ANÁLISIS EXPLORATORIO DE DATOS

1. INTRODUCCIÓN HISTÓRICA

1.1 La Estadística como ciencia

La Estadística actual es el resultado de la unión de dos disciplinas que evolucionaron de manera independiente hasta confluir en el siglo XIX, que son:

- Cálculo de probabilidades (nace en el s. XVII)
- Estadística (o ciencia del Estado, del latín Status) que estudia la descripción de los datos.

En el diccionario de la lengua aparece la palabra Estadística bajo el significado de "censo o recuento de la población, de la producción, del tráfico o de cualquier otra entidad colectiva".

Pero la Estadística como ciencia es mucho más amplia que las estadísticas que aparecen publicadas en los Medios de Comunicación. Es la ciencia encargada de recoger, analizar e interpretar los datos numéricos relativos a un conjunto de elementos, y como ciencia aplicada se ocupa del estudio de los métodos y procedimientos para efectuar esa recogida, clasificación y resumen de los datos.

1.2 Algunos problemas que resuelve la Estadística

Descripción de datos

Es el primer problema que aborda. Se trata de encontrar procedimientos para resumir la información obtenida de los datos.

Análisis de Muestras

En numerosas ocasiones no es posible estudiar todos los elementos de una población (ya sea por razones técnicas o económicas), por lo que se toma una muestra.

Contrastación de Hipótesis

Un objetivo frecuente en la investigación empírica es contrastar una hipótesis. Por ejemplo: ¿Es una nueva medicina eficaz para un catarro?.

La contrastación de hipótesis requiere una metodología para comparar las predicciones resultantes de la hipótesis con los datos observados y el diseño de experimentos para garantizar que las conclusiones que se extraigan de la experimentación no estén invalidadas por factores no controlados.

Predicción

Muchas variables, sobre todo económicas, tienen cierta inercia en su evolución y aunque son valores desconocidos, el estudio de su historia es informativo para prever su evolución futura.

2. INTRODUCCIÓN A LA ESTADÍSTICA

2.1. Concepto y Objetivo de la Estadística

La Estadística engloba por un lado las ciencias abstractas (utilizan el método deductivo, y mediante una serie de axiomas y proposiciones obtienen resultados) y también las ciencias empíricas (que son aquellas que obtienen resultados observando y experimentando todo aquellos que quieren estudiar).

La *Estadística* es la ciencia que estudia como debe emplearse la información que se tiene o se puede obtener y como dar una guía de acción en situaciones prácticas que entrañan incertidumbre. Su objetivo principal es obtener información, analizarla, examinarla y predecir.

2.1.1. Campo de Actuación

Hay dos tipos de fenómenos o experimentos que se pueden observar y estudiar, pero la Estadística no se ocupa del estudio de ambos, sólo de uno de ellos, que son los *fenómenos aleatorios o de azar.*

Los dos tipos de fenómenos son:

- Fenómenos causales o determinísticos.- Son los que están sometidos a leyes, repetidos en las mismas condiciones nos dan los mismos resultados. (Así, el tirar un objeto desde un cierta altura muchas veces, con la misma masa, tarda siempre el mismo tiempo).
- *Fenómenos aleatorios o de azar*. No están sometidos a leyes. Se caracterizan por la imposibilidad de conocer los resultados cuando se repiten en iguales condiciones. Sólo podemos predecir lo que sucederá después de repetir el fenómeno o experimento varias veces). (Ejemplo, lanzar una moneda o un dado al aire, sacar una carta de una baraja …).

2.1.2. Forma de Actuación

1. Causal: Analiza en cada caso los efectos, buscando las causas que producen dichos efectos.
2. Estadístico: Estudia los fenómenos aleatorios, analizando conjunto de individuos.

Las Ciencias Estadísticas

Podemos clasificar la Estadística en dos grandes bloques, que a su vez tienen más subdivisiones.

1. **Estadística Descriptiva:**
    - Se utiliza cuando los resultados del análisis estadístico no pretende ir más allá del conjunto de datos investigados.
    - Describe numéricamente, analiza y representa un conjunto de datos ordenados mediante la utilización de métodos numéricos, tablas y gráficas, simplificando y resumiendo la información.

2. **Estadística Inferencial:**
    - Se utiliza para predecir datos futuros a partir de los valores observados, pudiendo hacer una ley aproximada de lo que ocurra en el futuro.
    - Se apoya en el cálculo de probabilidades y a partir de unos datos, efectúa estimaciones, decisiones, predicciones y otras generalizaciones sobre un conjunto mayor de datos.

En resumen, una clase de Estadística se refiere a lo que ocurre y otra a lo que ocurrirá. Entre ambas hay un vacío que lo llena el concepto de probabilidad, el cual permite pasar de la estadística descriptiva a la inferencial, incorporando la medida de probabilidad.

Estudio Estadístico

Antes de hacer un estudio estadístico, tenemos que plantearnos bien que problema vamos a estudiar, y cuáles van a ser los objetivos de nuestra investigación, fijando los pasos a seguir, las clasificaciones que se van a realizar, las variables que debemos observar y cómo medirlas, los gráficos que vamos a representar …

Para la elaboración de los datos, hay que realizar una buena recogida de la información muestral, pues una mala recogida, puede anular el estudio. Tenemos dos procedimientos:

a) Muestreo. Consiste en elegir al azar una muestra mediante diferentes tipos de muestreo que estudiaremos más adelante y observar pasivamente esa muestra anotando los valores.
b) Diseño de experimentos. Consiste en fijar los valores de ciertas variables y observar las respuestas de otras, es decir, se plantean los objetivos que se pretenden delimitando la información que se va a estudiar.

Una vez que hemos recogido los datos, tenemos que:

1. Hacer una recopilación y reducción de dichos datos a unas pocas medidas representativas
2. Confeccionar tablas acompañadas de gráficos para una mejor visión de los datos
3. Interpretar los resultados y obtener conclusiones para predecir y tomar decisiones estadísticas.

3. Generalidades sobre la Estadística Descriptiva

3.1. Introducción

La Estadística Descriptiva, pretende dar una descripción numérica, ordenada y simplificada, a veces con la ayuda de representaciones gráficas, de la información obtenida en la recogida de datos de un fenómeno aleatorio.

Ya vimos qué era un fenómeno aleatorio (aquel que no está sujeto a leyes. Se caracteriza por la imposibilidad de conocer los resultados cuando se repiten en iguales condiciones. Solo podemos predecir lo que sucederá después de repetir el fenómeno o experimentos varias veces. Todos los fenómenos aleatorios pueden ser descritos estadísticamente, por ejemplo decimos que al lanzar una moneda tenemos las mismas posibilidades de sacar cara ,C, que de sacar cruz, X).

3.2. Terminología Estadística

Ya dijimos que el análisis estadístico, está formado por dos elementos fundamentales como ya hemos analizado:

- Descripción del conjunto de información
- Obtención de conclusiones de toda la población cuando sólo conocemos parte de ella, y predicción de consecuencias futuras.

Términos más usuales:

- **Población (colectivo o universo).** Conjunto de unidades, elementos o individuos sobre los que se realiza el estudio, y que cumplen una determinada característica o propiedad.

  A cada elemento de la población se le llama *individuos o unidades estadísticas.*

  El *tamaño* de la población es el número de individuos que tiene dicha población, y lo denotamos por **N**.

  Hay dos tipos de poblaciones: Finita e Infinita.

La población que se vaya a estudiar debe definirse con mucha precisión, por ejemplo, si queremos hacer un estudio de los hábitos de estudio de los estudiantes de una Universidad, debemos de saber a qué tipo de estudiantes nos referimos, si sólo de los que asisten a clase o también de los que no asisten…

- **Muestra:** es cualquier subconjunto de la población. La muestra es una representación de la población, por ello es importante su elección. El proceso mediante el cual se extrae una muestra se llama *muestreo*.
  Uno de los tipos de muestreo más utilizado es el *muestreo aleatorio simple* (m.a.s.) en el que cada individuo de la población tiene la misma probabilidad de ser incluido en la muestra.
  Existen varios motivos por las que se elige una muestra, una de ellas es el coste que supone hacer un estudio de una población entera … Pero cuidado! No siempre es necesario tomar una muestra, ya que si queremos estudiar el fracaso escolar de un curso determinado, deberemos tomar todos los alumnos de dicho curso, y no una muestra de ellos.

- **Caracteres estadísticos:** es una propiedad que permite clasificar a los individuos de una población. Se distinguen dos tipos:
  a) *Cualitativos.* Son aquellos cuya variación se recoge por la presentación de distintas cualidades, es decir, los que no se pueden medir.
  Ejemplo: estado civil, color de ojos, sexo, profesión de una persona, carrera que piensa elegir un alumno.
  Las modalidades son las diferentes situaciones de un carácter, por ejemplo, las modalidades del carácter profesión podrían ser: economista, psicólogo, informático, periodista …
  b) *Cuantitativos.* Son aquellos que se pueden medir o contar y están formadas por cantidades numéricas.
  Ejemplo: talla y peso de un individuo, número de acciones en Bolsa, número de alumnos matriculados en una universidad…

- **Variables estadísticas.** Cuando hablemos de variable, haremos referencia a un símbolo (X, Y, A, B..) que puede tomar cualquier modalidad (o valor) de un conjunto determinado, que llamaremos **dominio** de la variable o **rango**.
  En función del tipo de dominio, clasificamos las variables en:

  a) Variables cualitativas: cuando las modalidades posibles son de tipo nominal. Por ejemplo una variable de color, A $\in$ {“rojo”, “azul”, “verde”}
  b) Variables cuantitativas ordinales: son las que, aunque sus modalidades son de tipo nominal, es posible establecer un orden entre ellas. Por ejemplo, si estudiamos la llegada a la meta de un corredor en una competición de 20 participantes, su clasificación *C* es tal que $C \in \{1º, 2º, 3º, \ldots, 20º\}$.
  Otro ejemplo de variable cuantitativa ordinal es el nivel de dolor, *D*, que sufre un paciente ante un tratamiento médico: $D \in$ {“inexistente”, “poco intenso”, “moderado”, “fuerte”}.

c) Variables cuantitativas: son las que tienen por modalidades cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos grupos:

* Discretas: cuando no admiten siempre una modalidad intermedia entre dos cualesquiera de sus modalidades. Un ejemplo es el número de caras , obtenido en el lanzamiento repetido de una moneda. Es obvio que cada valor de la variable es un número natural.
* Continuas: cuando admiten una modalidad intermedia entre dos cualesquiera de sus modalidades, por ejemplo, el peso *X* de un niño al nacer. En este caso los valores de las variables son número reales, es decir, $X \in \mathbf{R}$.

Ocurre a veces, que una variable cuantitativa continua por naturaleza, aparece como discreta. Este es el caso en que hay limitaciones en los que concierne a la precisión del aparato de medida de esa variable, por ejemplo, si medimos la altura en metros de personas con una regla que ofrece dos decimales de precisión, podemos obtener $C \in$ {…, 1.50 , 1.51, 1.52, 1.53, …}. En realidad lo que ocurre es que con cada una de esas mediciones expresamos que el verdadero valor de la misma se encuentra en un intervalo de radio 0.005.

3.3. Distribuciones de Frecuencias

La forma de la distribución de los datos (de una variable) se denomina *distribución de frecuencias*.

El estudio de las distribuciones de frecuencias tiene por objeto la construcción de tablas de frecuencias que podrán utilizarse para una mejor presentación e interpretación de la información contenida en los datos observados en la muestra. En este apartado, nos referimos a las distribuciones unidimensionales de frecuencias, que son aquellas utilizadas para describir una variable individual sin tener en cuenta la información de otras variables que pudieran haberse incluido en el estudio.

Para poder obtener la forma general de una distribución de frecuencias unidimensional, es necesario introducir algunos conceptos previos.

Consideremos una población estadística de N individuos, descrita según una variable o carácter X, cuyas modalidades han sido agrupadas en un número n de clases, denotándolo como $x_1, x_2, \ldots, x_n$. Para cada una de estas clases x_i, i=1, . . ., n, vamos a definir:

- *Frecuencia absoluta de la clase x_i:* Es el número f_i de observaciones que existen en dicha clase. Dicho de otra forma, es el número de veces que se repite dicho valor. Se denota mediante f_i.
- *Frecuencia absoluta acumulada de la clase x_i:* Es el número de elementos de la población cuya modalidad es inferior o equivalente a las de la clase x_i. Se denota por F_i.

Además se cumple que: $F_i = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j$

- *Frecuencia relativa de la clase x_i:* Es el cociente entre las frecuencias absolutas de dicha clase y el número total de observaciones o datos que denotamos por N, es decir: $h_i = \frac{f_i}{N}$, *donde* $N = f_1 + f_2 + \ldots + f_n = \sum_{i=1}^{n} f_i = F_n$

  Si estamos interesados en trabajar con porcentajes, sólo tenemos que multiplicar la frecuencia relativa por 100 y así representamos el porcentaje (%) de la población que comprende a esa clase.

- *Frecuencia relativa acumulada de la clase x_i:* es el número de elementos de la población que están en alguna de las clases inferior o igual a la case x_i. Se denota por H_i. Resulta:

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

3.3.1. Tablas Estadísticas

Cuando se realiza un estudio y se recoge un conjunto de datos, como normalmente ese conjunto de datos es muy grande, es necesario disponer de alguna herramienta mediante la cual podamos visualizar esos datos. Para ello, una vez ordenados, hacemos un recuento de dichos datos y realizamos tablas estadísticas. En estas tablas, deberán figurar los valores de la variable de estudio, y sus frecuencias correspondientes. Vamos a ver el tipo de tablas que podemos utilizar según la variable estadística sea discreta o continua.

Modalidad (x_i)	**Frecuencia absoluta (f_i)**	**Frecuencia absoluta acumulada (F_i)**	**Frecuencia relativa (h_i)**	**Frecuencia relativa acumulada (H_i)**

La principal dificultad para la obtención de una distribución de frecuencias, reside en la construcción de las modalidades, ya que ésta variará de acuerdo con el tipo de variable que se pretende describir: si la variable es cualitativa, se tomarán como modalidades las distintas respuestas observadas de la muestra; si la variable es discreta (que tome pocos valores distintos), las modalidades coincidirán con los distintos valores medidos en la muestra; si la variable es continua (o bien discreta, pero toma muchos valores distintos), se tomarán como modalidades los intervalos de clase.

Intervalos de clase.-

Son los intervalos donde se encuentran los datos agrupados cuando se estudian variables estadísticas continuas, se denotan por $[L_{i-1}, L_i)$.

El número de clases o intervalos y la longitud que debemos considerar, depende de cada problema y de la utilización que se quiera dar a las tablas estadísticas. Lo normal es que todos los intervalos sean de la misma amplitud $(L_i - L_{i-1})$, aunque pueden existir múltiples razones donde se aconsejen tomar intervalos de amplitud variable, como puede ser el caso en el que existan uno o dos intervalos donde se concentren la mayoría de los datos.

Una vez construidos los intervalos de clase, se elige un representante en cada uno de ellos. Este representante es el valor medio de cada intervalo de clase, y se llama **marca de clase**. Luego la marca de clase para cada intervalo se calcula del siguiente modo:

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

La construcción de los intervalos de clase, introduce algunas cuestiones subjetivas, como son:

1) *¿Cuántos intervalos construir?*

Aunque no existe una regla general para usar, es evidente que el número de intervalos debe ser mayor al aumentar el tamaño muestral, por lo que se recomienda construir tantos intervalos como el número entero (entre 5 y 20) más próximo a $\sqrt{n}$.

2) *¿Qué valor se elige como extremo inferior del primer intervalo L_0?*

Se toma como L_0 un valor "un poco menor" que el mínimo de la muestra (o el mínimo).

Las tablas para datos continuos, quedan de la siguiente manera:

$L_{i-1} - L_i$	x_i Marca de clase	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (h_i)	Frecuencia relativa acumulada (h_i)

Aunque podamos ver la forma general de una distribución en una tabla de frecuencias, la mejor forma es mediante un *histograma de frecuencias* (que es una representación visual de los datos en la que pueden observarse tres propiedades esenciales de una distribución: forma, tendencia central o acumulación y dispersión o variabilidad).

3.4. Gráficos Estadísticos

Una de las herramientas más populares y utilizada dentro de la estadística descriptiva es, sin lugar a dudas, el análisis gráfico de los datos. Como hemos visto, las tablas estadísticas, resumen los datos de que disponemos sobre una población y dan toda la información necesaria, pero como se suele decir, *"Una imagen vale más que mil palabras"*, luego es conveniente expresar la información de que disponemos mediante un gráfico o diagrama, según proceda, con el fin de hacerla más clara y captar de un solo vistazo las características de los datos.

Gracias a los ordenadores y los programas que se han desarrollado en el campo de la informática se pueden realizar fácilmente todo tipo de representaciones gráficas y de gran calidad.

Veamos qué tipo de gráficos podemos realizar dependiendo de la variable estadística que utilicemos en cada caso.

3.4.1. Gráficos para variables cualitativas o atributos

(*) Diagrama de barras o bastones. Este tipo de gráficos se representan de forma cartesiana en un eje de coordenadas mediante unas barras que recorren el eje de ordenadas (Y) desde su origen hasta el valor del punto representado, colocando en el eje de abcisas (X) las diferentes modalidades de la variable y en el eje de ordenadas (Y) la frecuencia relativa o absoluta, según proceda.

Este tipo de gráficos también se puede hacer en el espacio, incorporando una nueva variable (Z) y realizando un dibujo tridimensional.

(*) Diagramas de sectores. Se utilizan para hacer comparaciones de las distintas modalidades de un carácter mediante sectores circulares. Para construirlos se divide un círculo en tantas porciones como modalidades existan de manera que el ángulo central de cada sector ha de ser proporcional a la frecuencia absoluta o relativa correspondiente. El arco de cada porción se calcula realizando una simple regla de tres:

$$\begin{array}{l} N \longrightarrow 360^{o} \\ f_i \longrightarrow \alpha \end{array}$$ Entonces $\alpha = \dfrac{360^{o}.f_i}{N}$ son los grados que tiene que tener cada modalidad según su frecuencia.

Este tipo de diagramas recibe también el nombre de *tartas o quesitos*, por la forma que tiene su representación.

(*) Pictogramas. Quizás es el tipo de gráfico más bonito a la vista, pues en él aparecen dibujos que hacen alusión al fenómeno estudiado, mediante su tamaño, forma …

Para realizarlos se representan a diferentes escalas un mismo dibujo teniendo en cuenta que el perímetro del dibujo tiene que ser proporcional a la frecuencia, pero esto puede incurrir en un efecto visual engañoso ya que a frecuencia doble corresponde un dibujo de área cuádruple, con lo cual tiene un inconveniente debido a la falta de precisión.

A pesar de este inconveniente este tipo de dibujos son muy utilizados por los medios de comunicación a la hora de hacer que el público no especializado comprenda temas complejos sin necesidad de dar una explicación complicada.

3.4.2. Gráficos para Variables Cuantitativas

Para este tipo de variables, tenemos diferentes gráficos según el tipo de frecuencia que usemos y además tenemos que tener en cuenta si la variable es discreta o continua.

Según el tipo de frecuencia usada se dividen en:

a) *Diagramas diferenciales.* Representan el número o porcentaje de elementos de una modalidad. Se representan a partir de las frecuencias absolutas o relativas.
b) *Diagramas integrales.* Representan el número de elementos de una modalidad inferior o igual a la dada. Se representan a partir de las frecuencias acumuladas. Este tipo de diagramas **no** tiene ningún sentido para variables cualitativas.

3.4.2.1. GRÁFICOS PARA VARIABLES CUANTITATIVAS DISCRETAS

(*) Diagrama de barras. Su representación es idéntica a la explicada para variables cualitativas, las barras deben de ser estrechas para mostrar que los valores que toma la variable son discretos. Se usan cuando se pretende hacer un diagrama diferencial utilizando variables discretas.

En el caso de realizar un diagrama integral, es decir, usando frecuencias acumuladas, las barras aparecen formando una escalera.

3.4.2.2. GRÁFICOS PARA BARIABLES CUANTITATIVAS CONTINUAS

(*) Histograma. Para construirlo se representa sobre el eje de abcisas los extremos de las clases definidas por los intervalos de clase $L_{i-1} - L_i$.

Se usan cuando se pretende hacer un diagrama diferencial utilizando variables continuas.

Sobre el eje de abcisas, se construyen rectángulos, tomando como base la amplitud del intervalo y como altura la frecuencia de cada intervalo, siempre que la amplitud de todos los intervalos sea la misma, puesto que el área se obtiene multiplicando la base por la altura.

Por lo tanto, en este caso, cada altura da idea de la densidad o concentración de datos en esa zona:

-Más altura → aparecen más valores de la variable.
-Menos altura → los datos que aparecen son más escasos.

Si los intervalos son de diferentes amplitudes, las alturas de los rectángulos deben ser calculadas teniendo en cuenta que el área de cada rectángulo tiene que ser proporcional a la frecuencia de cada intervalo.

El número de individuos de la muestra viene dado por el área del polígono que forma el histograma. Este tipo de gráficos representa frecuencias mediante áreas. Y si se expresan frecuencias relativas, el área total encerrada en el histograma es uno.

A diferencia del diagrama de barras, los rectángulos verticales, se representan contiguos para reflejar la idea de que la variable es continua. La forma del histograma refleja propiedades importantes de la variable estadística a la que se refiere.

A la hora de hacer un histograma es muy importante hacer una buena elección de la cantidad de clases a utilizar. Para este fin se utilizan distintas reglas, una de ellas consiste en tomar el número de clases igual al entero más próximo a la raíz cuadrada del número de observaciones que se estudian, N.

(*) Polígono de frecuencias. Se construye fácilmente una vez representado el histograma, y consiste en unir los puntos del histograma que corresponden a las marcas de clase de cada intervalo mediante una recta.

El diagrama integral para variables continuas se denomina también *polígono de frecuencias acumulado u ojiva*. En estos polígonos obtenidos se aprecian con claridad propiedades importantes, como si la curva es no creciente, de donde a donde se desplaza …

La diferencia esencial entre los histogramas y los polígonos de frecuencias es que estos últimos proporcionan una representación más suavizada de la distribución de frecuencias.

3.4.2.3. RESUMEN: DIAGRAMAS SEGÚN EL TIPO DE VARIABLES

Tipo de variable	Diagrama o gráfico
Cualitativa	Barras, sectores, pictogramas
Cuantitativa (discreta)	Diferencial (barras) Integral (escalera)
Cuantitativa (continua)	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulativos)

4. MEDIDAS DE CENTRALIZACIÓN, DISPERSIÓN, POSICIÓN Y FORMA.

Pasamos a estudiar las distintas formas de resumir las distribuciones de frecuencias estudiadas mediante **medidas de posición** (o de centralización), teniendo presente el error cometido en el e resumen mediante las correspondientes **medidas de dispersión**. A su vez, analizaremos la forma de la distribución mediante las **medidas de forma**. El histograma de frecuencias ya nos daba una representación visual de las tres propiedades anteriores. Se trata ahora de cuantificar estos conceptos.

4.1. Medidas de centralización

Se trata de encontrar medidas que sinteticen las distribuciones de frecuencias. En vez de manejar todos los datos sobre las variables, podemos caracterizar su distribución de frecuencias mediante algunos valores numéricos, eligiendo como resumen de los datos un valor central alrededor del cual se encuentran distribuidos los valores de la variable.

- **Media aritmética:** se define como la suma de los datos dividida por el número de ellos. Es decir:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{m} f_i x_i = \sum_{i=1}^{m} h_i x_i$$

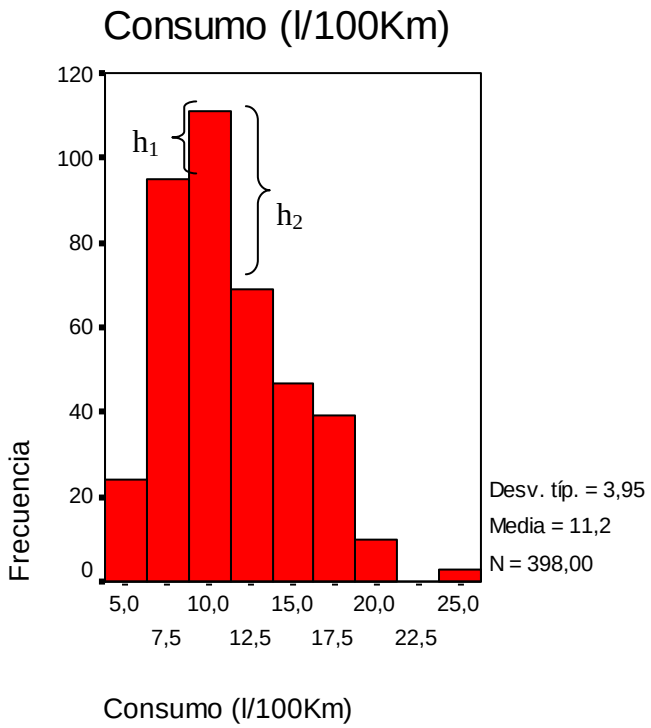
  donde las f_i son las frecuencias absolutas y h_i las relativas.
  Cuando la variable es continua, en vez de la observación, cogemos la marca de clase.

- **Media aritmética ponderada:** Es un caso particular de la media aritmética, que aparece cuando se otorga a cada valor de la variable x_i una ponderación o peso w_i. Así, tenemos:

$$\overline{X} = \frac{\sum_{i=1}^{m} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Moda:** Es el valor de la variable estadística que presenta mayor frecuencia. No tiene por qué ser única, y puede no poderse calcular. Según el número de modas, la distribución recibe el nombre de unimodal, bimodal, trimodal …

  Para variables continuas: $M_0 = a + c\frac{h_1}{h_1 + h_2}$, siendo c=b-a, [a,b) el intervalo modal. (El intervalo modal es el de mayor frecuencia, c= longitud del intervalo modal). Además, h_1, h_2 son las alturas del histograma:



Ejemplo.

L_{i-1}-L_i	f_i
[82-87)	3
[87-92)	7
[92-97)	8
[97-102)	8
[102-107)	7
[107-111)	7

En caso de no necesitar ni la marca de clase ni la frecuencia relativa, no las calculo.

Como dos intervalos tienen la misma frecuencia, el intervalo modal es la suma de los dos, es decir, M_o=[92-102), luego c=102-92=10.

$h_1 = 8 - 7$
$h_2 = 8 - 7$

Por tanto, $M_0 = 92 + 10\frac{1}{1+1} = 97 \in [92-102)$

- **Mediana (M_d):** es el valor de la variable que deja el 50% de los datos a un lado y a otro.
  Tenemos que diferenciar entre:

  Variables discretas:
  - Si el número de datos es impar → M_d es el valor central.
  - Si el número de datos es par → M_d es la semisuma de los valores centrales.

  Variables continuas (Veremos su valor cuando estudiemos los parámetros de dispersión).

4.2. Medidas de dispersión

Permiten *calcular la representatividad de una medida de posición*, para lo cual es preciso cuantificar la distancia entre los diferentes valores de la distribución respecto a dicha medida. (A esta distancia es a lo que se denomina ***variabilidad o dispersión de la distribución***).
La finalidad de estas medidas es estudiar *hasta qué punto para una determinada distribución de frecuencias, las medidas de tendencia central o de posición son representativas* como síntesis de toda la información de la distribución.
Medir la representatividad de una medida de posición equivale a cuantificar la separación de los valores de la distribución respecto a dicha medida.
A la mayor o menor separación de los valores de una distribución respecto del valor de posición se le llama ***dispersión o variabilidad***.

Se distinguen entre:

→ Medidas de dispersión relativas (no dependen de las unidades de medida)
→ Medidas de dispersión absolutas

(También se pueden distinguir las medidas en las formas anteriores, según sean medidas referentes a promedios o no lo sean).

4.2.1. Medidas de dispersión absolutas no referentes a promedios

- **Rango o Recorrido**: es la diferencia entre el mayor y menor valor de una distribución. (Para variables discretas será $\max\{x_i\}$- $\min\{x_i\}$, y para variables continuas será $L_p - L_o$).
- **Recorrido intercuartílico**: diferencia entre el tercer y el primer cuartel (ya lo veremos)

4.2.2. Medidas de dispersión relativas no referentes a promedios

- **Coeficiente de apertura**: cociente entre el mayor y menor valor de una distribución)
- **Recorrido relativo**: cociente entre el recorrido y la media.
- **Recorrido semicuartícilo**: cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartel $= \frac{recorrido\ \text{int}\ ercuartílico}{1^{er}\ cuartil + 3^{er}\ cuartil}$.

4.2.3. Medidas de dispersión absolutas referentes a promedios

Estas miden el error que cometemos usando el promedio en cuestión como resumen de datos.

- **Desviaciones medias:**

  .Desviación media respecto de la media aritmética:

$$D.M. = \frac{\sum_i | x_i - \bar{x} |\ f_i}{N}$$

  .Desviación media respecto de la mediana:

$$D.M. = \frac{\sum_i | x_i - M_d |\ f_i}{N}$$

(el problema de estas es que es mas complicado trabajar con valores absolutos).

- **Varianza, cuasivarianza, desviación típica y error estándar**

De todas las medidas de dispersión absoluta respecto de la media aritmética, la varianza y su raíz cuadrada son las más importantes.

  .Varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2 f_i}{N}$$

Propiedades de la varianza:

-Nunca puede ser negativa.
-Un cambio de origen en la variable no afecta a la varianza
-Al multiplicar los valores por una constante k, la varianza queda multiplicada por esa constante al cuadrado.

.Desviación típica. Así como las desviaciones medias vienen expresadas en las mismas unidades de medida que la distribución, la varianza **no** vendrá dada en las unidades correspondientes sino elevadas al cuadrado, así la desviación típica:

$$\sigma = +\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i}{N}}$$

.Cuasivarianza. Es un estadístico muy usado por sus propiedades muestrales, se define por:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i}{N-1}$$

.Cuasidesviación típica

$$S = +\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i}{N-1}}$$

.Error estándar

$$e = \frac{S}{\sqrt{N}}$$

4.2.4. Medidas de dispersión relativas

Se utiliza para comparar medidas de centralización o promedios.

.Coeficiente de variación de Pearson: compara medias aritméticas de varias distribuciones que pueden venir, en general, en unidades diferentes:

$$C.V. = \frac{\sigma}{\overline{X}}$$

(Al efectuar el cociente, eliminamos las unidades, luego C.V. es adimensional).

A menor C.V. mejor es la media.

Generalmente aparece en %, para ello multiplicamos C.V. por 100.

4.3. Parámetros de posición.

A todos ellos se les denomina **cuantiles**, y estos se dividen en:

- CUARTILES
- QUINTILES
- DECILES
- CENTILES **ó** PERCENTILES

(Distinguiremos el cálculo de ellos para variables discretas, i.e., datos sin agrupar, y para variables continuas, i.e., datos agrupados).

DATOS SIN AGRUPAR

- **Cuartiles**: son aquellos que dividen a la población en 4 partes iguales, por tanto son tres. Se denotan con Q_1, Q_2, Q_3.

  $$Q_i = N\frac{i}{4}$$

  Observa que Q_2=mediana.

- **Quintiles**: son los valores que dividen a la población en 5 partes iguales. (Por tanto hay 4). Se representan por K_i, i=1,2,3,4. Se calculan:

  $$K_i = N\frac{i}{5}$$

- **Deciles:** Son los valores que dividen a la población en 10 partes iguales. (Por tanto hay 9). Se preseentan por D_i, i=1, …, 9.
  Se calculan:

$$D_i = N\frac{i}{10}$$

  Observa que $D_2=K_1$, y $D_5=M_d$

- **Centiles o percentiles:** Son los valores que dividen a la población en 100 partes iguales. (Existen 99: $c_1, \ldots, c_{99} = p_1, \ldots, p_{99}$).

  Se calculan

$$c_i = N\frac{i}{100}$$

Nota: Una vez que hemos calcula el lugar que ocupa el cuantil, miramos el dato que corresponde al lugar F_i, al igual que hacíamos cuando calculamos la mediana.

DATOS AGRUPADOS

$$Cuantil_i = L_{i-1} + \frac{i\left(\frac{N}{p}\right) - F_{i-1}}{f_i} c_i$$

con p=4 → cuartil, p=5 → quintil, p=10 → decil, p=100 → percentil o centil.

(Primero se calcula i.(N/p), y se mira a ver en qué intervalo cae el número que salga, y es con este intervalo en el que nos tenemos que fijar para usar F_{i-1}, f_i,… c_i es la longitud de ese intervalo).

EJEMPLO 1.-

x_i	f_i	F_i
1	1	1
2	2	3
3	1	4
4	3	7
5	3	10
		10

Calcula el decil 3.

$D_3 = 3.\ (10/10) = 3$, por tanto $D_3 = 2$ (recuerda que el decil es el VALOR DE LA VARIABLE QUE DEJA …)

EJEMPLO 2.-

Calcular los todos los cuarteles y los percentiles 40 y 90, donde:

L_{i-1}-L_i	f_i	F_i
38-44	7	7
44-50	8	15
50-56	15	30
56-62	25	55
62-68	18	73
68-74	9	82
74-80	6	88

$Q_1=1.(N/4)=1.(88/4)=22$ → no llega a 30, luego $Q_1 \in [50,56)$, ahora ya, usamos la fórmula:

$Q_1=50 + (1.22-15)/15\ .\ 6 = 52,8$, es decir, antes de 52,8 están el 25% de los casos.

$Q_2=2.(N/4)=2.(88/4) = 44$ → luego $Q_2 \in [56,62)$, ahora ya, usamos la fórmula:

$Q_2=56 + (44-30)/25\ .\ (62-56) = 59,36$, es decir, antes de 59,36 están el 50% de los casos.

$Q_3=3.(N/4)=3.(88/4)=66$ → luego $Q_3 \in [62,68)$, ahora ya, usamos la fórmula:

$Q_3=62 + (66-55)/18\ .\ 6 = 65,67$

c_{40}=40.(N/100)=40.(88/100)=35,2 → $c_{40} \in$[56,62), ahora ya, usamos la fórmula:

c_{40}=56 + (35,2-30)/25 . 6 = 57,25

c_{90}=90.(N/100)=90.(88/100)=79,2 → $c_{90} \in$[68,74), ahora ya, usamos la fórmula:

c_{90}=68 + (79,2-73)/9 . 6 = 72,14

4.4. Medidas de Forma

Las medidas de forma de una distribución se basan en su representación gráfica sin llegar a la misma. Estas medidas las clasificamos en dos grupos:

→ *Medidas de asimetría:* cuya finalidad es elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución sin hacer su gráfica.

→ *Medidas de curtosis:* estudian la distribución de frecuencias en la zona central de la distribución.

Coeficientes de asimetría

Elaboran un indicador que permite establecer el grado de simetría (o asimetría) que presenta una distribución sin realizar su presentación gráfica.

Coeficiente de asimetría de Pearson:

Se define como:

$$A_p = \frac{\overline{X} - M_o}{\sigma}$$

Este valor se lee de la siguiente manera:

- Si $A_p > 0$ → asimetría a la derecha o positiva
- Si $A_p = 0$ → simétrica
- Si $A_p < 0$ → asimétrica a la izquierda o negativa

Coeficiente de asimetría de Fisher

$$A_F = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N}\sum (x_i - \bar{x})^3 f_i}{\sqrt{\left(\frac{\sum (x_i - \bar{x})^2 f_i}{N}\right)^3}}$$

Misma interpretación que el coeficiente de Pearson:

- Si $A_F > 0$ → asimetría a la derecha o positiva
- Si $A_F = 0$ → simétrica
- Si $A_F < 0$ → asimétrica a la izquierda o negativa

Coeficiente de Apuntamiento o de Curtosis

Aplicable a distribuciones campaniformes unimodales simétricas o con una ligera simetría.

Pueden adoptar las siguientes configuraciones y nombres:

Platicúrtica (achatada)	Mesocúrtica (normal)	Leptocúrtica

Bajo apuntamiento si y solo si gran aplastamiento.

La fórmula del coeficiente de apuntamiento o curtosis es:

$$g_2 = \frac{m_4}{\sigma^4} - 3$$

- Si $g_2 > 0$ → distribución leptocúrtica.
- Si $g_2 = 0$ → distribución mesocúrtica.
- Si $g_2 < 0$ → distribución platicúrtica.

Relación entre media y desviación típica.-

En el intervalo media ± σ caen el 68% de las observaciones.
En el intervalo media ± 2σ caen el 95% de las observaciones.
En el intervalo media ± 3σ caen el 99% de las observaciones

Relación entre media, moda y mediana

Cuando la MEDIA = MODA = MEDIANA, decimos que la distribución es simétrica.

Introducción al paquete estadístico "SPSS"

Tipos de ficheros:

El paquete estadístico SPSS permite manipular ficheros de una manera fácil y cómoda. Un *fichero de datos* (*nombrefichero.sav*) se estructura en variables (columnas) en las que se guardan las distintas observaciones que se han tomado para cada una de ellas. Cada fila corresponde a un *caso* (sujeto o unidad estadística). Estos ficheros además de los datos tienen la información necesaria para su procesamiento. Otro tipo son los *ficheros de resultados (nombrefichero.spo),* con posibilidad de exportar las tablas a otras aplicaciones bien como objeto o bien como tabla. Además se pueden modificar quitando o añadiendo cosas.

Variables:

Los ficheros de datos tienen dos modos. En el modo *vista de datos* es posible introducir o modificar los datos para cada una de las variables. En el modo *vista de variables* se puede dar formato a cada variable. Así, se puede dar nombre a la variable (nunca más de 8 caracteres ASCII y siempre serán consideradas como minúsculas), poner etiquetas de identificación (tanto para la variable, como para las categorías de la misma), definir los datos perdidos o ausentes (missing) o determinar la anchura de texto en variables cadena, la alineación y la anchura de visualización de una columna. Por último se puede definir el tipo (y escala de medida) de una variable:

- *Numérica*: Variable numérica usual delimitada la parte decimal con un punto o una coma, según esté configurado. Ejemplo: 12345.34 ó 12345,34 34 (doce mil trescientos cuarenta y cinco con treinta y cuatro).
- *Coma*: Variable numérica delimitada la parte decimal con un punto y en la parte entera una coma cada tres dígitos indicando los miles. Ejemplo: 12,345.34 (doce mil trescientos cuarenta y cinco con treinta y cuatro).
- *Punto*: Variable numérica delimitada la parte decimal con una coma y en la parte entera un punto cada tres dígitos indicando los miles. Ejemplo: 12.345,34 (doce mil trescientos cuarenta y cinco con treinta y cuatro).
- *Notación científica*: Variable numérica en la que los números vienen expresados con notación exponencial con base 10. Ejemplo: 1,234534 E+04 (doce mil trescientos cuarenta y cinco con treinta y cuatro).
- *Fecha*: Fechas en distintos formatos.
- *Dólar*: Moneda americana. Aparece con un $ a la izquierda de la cantidad.
- *Moneda personalizada*: Moneda de cada país definida previamente en las opciones.
- *Cadena*: Variable cualitativa. En algunas ventanas de diálogo cuando sea preciso dar el nombre de una categoría, esta habrá de ir entre comillas simples. Ejemplo: `nivel='BAJO'`. No es lo mismo utilizar mayúsculas o minúsculas, así `BAJO` y `bajo` se consideran categorías distintas.

Menús:

Es importante saber que en cada tipo de fichero aparece un menú distinto. En general el menú *Archivo* ofrece la posibilidad de abrir y guardar ficheros de diversos tipos.

La opción *Mostrar información de datos* proporciona información sobre un fichero de datos seleccionado.
El menú *Edición* ofrece la posibilidad de *Cortar*, *Copiar*, *Pegar* y *Borrar* datos.
Además en un fichero de datos permite *Buscar* determinados datos. En *Opciones* se puede configurar el formato genérico de nuestros ficheros. El menú *Ver* proporciona diversas posibilidades de visualización.
En los ficheros de datos el menú *Datos* ofrece opciones para la definición de las variables y manipulación de los datos.

i) Es posible generar fechas en el formato deseado. Esta opción se puede utilizar también para generar listas de números.

ii) Las opciones de *Insertar* permiten insertar columnas o filas en un fichero de datos determinado.

iii) *Ir a caso* y *Ordenar casos* permiten respectivamente ir a una fila determinada y ordenar los datos de acuerdo a una o más variables respectivamente.

iv) La opción *Transponer* trasforma filas en columnas y columnas en filas.

v) Es posible *Reestructurar* el fichero mediante un asesor. Esta opción es de interés cuando los datos provienen de otras aplicaciones que no tienen la estructura exigida por el SPSS para su tratamiento.

vi) *Fundir archivos* sirve para unir en un fichero variables o filas de dos ficheros dados. Puesto que el SPSS solamente permite tener un fichero activo esta operación crea un nuevo archivo que añade filas (columnas) de otro fichero, con la posibilidad de prescindir de algunas de las filas (columnas) del fichero activo.

vii) Con *Agregar* se hacen grupos de una o más variables (*Variable(s) de segmentación*) con referencia a una o más variables (Agregar v*ariable(s)*) asignando a cada grupo la media o la medida de posición o dispersión que se determine. Las variables obtenidas se guardan en un nuevo fichero. Puede ser útil cuando se tienen réplicas de un experimento y se quiere trabajar con las medias de cada uno.

viii) También es posible generar o mostrar *Diseños ortogonales* con los factores deseados y sus categorías.

ix) *Segmentar archivo* permite hacer grupos de casos de acuerdo a un criterio dado por una variable. Los análisis que se hagan posteriormente se realizarán para cada grupo y los resultados se mostrarán en una tabla comparativa o en varias tablas según se haya elegido la opción correspondiente.

x) Con *Seleccionar casos* se pueden eliminar, definitiva o temporalmente, algunas filas de acuerdo a algún criterio. Se creará una columna de filtros con unos para los casos seleccionados y ceros para el resto. Todos los análisis que se hagan a partir de entonces utilizarán solamente los casos seleccionados.

xi) Por último es posible *Ponderar casos* por una variable de pesos con el objeto de que los análisis estadísticos que se realicen mantengan dicha ponderación. Así un dato que se pondera por 4 tendrá doble valor (peso, ponderación) en los análisis correspondientes que otro que solamente sea ponderado por 2.

Con *Transformar* podemos realizar manipulaciones de las variables. Para ello utilizaremos las opciones:

i) *Calcular*, nos ofrece una ventana semejante a una calculadora que permite realizar operaciones entre las columnas. Además existe una lista de funciones, cada una de las cuales viene explicada en la ayuda del programa. La alternativa *Si* posibilita la inclusión de condicionales en el momento de hacer las operaciones deseadas. Cuando la condición se impone sobre los valores de una variable cualitativa, estos deben ponerse entre comillas simples. Ejemplo: raza = 'blanco'.

ii) *Semilla de aleatorización* permite asignar una semilla para la generación de números aleatorios. Si se fija la semilla, la secuencia que se obtiene es la misma, por ejemplo para dos usuarios distintos.

iii) *Contar apariciones* crea una nueva variable que asigna a cada caso el número de veces que se repite un valor o valores determinados en una fila para las variables seleccionadas.

iv) *Recodificar* crea una variable (*en variables diferentes*) o sustituye a la ya existente (*en las mismas variables*) con valores que se asignan de acuerdo a un criterio. Con *If* podemos seleccionar solamente los casos que interesa cambiar. Con *valores antiguos y nuevos* se determinan los cambios específicos a realizar.

v) *Categorizar variables* crea una nueva variable en la que los datos numéricos se convierten en un número prefijado de categorías. Los datos se categorizan según grupos percentiles; de modo que cada grupo contiene aproximadamente el mismo número de casos.

vi) La opción *Asignar rangos a casos* crea una variable que asigna rangos a cada uno de los casos. Es posible seleccionar el tipo de rangos que se desea y también el modo de tratar los empates. Además se pueden hacer grupos de acuerdo a un criterio proporcionado por una variable. En este caso se asignan rangos a cada grupo de manera independiente.

vii) *Recodificación automática* convierte los valores numéricos y de cadena en valores enteros consecutivos asignando un 1 al valor más bajo, 2 al siguiente, y así sucesivamente. También es posible hacerlo comenzando por el valor más alto. La nueva variable conserva las etiquetas de valor de la variable antigua. En el fichero de resultados se muestra una tabla con los valores antiguos, los nuevos y las etiquetas de valor. Los valores de cadena se recodifican por orden alfabético, con las mayúsculas antes que las minúsculas y los valores perdidos en primer lugar. En caso de empate

se asigna el mismo número a todos los valores empatados y se continúa en el siguiente.

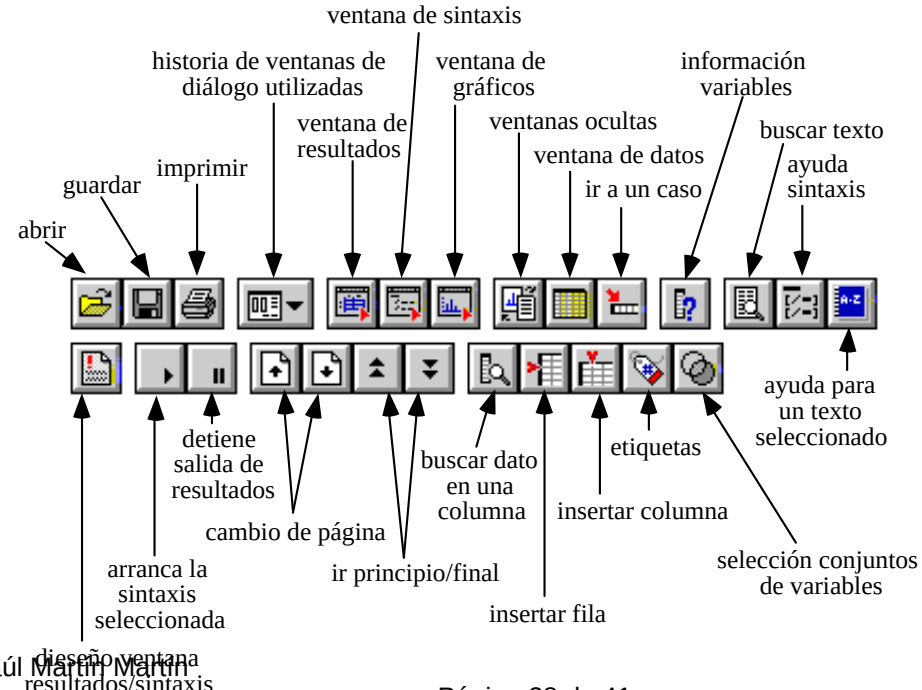
viii) *Crear serie temporal* genera variables basadas en funciones (de diferencias, medias móviles, medianas móviles, retardo o adelanto) de las variables de series temporales numéricas seleccionadas. Los nombres de las nuevas variables por defecto se componen de los seis primeros caracteres de la variable existente utilizada para crearlas, seguidos de un guión bajo y de un número secuencial.

ix) *Remplazar valores perdidos* asigna valores a los casos omitidos de acuerdo a un criterio determinado:

- *Media de la serie*: asigna la media de los casos existentes.
- *Media de los puntos adyacentes*: media de los puntos más cercanos, pudiéndose elegir el número de datos válidos por encima y por debajo que se desean incluir.
- *Mediana de los puntos adyacentes*: mediana de los puntos más cercanos, pudiéndose elegir el número de datos válidos por encima y por debajo que se desean incluir.
- *Interpolación lineal*: Hace interpolación lineal entre el último valor válido antes del valor perdido y el primer valor válido después del valor perdido.
- *Tendencia lineal en el punto*: Se hace regresión de la serie existente sobre una variable índice escalada de 1 al número de datos (filas) en la muestra y los valores perdidos se sustituyen con sus valores pronosticados.

Es importante puntualizar que los cuatro últimos métodos dependen de la ordenación de los datos. Por ejemplo el procedimiento puede ser adecuado cuando se han obtenido secuencialmente en el tiempo.

Funciones de la barra de herramientas del SPSS



PRÁCTICA I: MANEJO DE FICHEROS Y VARIABLES

¡Puede seguir las indicaciones que aparecen al final de cada apartado!

2. Construya un fichero de datos con el nombre `riesgo.sav` que contenga las variables siguientes correspondientes a un país ficticio:

AÑO	DEUDA	RENTA
1981	131.53	
1982	189.92	43.22
1983	149.84	35.84
1984	120.09	22.84
1985		52.99
1986	99.89	29.03
1987	81.08	29.81
1988	189.49	48.82
1989	1100.03	229.31
1990	254.11	55.32
1991	829.90	155.04
1992	283.94	82.53

AÑO	DEUDA	RENTA
1993	449.31	100.81
1994	308.52	48.35
1995	292.85	98.00
1996	525.09	119.82
1997	1389.19	
1998	5382.18	1093.02
1999	1418.42	345.99
2000	5815.87	1187.80
2001	2834.15	800.50
2002	3942.50	918.98
2003	2480.33	577.94
2004	12977.10	1710.52

2. Manipulación del fichero:

a) Haga una copia de seguridad del fichero anterior con la opción **`Archivo -> Guardar como`** con el nombre `copia_riesgo.sav`. Borre la variable `año` en la nueva copia y guarde de nuevo el fichero. A partir de ahora trabajaremos con esta nueva versión.

b) Cree una nueva variable cualitativa con el nombre `nivel` con tres valores: BAJO si `deuda` es menor de 200, MEDIO si está entre 200 y 700 y ALTO en el resto.

Indicaciones: Transformar -> Recodificar -> En distintas variables… -> Var. de entrada: deuda; Var. de resultado: nivel (pulse Cambiar) -> Valores antiguos y valores nuevos: (Seleccione la opción de variable cadena) Valor antiguo: del menor hasta 200 - valor nuevo: BAJO; valor antiguo: Rango: 200 hasta 700 - valor nuevo: MEDIO; valor antiguo: Rango: 700 hasta el mayor - valor nuevo: ALTO. Los rangos incluyen sus puntos finales y los valores definidos como perdidos por el usuario que estén dentro del rango.

c) Codifique la variable `nivel` con los valores 1, 2 y 3 respectivamente para BAJO, MEDIO y ALTO en una nueva variable llamada `nivel2`. Quite los decimales, si los tiene, de esta variable codificada.

Indicaciones: Transformar -> Recodificar -> En distintas variables… -> Var. de entrada: nivel; Var. de resultado: nivel2 -> Valores antiguos y valores nuevos: valor antiguo: BAJO - valor nuevo: 1; valor antiguo: MEDIO- valor nuevo: 2; valor antiguo: ALTO - valor nuevo: 3.

PRÁCTICA II: ANÁLISIS DESCRIPTIVO DE DATOS

Usar el fichero habitos.sav

Representación y gráficas de datos

Diagrama de tallos y hojas (steam and leaf plot)

Son procedimientos semigráficos, es decir, aparece un gráfico y una tabla. Representan la información para caracteres cuantitativos (*no vale si son cualitativos*).

Elementos del diagrama:

- Tallo: constituido por los primeros elementos o dígitos, y aparece puesto en vertical.
- Hojas: son los siguientes elementos de cada uno de los datos de nuestra variable. (Aparece en vertical).

Si lo giramos 90º es parecido a un histograma pero con más información.

El inconveniente de estos diagramas es que cuando tengamos un número elevado de datos, son difíciles de construirlos "a mano", y además, cuanto más grande sea el número de datos, la eficacia es menor.

(Si hay más de 100 datos, el diagrama de tallos y hojas no es eficiente).

Utilidad.

La utilidad de estos semi-gráficos es que podemos representar dos distribuciones a la vez, poniendo un tallo común y hojas a la derecha y a la izquierda y así compararlas.

Se utiliza para explicar el patrón de comportamiento.

Número de ramas (o filas)

Depende del analista. Se aconseja que si:

- $n > 100$ → $L = 10 \log_{10} N$
- $n < 100$ → $L = 2 \text{ sqrt}(2)$

En este tipo de diagramas, vamos a obtener y/u observar:

1) Rango
2) Localización de los valores centrales
3) Concentraciones o agrupaciones
4) Identifiación de valores (ej_ lo que no son frecuentes o al contrario)
5) Lagunas (o "gaps"): cuando no se han registrado valores (habrá huecos en los tallos)
6) Dispersión y simetría
7) Anomalías (datos extremos, outliers) Además identificarán qué dato es.

Pasos para dibujar un diagrama de tallos y hojas

1º) Escoger el intervalo de unidades (tronco) que cubra la totalidad de los datos, para ello reordenamos dichos datos. (¡ Los tallos nunca pueden ser números decimales, siempre enteros !).
(Es conveniente hacer más de uno, empleando distintas unidades)

2º) Suprimir la última fila de cada datos (ej. Si tenemos 112, quitamos el 2). (Esto no tiene por qué ser así, cada analista hace lo que cree). Después ordenamos de forma creciente, de menor a mayor, y eliminamos las repeticiones. Con eso tenemos los tallos.

3º) Trazamos una línea vertical que me separe los tallos y las hojas. (Anotar la unidad representada en el tronco).

Cada datos se anota en la fila correspondiente al tallo. (Sólo escribo la última cifra).
Frecuencia absoluta = número de hojas de cada tallo.

4º) Ordenar las hojas y añadir una columna de recuento, que se añade a la izquierda del diagrama (esta indica la frecuencia acumulada de cada tallo)

Dependiendo del autor, esta columna es:
- de tipo ascendente (si la clase no supera a la clase mediana)
- de tipo descendente (en caso contrario)

La mediana se pone entre paréntesis.

(Mediana) → Presenta el mayor recuento de las hojas.

Variantes del diagrama de tallos y hojas

- Agrupaciones por intervalos (reduciendo casos)

Reducir a la mitad la amplitud subdividiendo cada tallo en dos:
Notación 1ª, 1b ó 1_*, 1_o

a, $_*$ → 0 – 4
b, o → 5 – 9

Otra subdivisión (para ejemplos grandes) es dividir el tallo en 5 partes:

$_*$ → 0 - 1
t → 2 – 3
f → 4 – 5
s → 6 – 7
o → 8 – 9

Ejemplo.- Muestra del espesor de suelos de hormigón dad en cm:
11,357 ; 12,542 ; 11,384 ; 12,431; 14,212 ; 15,213 ; 13, 300 ; 11,300 ;
12,710 ; 13,455 ; 16,743 ; 12,162 ; 12,721 ; 13, 420; 14,698

1) **Reordenamos los datos (pues el tallo debe ser de 2 a 3 dígitos, cuantos menos mejor)**

Reordeno para quedarme con 3 cifras:

114 125 114 124 142 152 133 113 113 127 135 167 122 127 134 147

(El tallo es grande)

2) **Elimino la última cifra de los datos**

11 12 11 12 14 15 13 11 12 13 16 12 12 13 14

Tachamos cifras que se repiten y obtengo el tallo

TALLO : 11 12 13 14 15 16 (reordenados)

3) Coloco el tallo

11	4 4 3
12	5 4 7 2 7
13	3 5 4
14	2 7
15	2
16	7

Unidad: 10 000 dígitos de mi muestra original ej: 11,357
O también 11/1 indicando que va desde 11100 – 11199

4)

Recuento	F_i	tr	Hojas
3	3	11	3 4 4 (ordenadas)
5	$M_e(5)$	12	2 4 5 7 7
3	7	13	3 4 5
2	4	14	2 7
1	2	15	2
1	1	16	7

Ejercicio. Construye el diagrama de tallos y hojas para los datos

112 112 115 212 213 213 215 342 358 361 362 383 433 436
438 513 568

Ejercicio. Idem para las siguientes calificaciones de alumnos y alumnas:

Alumnos:

68 65 65 74 73 72 70 79 79 79 80 81 82 84 85 88 89 90 91 91 92 96

Alumnas:

65 73 78 78 82 83 87 88 89 90 91 91 93 94 95 95 96 97 98

DIAGRAMA DE CAJAS Y BIGOTES (BOX –and- WHISHER PLOT)

Es una forma de representar gráficamente un conjunto de estadísticos descriptivos. Esto nos permitirá detectar datos extraños (outliers) y asimetrías ya que el gráfico se divide en cuatro áreas de igual frecuencia.

Las características del gráfico son:

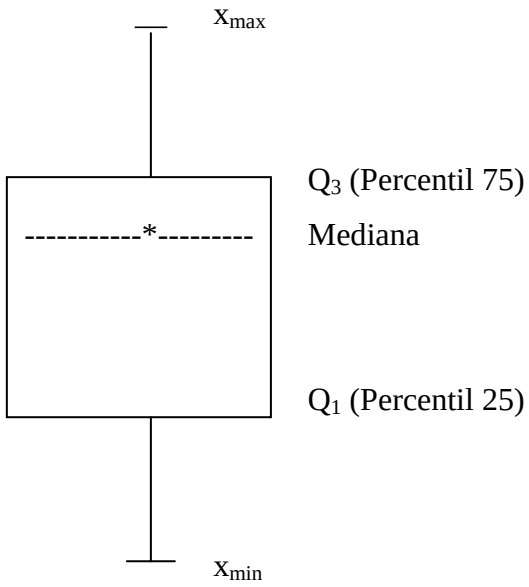
1. Tamaño no muy grande. (Si tengo muchos datos, este diagrama no es significativo. Se utiliza cuando hacemos un histograma y no vemos muy clara su interpretación).
2. Gráfico basado en las medidas de posición.
3. Intervienen 5 cantidades: Mediana (Q_2), cuartiles (Q_1 y Q_3), mínimo (x_{min}) y máximo (x_{max}).
4. Ofrece un resumen de la información más relevante de la distribución (SIN QUE LOS DATOS APAREZCAN)
5. Da los valores de extremos y los outliers.
6. Sirve para comparar distribuciones de dos variables.

Valores indicativos principales:

- Localizaciones
- Agrupaciones significativas de valores
- Zonas en las que predomina la dispersión
- Relación entre agrupaciones y dispersión
- Referencia visual de la simetría central y de los extremos
- Referencia visual de la curtosis (relacionando la longitud de la caja y patillas o bigotes)
- Longitud de colas
- Rango
- Outliers, anomalías o valores alejados del grupo central de los datos

E valores + 3.0 (Extreme)

O valores + 1.5 (outliers)



O valores – 1.5 (outliers)

E valores + 3.0 (extremos)

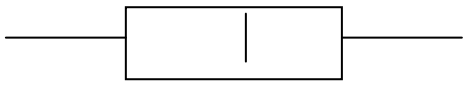
En la caja está el Q_1 y el Q_3, por tanto, en el rectángulo se encuentra el 50% de los datos.

La longitud de la caja es el rango intercuartílico IQR = $Q_3 - Q_1$ (es decir, indica la dispersión de los datos CENTRALES).
(Si tengo 2 distribuciones, el que tenga IQR mayor, es el que tiene los datos más dispersos).

Mediana: mide la tendencia central, es decir, me indica donde está el centro de los datos.

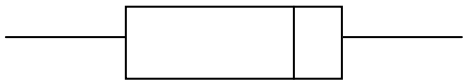
Simetría – Asimetría

1) Si la mediana está justamente en el centro, entonces la distribución es simétrica.



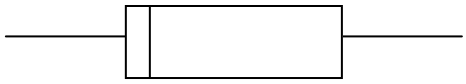
2) Si la mediana no está en el centro, entonces la distribución es sesgada (asimétrica).

2.1. Si la mediana está



entonces la distribución es asimétrica negativa. (izquierda)

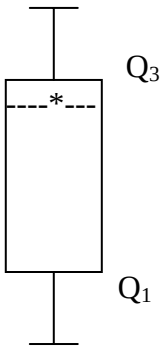
2.2. Si la mediana es



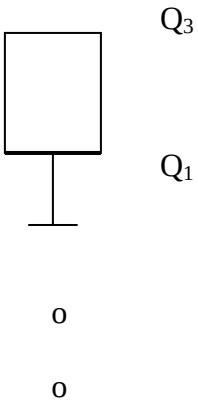
Entonces la distribución es asimétrica positiva. (derecha)

Ejercicio.-

Supongamos que tenemos dos distribuciones de variables, cuyos diagramas BOX-PLOT son los siguientes:



Variable 1



Variable 2

1) Tienen distinta variabilidad
Los datos son más dispersos en Var 1 que en la Var 2 pues $IQR_1 > IQR_2$

2) La variabilidad de VAR1 se da entorno al centro, mientras que en VAR2 se produce alrededor de un extremo.

3) En VAR2, los valores inferiores son considerados casos outliers (peso negativo para calcular la media)

4) VAR1 es asimétrica positiva. En VAR 2 $Q_1 = M_e$, es mucho más asimétrica que VAR1, pues no tiene patilla.

PRÁCTICA III
PROCEDIMIENTOS GRÁFICOS EN EL AED

1) Representar un gráfico de caja y bigotes que resuma los datos dados por la variable potencia de los automóviles (cv) del fichero de datos sobre coches (COCHES). Representar sobre este gráfico la media, la mediana y los valores atípicos. Interpretar los resultados y analizar gráficamente la simetría.

Indicaciones: Gráficos -> Diagramas de Caja -> Simple y Resúmenes para distintas variables -> Definir

2) Representar un gráfico de caja y bigotes que resuma los datos dados para la variable potencia de los automóviles (cv) del fichero de datos sobre coches (COCHES), clasificados en tres gráficos simples de caja y bigotes. Esta clasificación vendrá dada por los valores 1,2 y 3 de la variable origten (región de origen de los coches), cuyas etiquetas respectivas son: EE.UU., Europa y Japón.

Indicaciones: Gráficos -> Diagramas de Caja -> Simple y Resúmenes para grupos de casos -> Definir

3) La encuesta de población activa elaborada por el INE referente al 4º trimestre de 1970 presenta para el número de activos por ramas los siguientes datos:

RAMA DE ACTIVIDAD	MILES DE ACTIVOS
Agricultura, caza y pesca	3706,3
Fabriles	3437,8
Construcción	1096,3
Comercio	1388,3
Transporte	648,7
Otros servicios	2454,8

Realizar un gráfico de sectores con las etiquetas de las ramas de actividad sobre los sectores, otro con porcentajes del número de activos por ramas y etiquetas, y otro con porcentajes del número de activos por ramas, etiquetas y valores.

Indicaciones: Gráficos -> Sectores …