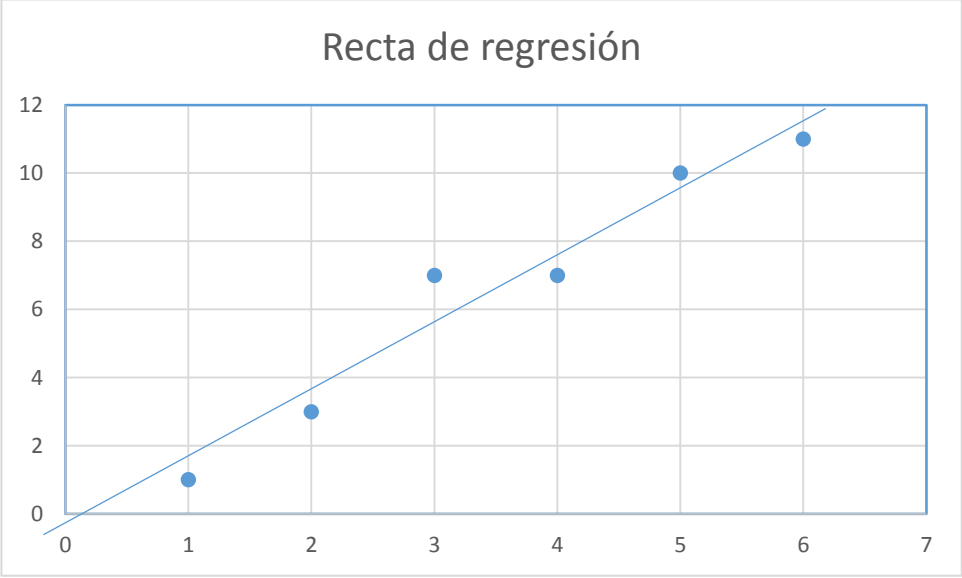
- En el caso de que su valor sea 0, las dos variables no tienen ningún tipo de correlación lineal, por lo que serán linealmente independientes.

7.2 Recta de regresión

En los epígrafes anteriores de la Unidad, hemos mencionado algunas expresiones matemáticas que nos permiten medir el grado de asociación lineal entre pares de variables. Ahora bien, en el caso de que dicha asociación existiera, lo que nos interesa es poder establecer una relación entre ambas que permita predecir los valores de una en función de la otra.

Para este propósito están las técnicas de regresión, que pretenden establecer previsiones de futuro sobre la evolución de una variable haciendo supuestos o conociendo la evolución de otras variables que explican su comportamiento.

En nuestro caso vamos a ver la regresión lineal, es decir, partimos de una nube de puntos en una distribución bidimensional y queremos determinar la recta que mejor se ajuste a esta nube de puntos.



7.2.1 Ajuste por mínimos cuadrados

Sean X e Y dos variables cuya distribución conjunta de frecuencias es $(x_i, y_j; n_{ij})$. Para determinar las rectas de regresión procedemos por el método de mínimos cuadrados.

El propósito es hallar la recta que mejor exprese el comportamiento de Y en función de X , de tal forma que se consiga predecir con el mínimo error los valores en Y para valores dados en X. La recta tendrá una expresión del tipo $Y = a + bX$, donde es preciso encontrar los valores de a y b.

De esta forma hay:

- Un valor observado o real, que denomino y_j
- Un valor teórico y_j^* que es el obtenido al sustituir x_i en la recta de regresión.

Así, cada punto tiene una distancia a la recta, y esta distancia es el error de cada punto. El error cometido en la predicción en este caso vendría dado por la siguiente expresión:

$$e_j = y_j - y_j^* = y_j - ax_i - b$$

El método de mínimos cuadrados busca determinar los parámetros a y b, de forma que los errores sean mínimos. Los valores de a y b que minimizan el error son:

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

Así la recta de regresión de Y sobre X es: $y = \bar{y} + \frac{S_{xy}}{S_x^2}(x - \bar{x})$

En el caso de que sea la regresión de X/Y, (Y explicativa y X a explicar), se formularía de manera análoga obteniendo la siguiente expresión de la recta

$$x = \bar{x} + \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

Para medir el ajuste y por lo tanto la fiabilidad de las estimaciones, se puede hacer uso del coeficiente de correlación r. Cuanto más se aproxime éste a 1 o -1, mayor será la asociación lineal entre las variables, el ajuste de la recta a la nube de puntos será mejor, y las predicciones que se realicen a través de la recta de regresión serán más fiables.

Unidad 8. Tasas de variación. Números índices simples. Índices complejos. Índices de precios y de cantidad.

Es frecuente cuando se trata con datos que surja la necesidad de comparar los valores de la variable entre dos periodos de tiempo (exportaciones de naranjas en mayo de 2016 y mayo de 2019) o dos referentes de espacio (exportaciones de naranjas de Marruecos y de España). Asimismo, también puede ser de interés el análisis a lo largo de una serie temporal o entre una serie de regiones.

En lo que sigue nos vamos a centrar en definir conceptos que nos permiten llevar a cabo este análisis de los valores que una variable toma en distintos puntos del tiempo o espacio.

8.1 Tasas de variación

Siempre que estudiamos algún dato estadístico, resulta interesante hacer una comparativa para que sea más rica la información. El dato de población de España a 1 de Enero de un año cualquiera, por sí mismo no aporta gran información sobre la evolución del tamaño y composición de la misma. Por ello, lo habitual es comparar este dato con el del año previo para medir los posibles incrementos o disminuciones.

Así para llevar a cabo un análisis de un dato de la variable, necesitamos compararlo con otros datos anteriores y a veces no solo con respecto al valor inmediatamente anterior. Para hacer estas comparaciones se usa lo que conocemos con el nombre de **tasas de variación** que proporcionan una medida de la evolución de una magnitud entre dos instantes temporales.

Definición: Se define tasa de variación de una variable X en el punto t respecto al punto $t-1$ como

$$T_t^{t-1} = \frac{X_t - X_{t-1}}{X_{t-1}} = \frac{X_t}{X_{t-1}} - 1$$

Esta tasa representa la variación en tanto por uno desde t-1 a t de la variable. Si multiplicamos por 100 expresaremos la variación en términos porcentuales:

$$T_t^{t-1} = \frac{X_t - X_{t-1}}{X_{t-1}} * 100 = \left(\frac{X_t}{X_{t-1}} - 1\right) * 100$$

***Observación**: no debemos confundir el concepto de tasa con el de tasa de variación. El concepto de tasa se define como la relación entre dos magnitudes aunque medidas en el mismo momento del tiempo. Por ejemplo, podemos hablar de tasa de desempleo como la relación entre desempleados y población activa (ambas medidas en el mismo instante del tiempo). En cambio la tasa de variación es la relación de la misma magnitud pero entre dos momentos diferentes del tiempo.*

La ventaja de las tasas de variación es que son adimensionales, por lo que se puede comparar la evolución de dos variables con magnitudes diferentes. Por ejemplo, si queremos medir el cambio de población entre 2017 y 2018 de dos países como India y Arabia Saudí, tendríamos los datos siguientes:

	2018	2017	Var.absoluta	Tasa
India	1.352.617.328	1.338.658.835	13.958.493	1,04%
Arabia Saudí	33.699.947	32.552.000	1.147.947	3,52%

Se aprecia que la población de India aumentó en un año casi 14 millones de personas mientras que la población de Arabia Saudí aumentó en poco más de un millón. Pero calculando la tasa de variación, por cada 100 personas de India, creció en un habitante, mientras que por cada 100 residentes en Arabia Saudí hubo 3 habitantes nuevos en 2018 respecto al año anterior. Es decir en términos relativos, el incremento de población en arabia Saudí triplicó al de India.

1. Tipos de tasas de variación

El concepto de tasa de variación se puede generalizar considerando la variación de un punto t respecto a otro punto t', no teniendo que ser entre dos puntos consecutivos, como:

$$T_t^{t'} = \frac{X_t - X_{t'}}{X_{t'}} = \frac{X_t}{X_{t'}} - 1$$

A partir de esta definición general, vamos a definir distintos conceptos de tasas de variación según:

- el tiempo transcurrido entre dos puntos
- la acumulación de la magnitud

1. Tasa de variación intermensual (o mensual)

Cuando se comparan los valores de una variable entre un mes y el siguiente.

2. Tasa de variación intertrimestral (trimestral)

Cuando se disponen de los datos en periodos trimestrales, la evolución de la variable se mide entre dos trimestres consecutivos y la tasa se denomina intertrimestral.

$$T_{trim}^{trim-1} = \frac{X_{trim} - X_{trim-1}}{X_{trim-1}} = \frac{X_{trim}}{X_{trim-1}} - 1$$

3. Tasa de variación interanual

Cuando se mide la tasa de variación entre dos momentos iguales de dos años consecutivos.

4. Tasas de datos acumulados.

En ciertos casos querremos comparar la evolución del acumulado de varios periodos dentro de un año respecto al año anterior. Con esta medida se suavizan posibles datos anómalos o de difícil comparabilidad entre varios años. Por ejemplo, el número de viajeros transportados puede depender de los fines de semana, por lo que si en un mes tiene 5 fines de semana en un año y el mismo mes del año anterior dispone de solo 4, estos meses no son comparables, con lo que la tasa de variación interanual no proporciona una información fiable. O por ejemplo, 2012 fue bisiesto, con lo que el mes de febrero tiene un día más y obviamente habrá más viajeros que en febrero de 2011. Para conseguir ganar en comparabilidad se puede ir agrupando los meses de un año y comparándolos con los mismos del año anterior, de esta manera disponemos de series de datos más "suavizadas" ya que las anomalías de un mes se compensan con las de otros, o se "diluyen" en la suma de varios meses.

Tasa interanual acumulada hasta el mes k

$$T_{\text{mes k año t}}^{\text{mes k año t-1}} = \frac{\sum_{i=1}^{k} X_{\text{mes i año t}} - \sum_{i=1}^{k} X_{\text{mes i año t-1}}}{\sum_{i=1}^{k} X_{\text{mes i año t-1}}} = \frac{\sum_{i=1}^{k} X_{\text{mes i año t}}}{\sum_{i=1}^{k} X_{\text{mes i año t-1}}} - 1$$

Ejercicio 20.1: El número de vehículos matriculados en España durante los años 2017 y 2018 son:

Año	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
2017	100.716	114.344	147.472	118.651	147.608	153.237	130.230	86.908	100.861	115.262	125.635	121.306
2018	120.327	129.990	149.639	134.895	158.494	165.741	152.457	124.802	85.008	109.337	111.256	118.886

Se pide:

a) Calcular la tasa interanual acumulada del primer semestre

b) Calcular la tasa interanual acumulada

Solución 20.1: Para calcular las tasas acumuladas primero se obtienen los datos acumulados de vehículos matriculados en ambos períodos (frecuencia acumulada):

	Año 2017	Año 2018	Acumulado Año 2017	Acumulado Año 2018
Enero	100.716	120.327	100.716	120.327
Febrero	114.344	129.990	215.060	250.317
Marzo	147.472	149.639	362.532	399.956
Abril	118.651	134.895	481.183	534.851
Mayo	147.608	158.494	628.791	693.345
Junio	153.237	165.741	782.028	859.086
Julio	130.230	152.457	912.258	1.011.543
Agosto	86.908	124.802	999.166	1.136.345
Septiembre	100.861	85.008	1.100.027	1.221.353
Octubre	115.262	109.337	1.215.289	1.330.690
Noviembre	125.635	111.256	1.340.924	1.441.946
Diciembre	121.306	118.886	1.462.230	1.560.832

Finalmente se van calculando las tasas de variación de los acumulados:

La tasa interanual acumulada del primer semestre sería comparar los datos acumulados de los 6 primeros meses del año:

$$T_{06/2017}^{06/2018} = \frac{859.086 - 782.028}{782.028} * 100 = 9.85\%$$

La tasa interanual acumulada (es decir, sin explicitar meses que se acumulan) se referiría a la tasa del acumulado de todo el año:

$$T_{12/2017}^{12/2018} = \frac{1.560.832 - 1.462.230}{1.462.230} * 100 = 6.74\%$$

8.2 Números índices

Los índices son una expresión cuantitativa que permite comparar una variable a lo largo del tiempo o el espacio, es decir, estudiar la variación relativa.

Para su construcción se procede dividiendo una magnitud en un momento cualquiera del tiempo entre el valor de la variable en el momento o período inicial que llamaremos periodo base o de referencia. Generalmente el momento base, como ocurre con los indicadores más conocidos: IPC, IPI, IPRI,..., es un año. Observemos que los índices son adimensionales, no tienen unidades de medida.

Según su composición y forma de construcción, se puede hablar de números índices simples y números índices complejos. Desarrollemos todas estas ideas a continuación.

8.2.1 Números índices simples

Los **índices simples** son una expresión matemática que proporciona la variación sufrida por una magnitud entre dos períodos. Lo que les caracteriza es que no se pueden desagregar más aquellos componentes que integran su cálculo.

Definición: Consideramos X_i una magnitud, y sean X_i^0 y X_i^t los valores que toman en los períodos base y actual. Definimos el índice simple como:

$$I_0^t(i) = \frac{X_i^t}{X_i^0}$$

Habitualmente el índice se expresa en tanto por ciento, esto es

$$I_0^t(i) = \frac{X_i^t}{X_i^0} * 100$$

Vamos a mencionar algunas de las **propiedades** de los índices simples:

- **Identidad**: si los períodos base y actual son coincidentes, el valor del índice es la unidad (o 100, si está multiplicado por 100)

$$I_0^0(i) = \frac{X_i^0}{X_i^0} = 1$$

- **Reversibilidad**: Si se intercambian los períodos base y actual, los índices deben ser los valores recíprocos.

$$I_0^t(i) = \frac{1}{I_t^0(i)}$$

- **Homogeneidad**: no queda afectado por un cambio en las unidades de medida.
- **Proporcionalidad**: si en el período actual sufre una variación proporcional k, el número índice debe quedar afectado por k.
- **Circularidad:** El producto de índices que tienen sucesivamente como base la situación objeto del anterior, debe ser igual al índice con base la del primero y objeto la del último.

$$I_0^t(i) = I_{t-1}^t(i) * I_{t-2}^{t-1}(i) * \ldots * I_0^1(i)$$

8.2.2 Números índices complejos

En la mayoría de situaciones, el fenómeno cuya evolución se quiere medir requiere la integración de varios índices simples, porque no sólo estamos interesados en comparar el precio o la cantidad de un bien individual, sino comparar dichas magnitudes para un conjunto de ellos.

Los **índices complejos**, o índices agregados, se calculan como agregaciones de índices simples. Para obtener dicha agregación se utilizan fórmulas variadas (generalmente medias simples o medias ponderadas)

Un índice complejo tiene la siguiente expresión matemática:

$$I_t = f(I_{it}, W_i)$$

Se trata, pues, de una función de índices simples en el momento t y sus respectivas ponderaciones W_i (éstas pueden estar referidas o no al mismo momento t). Se dice que son índices complejos no ponderados cuando la ponderación es la misma para todos ellos

Entre los **índices complejos no ponderados** se pueden considerar varios tipos según sea la fórmula general utilizada para agregar:

- Medias aritméticas de índices simples.

$$I_t = \frac{I_0^t(1) + I_0^t(2) + \cdots + I_0^t(N)}{N} = \frac{\sum_{i=1}^{N} I_0^t(i)}{N}$$

Donde I_t es el índice agregado en el momento t, I_{it} es el índice simple del elemento i en el momento t y N es el número de elementos que se agregan.

- Medias geométricas de índices simples.

$$I_G = \sqrt[N]{I_0^t(1) * I_0^t(2) *\ldots* I_0^t(N)} = \sqrt[N]{\prod_{i=1}^{N} I_0^t(i)}$$

Cuando se asignan ponderaciones a cada elemento y cada una de ellas diferentes las **fórmulas** toman las siguientes formas:

- Medias aritméticas ponderadas.

$$I^* = \frac{I_0^t(1)w_1 + I_0^t(2)w_2 + \cdots + I_0^t(i)w_i + \cdots + I_0^t(N)w_N}{w_1 + w_2 + \cdots + w_i + \cdots + w_N}$$

Donde I^* es el índice agregado, $I_0^t(I)$ es el índice simple del elemento i en el momento t y w_i es la ponderación del elemento i

8.3 Índices de precios y de cantidad

Según la variable objeto de estudio, se distingue entre tres tipos de índices

- **Índices de precios**: miden la evolución en los precios, ya sea de uno o varios bienes.
- **Índices de cantidad:** miden la evolución del consumo de uno o varios bienes o de la producción.
- **Índices de valor:** miden la evolución del valor de uno o varios bienes.

1. Índices Simples:

- **Índices de precios**: definido como la razón entre el precio de un bien en el periodo actual, y el precio del mismo en el periodo base.

$$I_0^t(i) = \frac{p_i^t}{p_i^0}$$

- **Índices de cantidad:** definido como la razón entre la cantidad consumida (o producida) de cierto producto en el periodo actual y la cantidad consumida (o producida) del mismo en el periodo base.

$$I_0^t(i) = \frac{q_i^t}{q_i^0}$$

- **Índices de valor:** es la razón entre el valor del bien en el periodo actual y el valor del mismo en el periodo base.

$$I_0^t(i) = \frac{v_i^t}{v_i^0}$$

2. Índices Complejos ponderados:

Existen diversas fórmulas para calcular los índices de precios y cantidad complejos ponderados.

2.1 Precios

Dados dos periodos, t y 0 (periodo base), los precios P_i en ambas situaciones y las cantidades Q_i que determinan la estructura de consumo dada, el índice agregativo será:

$$\sum_i I_i W_i \quad donde \ W_i = \frac{Q_i P_{i0}}{\sum Q_i P_{i0}}$$

Es decir: $I_a = \frac{\sum I_i Q_i P_{io}}{\sum Q_i P_{i0}}$

Que también puede expresarse como: $I_a = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i}$

lo que nos permite definir el índice agregativo como la relación de los costes de una determinada estructura de consumo a precios de la situación objeto respecto de la situación base.

La diferente definición del conjunto $\{ Q_i \}$ (conjunto de ponderaciones) da lugar a las numerosas fórmulas existentes en la literatura de índices. A destacar los siguientes

- Fórmula de LASPEYRES:media agregativa de precios ponderados por las cantidades del periodo base $I_L = \frac{\sum P_{it} Q_{i0}}{\sum P_{i0} Q_{i0}}$

- Fórmula de PAASCHE: media agregativa de precios ponderados por las cantidades del periodo corriente. $I_P = \frac{\sum P_{it} Q_{it}}{\sum P_{i0} Q_{it}}$

2.2 Cantidades:

Realizando un razonamiento análogo que el utilizado para determinar las distintas expresiones de índices complejos de cantidad ponderados, obtenemos los de cantidad. Algunos de los más conocidos son:

- Fórmula de LASPEYRES:media agregativa de cantidades ponderados por los precios del periodo base $I_L = \frac{\sum Q_{it} P_{i0}}{\sum Q_{i0} P_{i0}}$

- Fórmula de PAASCHE: media agregativa de cantidades ponderados por los precios del periodo corriente. $I_P = \frac{\sum Q_{it} P_{it}}{\sum Q_{i0} P_{it}}$

Anexo ejercicios

Ejercicio 1. Se ha entrevistado a 300 familias y se ha recogido entre otras características el número de unidades adquiridas de cierto producto y obteniendo la siguiente distribución:

$L_{i-1} - L_i$	n_i
(10,20]	25
(20,30]	36
(30,40]	42
(40,50]	18
(50,60]	27
(60,70]	43
(70,80]	31
(80,90]	37
(90,100]	41

Se pide:

a) Hallar las frecuencias acumuladas, frecuencias relativas y frecuencias relativas acumuladas de la distribución anterior.

b) Obtener la media de adquisiciones de este producto por familia, la mediana y la moda.

Solución 1:

a) En primer lugar, obtenemos la tabla de distribución de frecuencias

$(L_{i-1} - L_i]$	x_i	n_i	$x_i n_i$	N_i	f_i	F_i
(10,20]	15	25	375	25	0,08	0,08
(20,30]	25	36	900	61	0,12	0,20
(30,40]	35	42	1470	103	0,14	0,34
(40,50]	45	18	810	121	0,06	0,40
(50,60]	55	27	1485	148	0,09	0,49
(60,70]	65	43	2795	191	0,14	0,64

(70,80]	75	31	2325	222	0,10	0,74
(80,90]	85	37	3145	259	0,12	0,86
(90,100]	90	41	3690	300	0,14	1,00
		N=300	$\sum x_i n_i$ =16995		1,00	

Nota: x_i es la marca de clase de los intervalos

b)

- La media aritmética la obtenemos de la siguiente forma:

  $\bar{x} = \frac{1}{N}\sum x_i n_i = \frac{1}{300}$*16995=56,65

- Para obtener la mediana

  - $\frac{N}{2} = \frac{300}{2} = 150$

  - Se busca la primera frecuencia absoluta acumulada que supere a $\frac{N}{2}$, que este caso es $N_6 = 191$, luego la mediana pertenece al intervalo (60,70].

    De acuerdo con la fórmula:

    Me=60+((150-148)/43)*10=60,47

- Se pide la moda que es el valor más frecuente de la variable. Se observa que el intervalo modal es el (30,40]

Ejercicio 2. En un equipo de fútbol infantil la altura (X) y peso (Y) de 20 componentes del equipo son los siguientes, donde la variable X viene expresa en metros y la Y en kilogramos:

X/Y	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]
(1.65,1.70]	1	2	4		
(1.70,1.75]	1	4	2	1	1
(1.75,1.80]		1	3		
Total	2	7	9	1	1

Obtener la media y la desviación de las distribuciones marginales.

Solución 2:

Partimos de la tabla de contingencia

X/Y	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	Total
(1.65,1.70]	1	2	4			7
(1.70,1.75]	1	4	2	1	1	9
(1.75,1.80]		1	3			4
Total	2	7	9	1	1	20

La distribución marginal de Y, sería la que sigue a continuación que contiene además todos los cálculos necesarios para obtener la media y la desviación:

$(L_{i-1}, L_i]$	n_i	y_i	$y_i n_i$	$y_i^2 . n_i$
(1.65,1.70]	7	1,675	11,725	19,639375
(1.70,1.75]	9	1,725	15,525	26,780625
(1.75,1.80]	4	1,775	7,1	12,6025
	20		34,35	59,0225

- La media $\bar{y} = \frac{1}{N}\sum y_i n_i = \frac{34{,}35}{20} = 1{,}7175$
- La varianza $S_y^2 = \frac{1}{N}\sum {y_i}^2 n_i - \bar{y}^2 = \frac{59{,}023}{20} - (\frac{34{,}350}{20})^2 = 0{,}0013$
- La desviación típica $S_y = \sqrt{0{,}0013} = 0{,}037$

La distribución marginal de X, sería la que sigue a continuación que contiene además todos los cálculos necesarios para obtener la media y la desviación:

$(L_{i-1}, L_i]$	n_i	x_i	$x_i n_i$	$x_i^2 . n_i$
(60,65]	2	62,5	125	7812,5
(65,70]	7	67,5	472,5	31893,75

(70,75]	9	72,5	652,5	47306,25
(75,80]	1	77,5	77,5	6006,25
(80,85]	1	82,5	82,5	6806,25
	20		1410	99825

- La media $\bar{x} = \frac{1}{N}\sum x_i n_i = \frac{1410}{20} = 70{,}50$
- La varianza $S_x^2 = \frac{1}{N}\sum {x_i}^2 n_i - \bar{x}^2 = \frac{99.825}{20} - (\frac{1410}{20})^2 = 21$
- La desviación típica $S_x = \sqrt{21} = 4{,}58$

Ejercicio 3. En una encuesta se han recogido los datos relativos a los ingresos de las familias en un año en relación con el tamaño del hogar en el que residen. Sea la variable X : "número de personas que residen en el hogar" y la variable Y: "ingresos en el hogar en miles de euros":

X/Y	(0,20]	(20,40]	(40,60]	(60,80]
1	100	130	.	.
2	180	110	.	.
3	170	120	60	30
4	140	120	80	40
5	10	20	60	30

a) Obtenga un modelo lineal que permita predecir los ingresos de las familias a partir del tamaño del hogar.

b) Juzgue la bondad del modelo estimado y de la predicción realizada.

Solución 3:

a) Para obtener el modelo lineal que permita explicar los ingresos anuales de las familias (variable Y) en función del tamaño del hogar (variable X), recurrimos al ajuste por mínimos cuadrados de la recta de regresión:

$$Y = a + bX$$

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

Transformamos la tabla de doble entrada, sustituyendo los intervalos por marcas de clase y obtenernos en la tabla todas las operaciones que nos van a ser de utilidad

x_i	y_i	n_i	$m.c(x_i)$	$y_i n_i$	$x_i n_i$	$x_i y_i n_i$	$y_i^2 n_i$	$x_i^2 n_i$
1	0-20	100	10	1000	100	1000	10000	100
1	20-40	130	30	3900	130	3900	117000	130
2	0-20	180	10	1800	360	3600	18000	720
2	20-40	110	30	3300	220	6600	99000	440
3	0-20	170	10	1700	510	5100	17000	1530
3	20-40	120	30	3600	360	10800	108000	1080
3	40-60	60	50	3000	180	9000	150000	540
3	60-80	30	70	2100	90	6300	147000	270
4	0-20	140	10	1400	560	5600	14000	2240
4	20-40	120	30	3600	480	14400	108000	1920
4	40-60	80	50	4000	320	16000	200000	1280
4	60-80	40	70	2800	160	11200	196000	640
5	0-20	10	10	100	50	500	1000	250
5	20-40	20	30	600	100	3000	18000	500
5	40-60	60	50	3000	300	15000	150000	1500
5	60-80	30	70	2100	150	10500	147000	750
		1400		38000	4070	122500	1500000	13890

$$\bar{y} = \frac{\sum_{i=1}^{k} y_i * n_i}{N} = \frac{38.000}{1.400} = 27{,}14$$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i * n_i}{N} = \frac{4.070}{1.400} = 2{,}91$$

$$S_y^2 = \frac{\sum_{i=1}^{k} y_i^2 \, n_i}{N} - \bar{y}^2 = \frac{1.500.000}{1.400} - \left(\frac{38.000}{1.400}\right)^2 = 334{,}70$$

$$S_x^2 = \frac{\sum_{i=1}^{k} x_i^2 \, n_i}{N} - \bar{x}^2 = \frac{13.890}{1.400} - \left(\frac{4.070}{1.400}\right)^2 = 1{,}47$$

$$S_{xy} = \frac{\sum_{i=1}^{k} x_i * y_i * n_i}{N} - \bar{x} * \bar{y} = \frac{122.500}{1.400} - 27{,}14 * 2{,}91 = 8{,}526$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{8{,}526}{1{,}47} = 5{,}8$$

$$a = \bar{y} - b\bar{x} = 27{,}14 - 5{,}8 * 2{,}91 = 10{,}26$$

La recta de regresión es:

$$y = 10.26 + 5.8x$$

b) Obtenemos el coeficiente de correlación

$$r = \frac{S_{xy}}{S_x S_y} = \frac{8.526}{1.21 * 18.29} = 0.384$$

Concluimos que el ajuste realizado explica un 38% de la varianza de la variable dependiente.

Ejercicio 4. Las cifras de ventas de una compañía de coches en millones de euros es la siguiente.

Año	Ventas
2000	12
2001	14
2002	15
2003	16
2004	18
2005	20

i) Hallar los índices de ventas tomando como base el año 2000

ii) Hallar los índices de ventas tomando como base el año 2003

Solución 4:

El año 2000 se toma como base del índice 100, para calcular los otros años hay que dividir las ventas de un año entre las del año base y multiplicar por 100.

- $I_{2000}^{2000} = \frac{12}{12} * 100 = 100$
- $I_{2000}^{2001} = \frac{14}{12} * 100 = 116{,}6$
- $I_{2000}^{2002} = \frac{15}{12} * 100 = 125$
- $I_{2000}^{2003} = \frac{16}{12} * 100 = 133{,}33$
- $I_{2000}^{2004} \frac{18}{12} * 100 = 150$
- $I_{2000}^{2005} = \frac{20}{12} * 100 = 166{,}67$

Año	Ventas	$I(2000 = 100)$
2000	12	100,00
2001	14	116,67
2002	15	125,00
2003	16	133,33
2004	18	150,00
2005	20	166,67

ii. En este caso el año 2003 se toma como base del índice 100, para calcular los otros años hay que dividir las ventas de un año entre las del año base y multiplicar por 100

- $I_{2003}^{2000} = \frac{12}{16} * 100 = 75$
- $I_{2003}^{2001} = \frac{14}{16} * 100 = 87{,}50$
- $I_{2003}^{2002} = \frac{15}{16} * 100 = 93{,}75$
- $I_{2003}^{2003} = \frac{16}{16} * 100 = 100$
- $I_{2003}^{2004} = \frac{18}{16} * 100 = 112{,}5$
- $I_{2003}^{2005} = \frac{20}{16} * 100 = 125$

Año	Ventas	$I(2003 = 100)$
2000	12	75,00
2001	14	87,50
2002	15	93,75
2003	16	100,00
2004	18	112,50
2005	20	125,00