DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Sistemas Inteligentes de Gestión

Guión de Prácticas de Minería de Datos

Práctica 0

Estadística Descriptiva y Preprocesamiento

Introducción a SPSS

© *Juan Carlos Cubero & Fernando Berzal*

NOTA:

Todos los ejercicios de esta primera práctica son de tipo C salvo los dos últimos; de tipo B y A, respectivamente.

Las partes marcadas como "Ampliación" son de lectura opcional y no contribuyen a la calificación final de las prácticas de la asignatura.

Ficheros que se han de entregar:

```
P0_Preprocesamiento_SPSS.spo
P0_Preprocesamiento_SPSS.pdf
```

Introducción

La Estadística es fundamental en minería de datos:

- Proporciona técnicas muy utilizadas en minería de datos (como, por ejemplo, el análisis de componentes principales, las técnicas de regresión o el análisis factorial).

- Sirve de filtro previo a la realización de distintos estudios de minería de datos. Por ejemplo, en un estudio que analiza qué variables son importantes para predecir el comportamiento de otra (clasificación), ¿hay variables correladas que se pudiesen suprimir antes de proceder a dicho estudio?

- Se utiliza como parte de las técnicas propias de minería de datos (p.ej. test de la Chi cuadrado como medida de implicación entre dos ítems de una regla de asociación).

A la hora de aplicar técnicas estadísticas, hemos de tener en cuenta lo siguiente:

1. Las técnicas estadísticas suelen requerir que el experto diga exactamente lo que quiere comprobar.

2. Cuando se aplican técnicas estadísticas "clásicas", hay que tener cuidado de que se cumplan ciertos "requerimientos" o "hipótesis de partida". En caso contrario, hay que aplicar técnicas "no paramétricas".

SPSS

En esta práctica utilizaremos SPSS debido a su uso generalizado en la realización de estudios de tipo estadístico.

- SPSS es un paquete software para realizar análisis estadísticos.

- SPSS utiliza menús descriptivos y cuadros de diálogo simples para realizar las funciones solicitadas por el usuario.

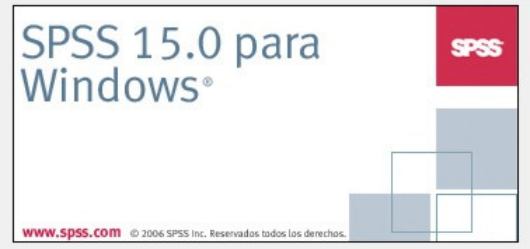
- SPSS ofrece la posibilidad de ejecutar una serie de comandos especificados en los denominados ficheros de sintaxis.

- SPSS posee una estructura tipo modular.

- El módulo base forma el núcleo del sistema e incluye, tanto comandos de lectura y transformación de datos y ficheros, como procedimientos estadísticos básicos.

- En estas prácticas, utilizaremos como ejemplo la versión 15.0.

Ejecución de SPSS

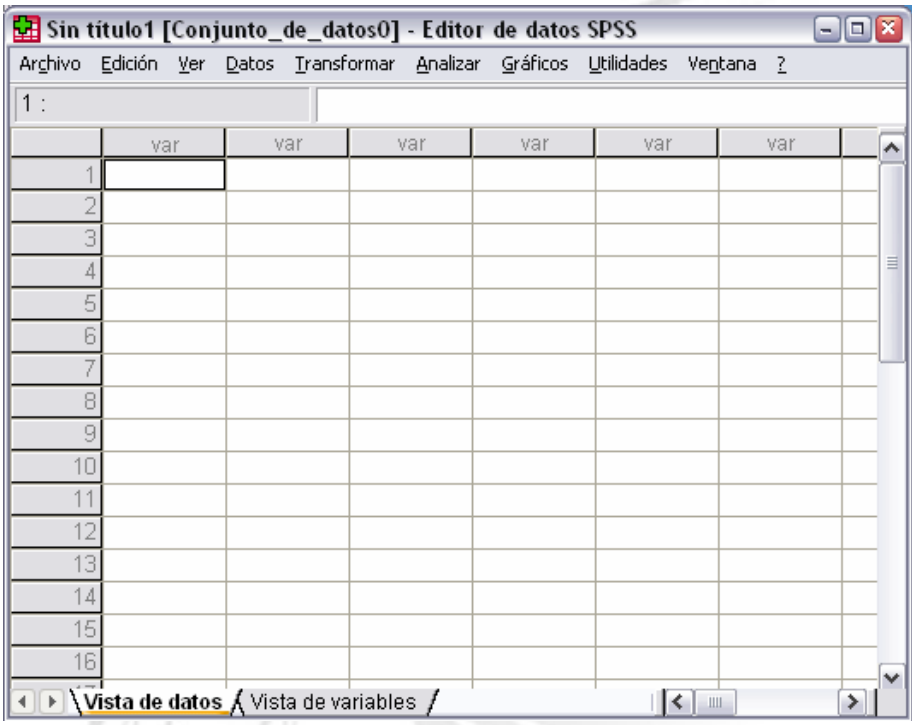


Al ejecutar el programa desde el menú de inicio, se muestra una ventana desde la que se nos ofrecen diversas opciones para abrir ficheros de datos, introducir nuevos datos o ejecutar un tutorial.

Se pueden crear ficheros de datos nuevos, importar hojas de cálculo desde Excel, bases de datos desde Oracle o leer ficheros de texto (en formato CSV, por ejemplo)

NOTA: La extensión de los ficheros con los que trabaja SPSS es `.sav`.

 Cargue "*Datos de Empleados*"



SPSS nos ofrece una vista de datos y una vista de variables (tipo hoja de cálculo) - desde la versión 10 en adelante -

Definición de variables

En SPSS, los nombres de las variables no pueden tener más de 8 letras, pero se les puede poner una etiqueta más larga que luego saldrá en los gráficos (columna *Etiqueta*).

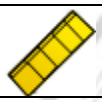
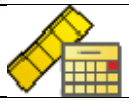
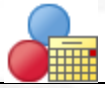
A la hora de declarar variables, es muy importante escoger adecuadamente la combinación Tipo de dato – Medida

Medidas (establecen qué mide la variable):

- Nominal: Una variable que toma valores no ordenados (p.ej. color de pelo).
- Ordinal: Una variable que toma valores ordenados (p.ej. nivel de satisfacción, medido de 0 a 5).
- Escala: Una variable que toma valores numéricos, para los que tiene sentido la operación de resta (p.ej. edad).

Tipos (establece cómo codificamos lo que la variable mide):

- Numérico (con una precisión determinada).
- Cadena (típica cadena de caracteres)
- Otros: Dólar, fecha, etc.

Nivel de Medida	**Tipo de datos**			
	Numérico	Cadena	Fecha	Tiempo
Escala		n/a		
Ordinal				
Nominal				

Ejemplos

- Color de pelo de una persona: Lo normal sería definir una medida nominal de tipo cadena, pero también podríamos usar una medida nominal con un tipo numérico (con 1 dígito de precisión para codificar los colores: 0 para el rojo, 1 para el negro, etc.)

- Ingresos de una persona: Medida de escala, tipo numérico.

- Grado de satisfacción del usuario: Medida ordinal, tipo cadena ("bajo", "alto", "medio") o numérico (0,1,2).

- Sexo: Medida nominal, tipo cadena ("hombre", "mujer") o numérico (1, 2).

- Categoría laboral: Si consideramos que existe una jerarquía en la que es más ser directivo que administrativo, usaríamos una medida ordinal, y un tipo cadena o numérico.

  NOTA: Usualmente, a un tipo de cadena siempre le pondremos una medida nominal, pero también podríamos asignarle una medida ordinal (consistente en utilizar el orden lexicográfico, si bien no es demasiado es usual).

Puede que queramos restringir los posibles valores que pueda tomar una variable.

Ejemplo

Si usamos el tipo cadena con 1 único carácter para la variable Sexo, podemos desear que sólo pueda tomar los valores "h" y "m". Para ello, usaremos la columna de valores en la vista de variables. Obsérvese que, por una parte, aparecen los valores ("h", "m") que deben corresponder al tipo de la variable y, por otra parte, figuran las etiquetas de los valores (las cadenas de caracteres que luego aparecerá en los resúmenes e informes que SPSS genere).

Observe la categoría laboral. Lo más habitual sería una medida ordinal con un tipo de cadena con valores "d", "a", "s" (o, incluso, "directivo", "administrativo", "seguridad"). Sin embargo, se utiliza un tipo numérico dado que algunos tests estadísticos necesitan que la variable sea numérica para poder trabajar con ella (aunque corresponda a una medida ordinal y no de escala). Observe que, como posibles valores, tiene {1, 2, 3} pero luego, como etiquetas, tiene "Administrativo", "Seguridad", "Directivo".

Análisis exploratorio de datos: Gráficos y Estadística Descriptiva

Un buen punto de partida para el análisis exploratorio de datos es echar un vistazo por separado a cada una de las variables que describen nuestro conjunto de datos. Esto nos permitirá conocer características básicas de nuestros datos que nos serán de gran en análisis posteriores. Para ello, usaremos estadísticos básicos y gráficos. Dependiendo del tipo de variable, usaremos unas técnicas u otras:

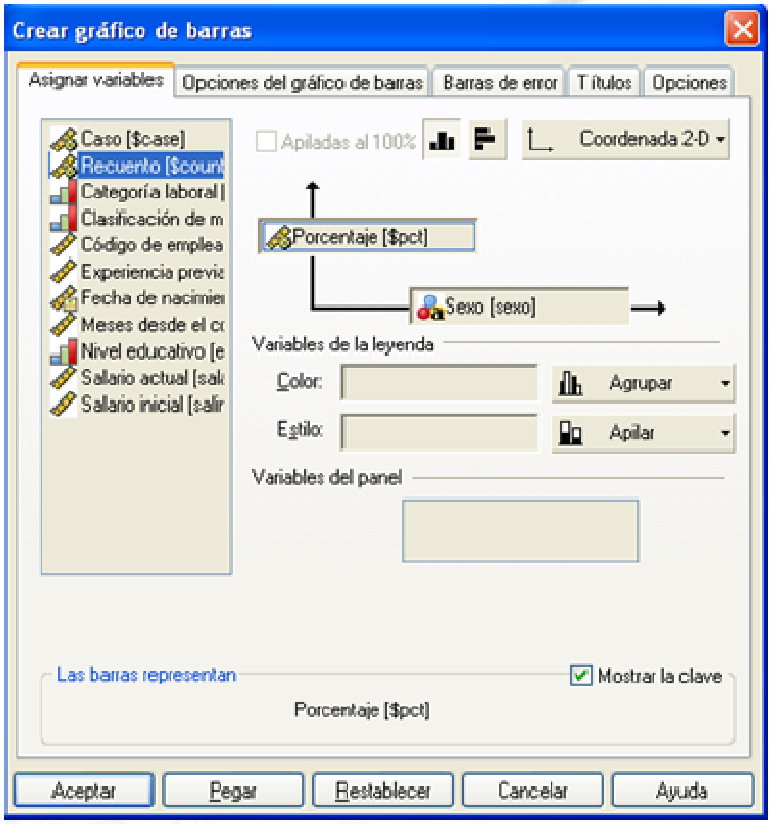
Sobre una variable nominal

Queremos responder a la pregunta: ¿cómo se distribuye una variable nominal?

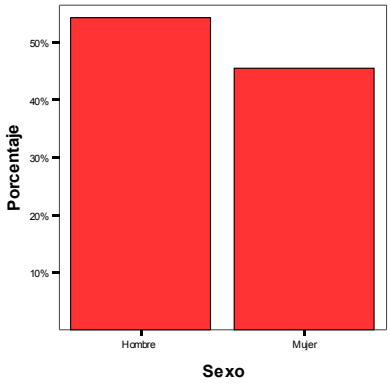
SPSS no ofrece demasiadas facilidades para las variables nominales (por ejemplo, que liste automáticamente los valores distintos). Para ello, tendremos que construir un gráfico de puntos o de barras y verlo en él.

Gráficos / Interactivos / Barras

Arrastramos con el ratón la variable *Sexo* al cuadro del eje de abscisas y *Porcentaje* al del eje de ordenadas.



Aparece el visor de resultados:

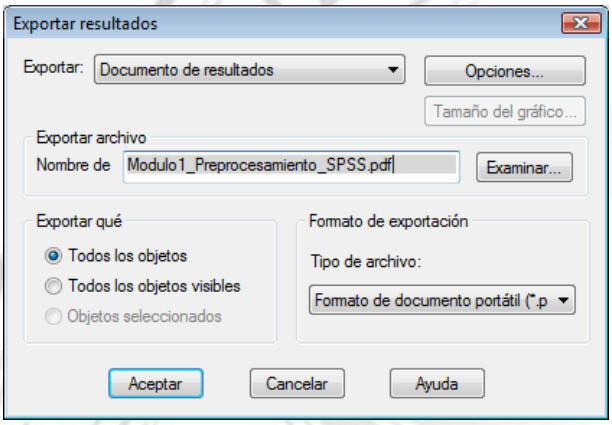


Al hacer doble click sobre el gráfico anterior, se abre un marco interactivo en el que podemos editar y cambiar algunos de los elementos del gráfico, cambiar las variables indicadas o incluso añadir cajas de texto con nuestros propios comentarios (*Insertar* > *Nuevo Texto*).

Todos los análisis que se vayan realizando se van guardando en el mismo sitio. Si quisiésemos suprimir cualquier elemento del visor de resultados, basta con borrarlo del panel izquierdo. El contenido del visor se guarda en un fichero con extensión `spo`.

A partir de ahora, todos los resultados que se obtengan como resultado de la ejecución de los análisis de este guión, se guardarán en un fichero con nombre `P0_Preprocesamiento_SPSS.spo` que habrá que entregar.

Cuando se indique "analice" o "comente" el resultado, habrá que añadir una caja de texto con la discusión pertinente. Una vez completados todos los ejercicios, habrá que crear el fichero `P0_Preprocesamiento_SPSS.pdf`, que también habrá que entregar. Este fichero se crea desde *Visor de Resultados* > *Archivo* > *Exportar*.



Una vez que hemos visto cómo obtener una representación gráfica, veamos algunos estadísticos que nos informen de cómo es la muestra. Para una variable de tipo nominal, no hay mucha información que ofrecer: su moda, las frecuencias relativas de los distintos valores y poco más (obviamente, la media no tiene sentido, por ejemplo).

- **Gráficos**: *Gráficos de barras / Porcentajes*

  (nos muestra el gráfico anterior si seleccionamos *Sexo*).

- **Estadísticos**:

  Aunque puede marcarse, ningún estadístico aparece en el resultado si seleccionamos *Sexo* (ni siquiera la moda). Esto ocurre porque se definió con el tipo de cadena de caracteres, a pesar de que la moda (el valor que más se repite) sería un estadístico perfectamente aplicable a *Sexo* :-(

  Si hacemos lo mismo con *Categoría Laboral*, ahora sí puede verse la moda y los demás estadísticos, ya que se usó un tipo numérico para representar dicha variable (que es de medida nominal).

Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hombre	258	54,4	54,4	54,4
	Mujer	216	45,6	45,6	100,0
	Total	474	100,0	100,0	

Estadísticos

Sexo

N	Válidos	474
	Perdidos	0

Estadísticos

Categoría laboral

N	Válidos	474
	Perdidos	0
Moda		1

Sobre una variable de escala

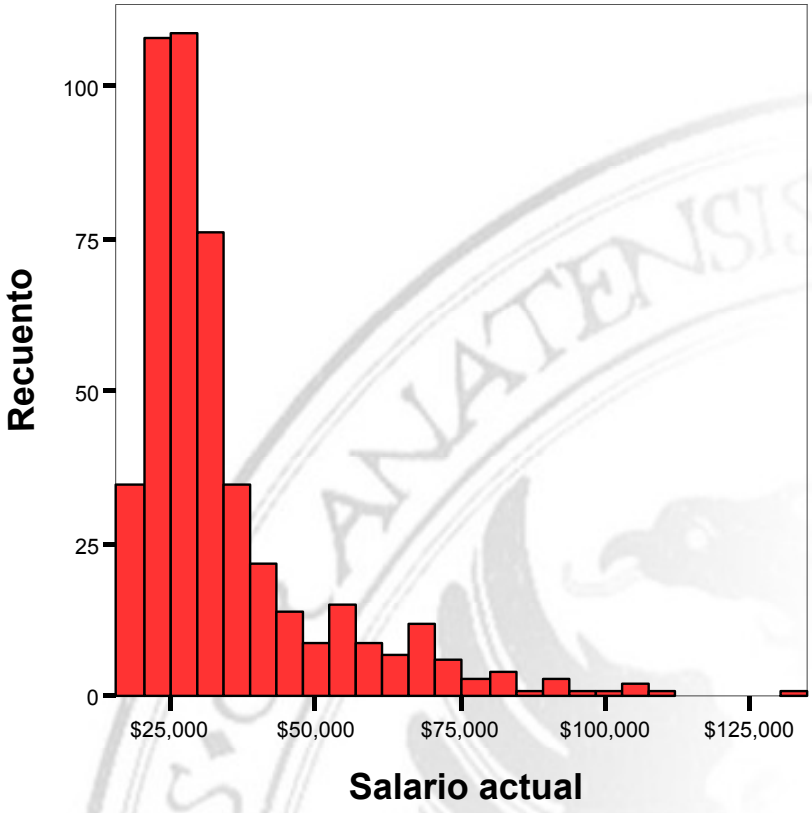
Queremos responder a la pregunta: ¿Cómo se distribuye una variable numérica?

Para las variables de escala, usaremos un **histograma**. Los histogramas de frecuencias proporcionan una forma de visualizar distribuciones para una sola variable. Para construirlos, se divide el rango entre el menor y el mayor valor de la variable en intervalos del mismo tamaño y se representa en ordenadas el número de casos cuyo valor de la variable está contenido en el intervalo correspondiente (usualmente, mediante una barra).

¿Cómo se distribuye el salario entre los empleados?

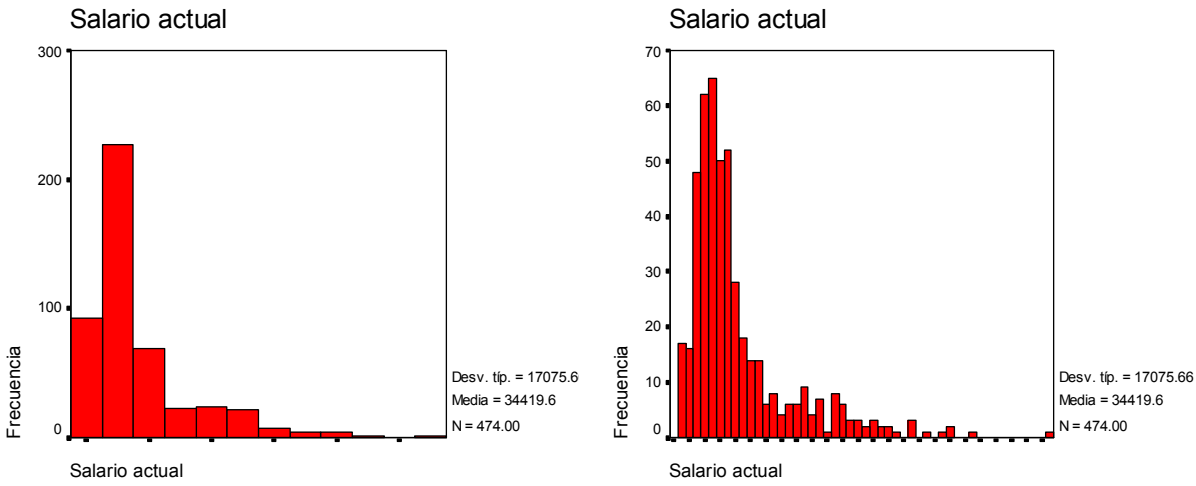
 *Gráficos / Interactivos / Histograma*

Seleccione *Salario Actual* en las abscisas y *Recuento* en las ordenadas:

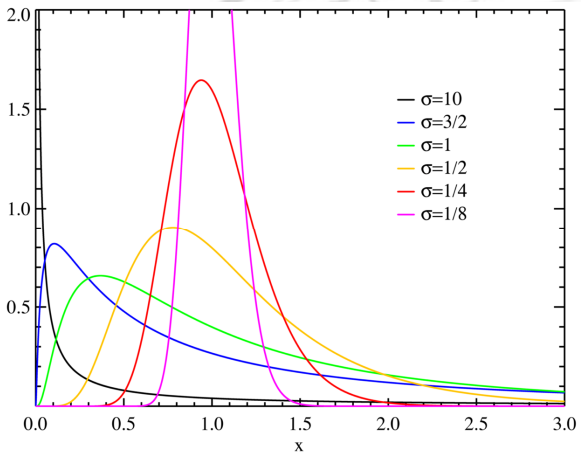


- Comente el histograma (*Insertar* > *Nuevo Texto*). ¿Qué se puede apreciar?

El ancho de los intervalos, que determina su número, afecta a la información que muestra el histograma y, en particular, puede afectar a su apariencia. Cambiando esta característica desde la pestaña *Histograma*, podemos obtener información más precisa y detallada. Por ejemplo, podemos partir de pocos rectángulos e ir introduciendo un mayor grado de detalle progresivamente si lo consideramos necesario. Para ello, una vez generado el histograma, seleccione haciendo click con el botón derecho del ratón la opción "*Herramientas para intervalos*":



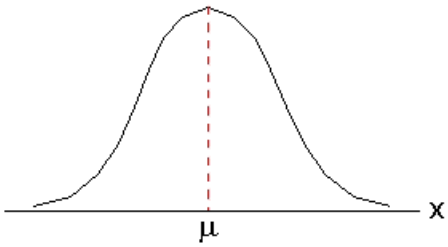
Si extrapolamos el histograma, obtendríamos una función matemática que determinaría la probabilidad con la que se da cada valor. En Estadística, se han estudiado muchas distribuciones de probabilidad. En el caso de la anterior, su forma se asemeja a una distibución log-normal:



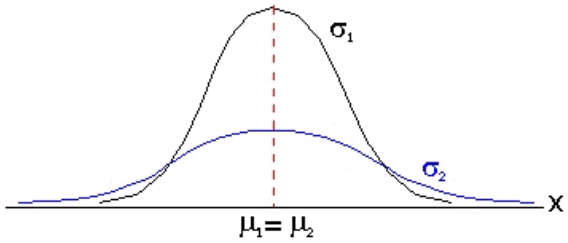
$$f(x;\mu,\sigma) = \frac{e^{-(\ln x-\mu)^2/(2\sigma^2)}}{x\sigma\sqrt{2\pi}}$$

Otra distribución muy conocida es la distribución normal:

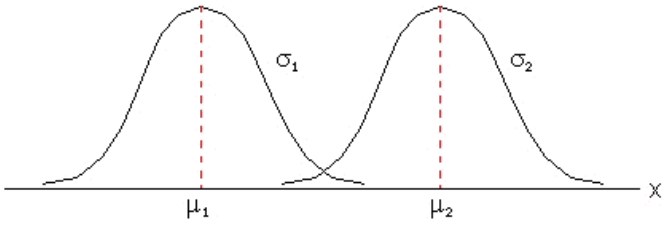
$$N(\mu,\sigma) \quad f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$



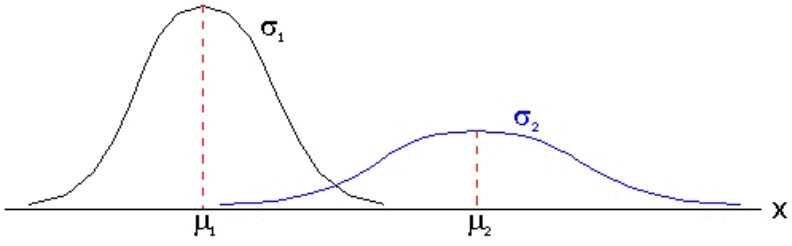
The normal curve



The normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$.



The normal curves with $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$.



The normal curves with $\mu_1 \neq \mu_2$ and $\sigma_1 < \sigma_2$.

Estadísticos de localización y dispersión

Para obtener una visión global de la distribución, utilizaremos medidas de resumen. Estas medidas de resumen se calculan a partir de los propios datos de la muestra y se denominan estadísticos.

Un estadístico básico es el tamaño de la muestra:

- **Tamaño de la muestra** (N): El número de casos en la muestra.

  Para nuestro conjunto de datos de ejemplo, el valor de N para administrativo, seguridad y directivo es 363, 27 y 84, respectivamente.

Aparte del tamaño muestral, hay dos tipos importantes de estadísticos:

- Los **estadísticos de localización** dan una idea de cuáles son los valores habituales de la distribución (en cierto modo, nos dicen dónde la distribución es más densa).
- Los **estadísticos de dispersión** dan una idea de cuál es la variabilidad en los datos.

Estadísticos de localización:

- **Media muestral**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  Problema: Sensible a casos aislados (outliers).

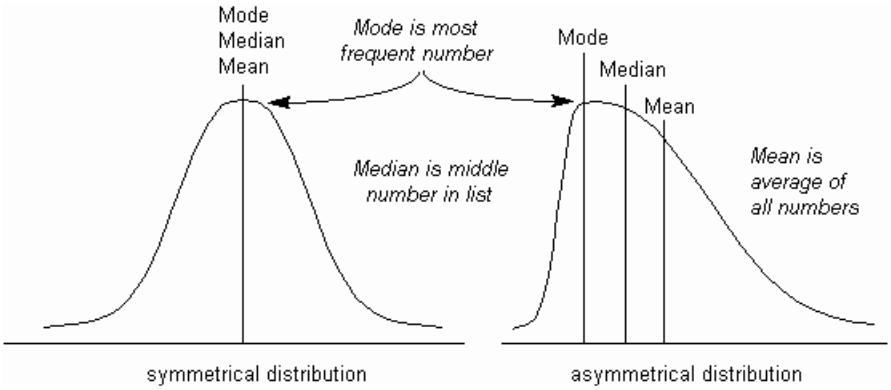
- **Mediana muestral**: Valor central de la lista ordenada de valores. El 50% de los valores están a su derecha y el otro 50% a la izquierda.

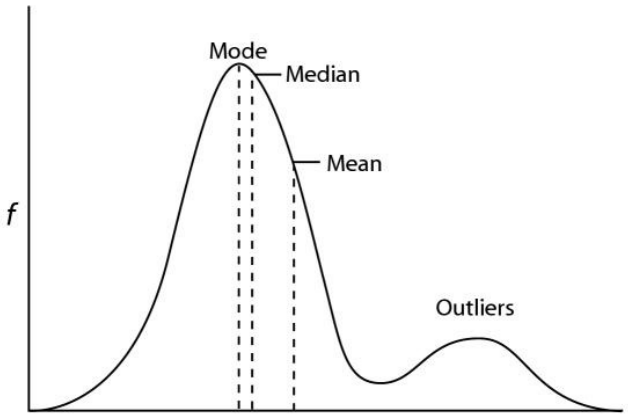
  Cálculo: Se ordenan todos los valores y se escoge el central. Si el número de valores es par, se toma la media aritmética de los dos valores centrales.

  Ventaja: Menos sensible a outliers.

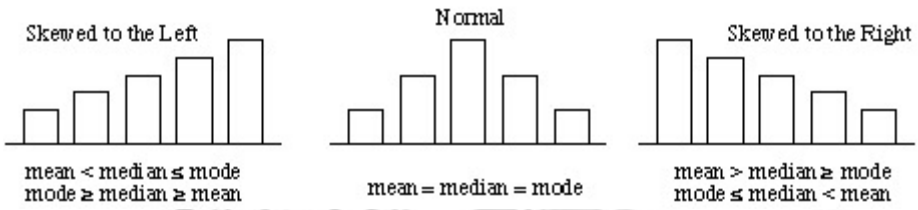
- **Moda muestral**: Valor más común.

  Si la distribución tiene una única moda, se dice que la distribución es unimodal. En ocasiones, no obstante, una distribución tiene más de una moda (distribución multimodal).





Summary
Typical Relationships Between Mean, Median and Mode
For Three Special Distributions

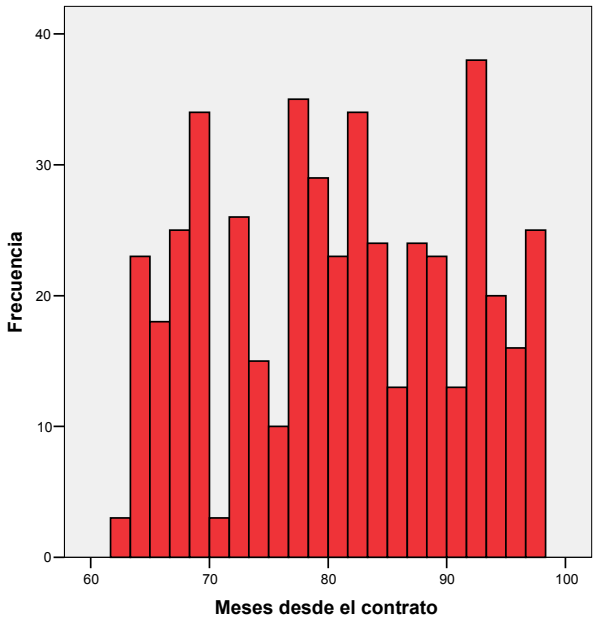


- **Desviación típica muestral**: Es una medida global que representa cómo de dispersos están los datos con respecto a su media aritmética.

$$S = +\sqrt{\frac{1}{n-1}\sum (X_i - \overline{X})^2}$$

Para una amplia mayoría de distribuciones, la mayor parte de los valores están comprendidos entre 2 desviaciones de la media (media ± 2 S) y el 70% de los casos están a una distancia de la media no mayor a 1 desviación típica.

- La **varianza** es el cuadrado de la desviación típica

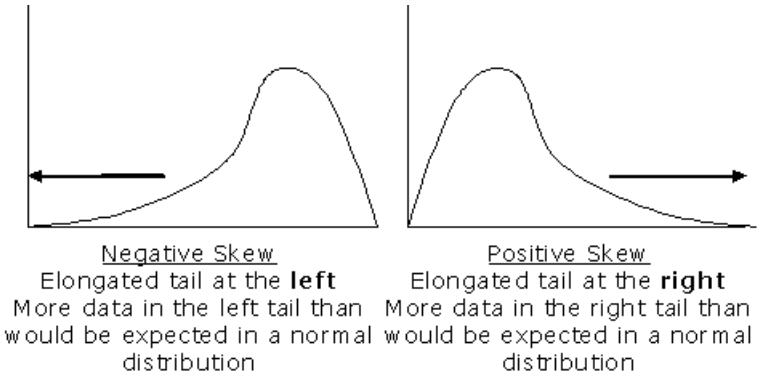


Ejemplo: Una distribución con una varianza elevada

Ampliación: Estadísticos de forma

Otros estadísticos dan una idea de cuál es la forma de la distribución:

- El **skew (asimetría)** representa si hay más datos en la parte superior de la distribución o en la parte inferior:



  Se define formalmente como el tercer momento tipificado de la distribución, aunque Pearson dio una forma aproximada de calcularlo como

  (media - moda) / desviación típica

  La distribución normal es simétrica, por lo que tiene un valor de asimetría 0. Una distribución que tenga una asimetría positiva significativa tiene una cola derecha larga. Una distribución que tenga una asimetría negativa significativa tiene una cola izquierda larga. Un valor de asimetría mayor que 1, en valor absoluto, indica generalmente que una distribución difiere de manera significativa de la distribución normal.

- La **kurtosis (curtosis)** es una medida del grado en que las observaciones están agrupadas en torno al punto central.

  Para una distribución normal, el valor del estadístico del coeficiente de curtosis se suele definir de forma que valga 0 (distribución mesocúrtica).

  Una curtosis positiva indica que las observaciones se concentran más en torno a la media que las de una distribución normal (distribución leptocúrtica).

  Una curtosis negativa indica que las observaciones se agrupan menos en torno a la media que las de una distribución normal (distribución platicúrtica).

NOTA: En Estadística, el momento central o centrado de orden k de una variable aleatoria X es la esperanza matemática $E[(X - E[X])^k]$ donde E es el operador de la esperanza. El primer momento central es cero y el segundo se llama varianza (σ^2) donde σ es la desviación estándar. Los tercer y cuarto momentos centrales sirven para definir los momentos estándar denominados de asimetría y de curtosis.

¿Cuál es la media aritmética del salario de los empleados?
¿Qué dispersión ó varianza presenta el salario entre los empleados?

 *Analizar / Estadísticos Descriptivos / Frecuencias*

Seleccione *Salario Actual*, quite "*Mostrar tablas de frecuencias*", en gráficos seleccione *Histograma* y, en *Estadísticos*, media, desviación típica, mínimo y máximo.

Si aparece **********, tendrá que agrandar convenientemente la tabla de resultados.

Con apenas cuatro valores de resumen, nos podemos hacer una idea muy aproximada de cuál es la distribución de los datos. La media está en torno a los $34.400. La mitad de los trabajadores ganan menos de $28.875 y la otra mitad gana más.

Resúmenes de casos

	Salario actual	Experiencia previa (meses)
N	474	474
Media	$34,419.57	95,86
Desv. típ.	$17,075.661	104,586

En cuanto a la variabilidad, el 70% de los individuos tienen un salario en el intervalo
[$34.419 - $17.075, $34.419 + $17.075] ≈ [$17.500, $51.500]
y la mayor parte (95%) de los individuos tienen un salario en el intervalo
[$34.419 – 2*$17.075, $34.419 + 2*$17.075] ≈ [$300, $68.500]

 Realice el mismo análisis (descriptivo y gráfico) con las variables *Salario Inicial* y *Meses desde el contrato*. Comente los resultados

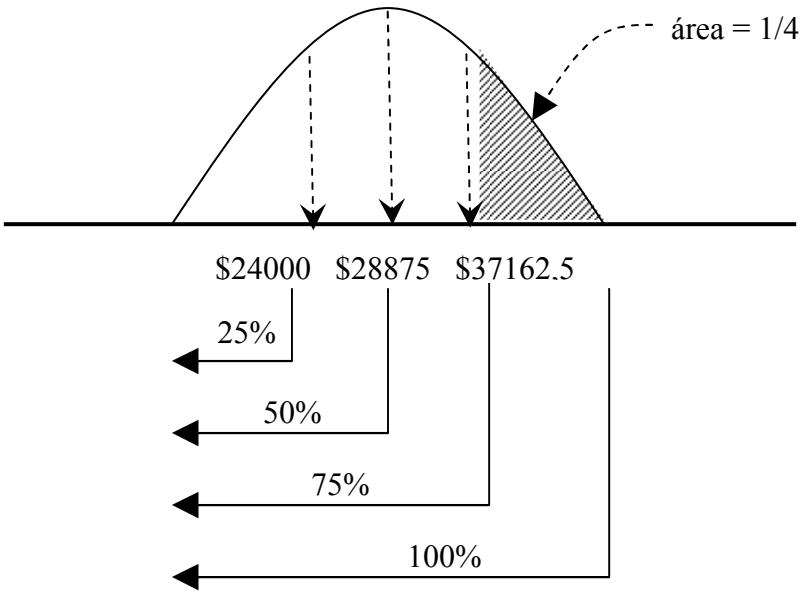
 Abra el fichero `Mundo95.sav`, elija un par de variables cualesquiera y realice el mismo análisis (descriptivo y gráfico). Comente los resultados.

NOTA: Desde SPSS, se pueden obtener los mismos estadísticos desde distintos sitios (confunde un poco el hecho de que SPSS permita realizar los mismos análisis desde distintos menús):

- *Analizar / Informes / Resúmenes de casos*
- *Analizar / Estadísticos Descriptivos / Descriptivos*
- *Analizar / Informes / Resúmenes de Casos*

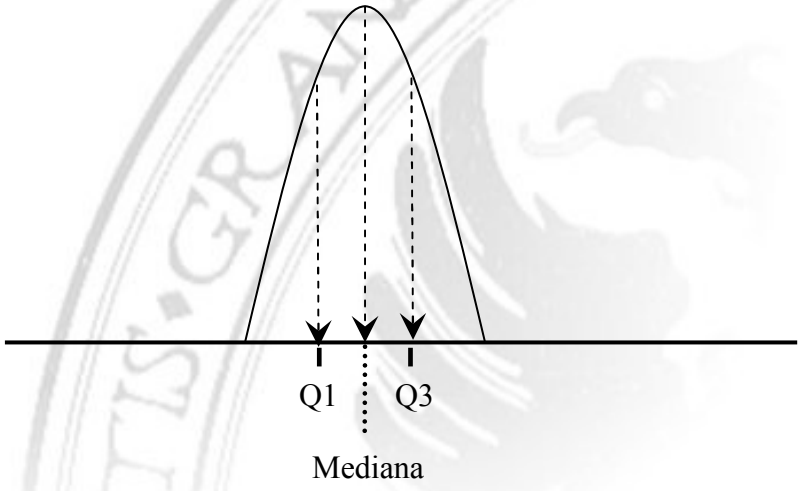
Los **percentiles** son unos estadísticos de tendencia central, pero que también ofrecen información sobre la dispersión de los datos.

El percentil 25 es un valor tal que el 25% de los valores de la muestra son menores que él (obviamente, el percentil 50 es la mediana).



Estos percentiles (25-50-75) se denominan **cuartiles**.

Si dos cuartiles estuviesen muy próximos (imaginemos $27500 y $27900), esto indicaría que un 25% de la muestra tiene salarios muy parecidos, por lo que hay una elevada concentración de individuos en ese intervalo.



Analizar / Estadísticos Descriptivos / Frecuencias

Seleccione *Salario Actual* y, en *Estadísticos*, incluya los cuartiles:

Estadísticos

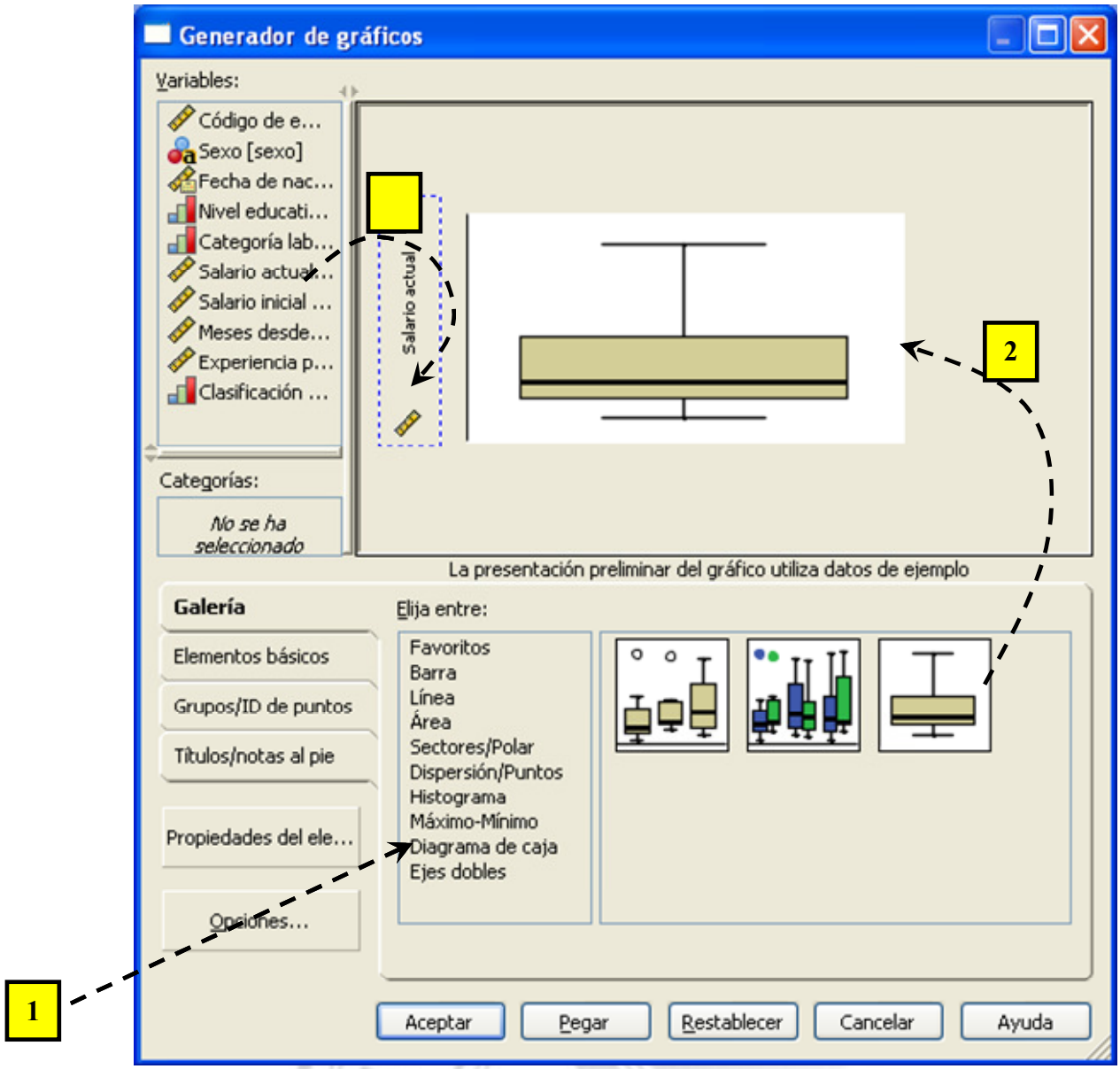
Salario actual

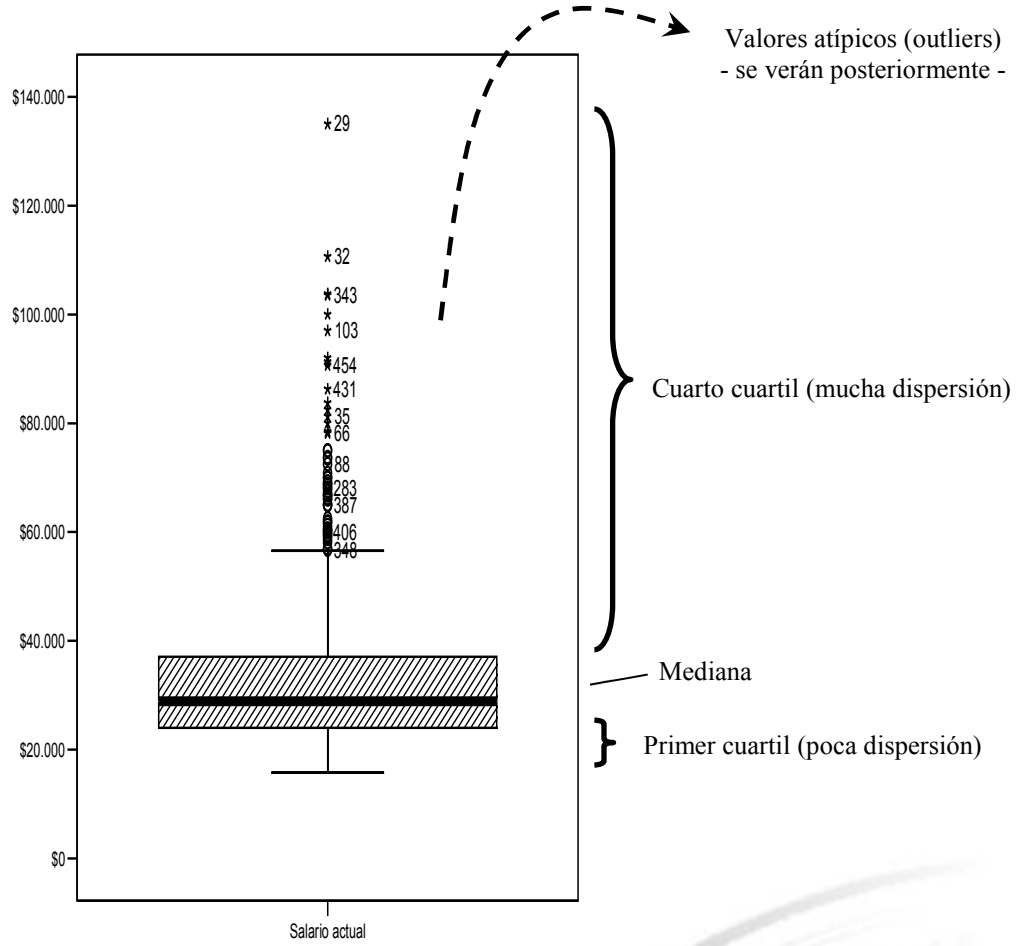
N	Válidos	474
	Perdidos	0
Media		$34,419.57
Mediana		$28,875.00
Moda		$30,750
Desv. típ.		$17,075.661
Asimetría		2,125
Error típ. de asimetría		,112
Mínimo		$15,750
Máximo		$135,000
Suma		$16,314,875
Percentiles	25	$24,000.00
	50	$28,875.00
	75	$37,162.50

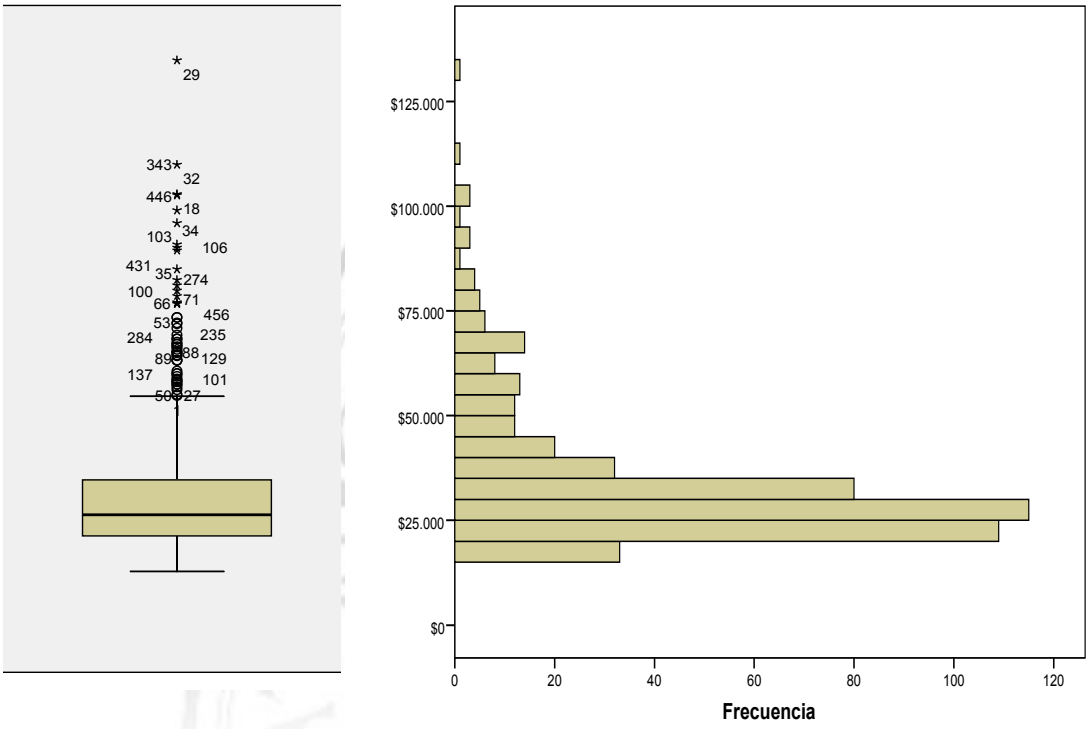
Diagramas de caja [box plots]

Una forma de representar gráficamente los cuartiles:

Gráficos / Generador de gráficos







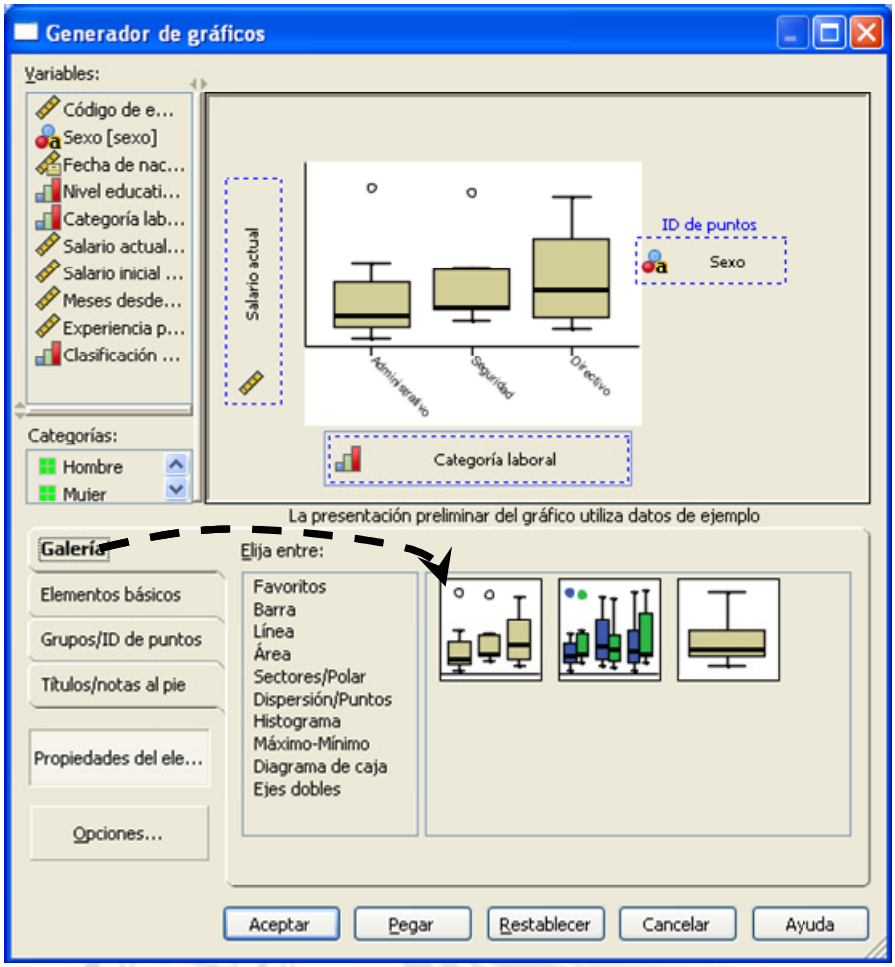
Se puede apreciar que la mitad de los empleados ganan entre $15000 y $30000 mientras que en la otra mitad hay mucha más variación de salarios (entre $30000 y $140000).

Incluya otro gráfico con la variable "*Experiencia Previa*". Aquí se verá que la mitad de los datos están agolpados en un intervalo de valores muy pequeño, mientras que la otra mitad está mucho más dispersa.

Si utilizmos la variable "*Meses desde el contrato*", no hay apenas dispersión (véase el histograma correspondiente)

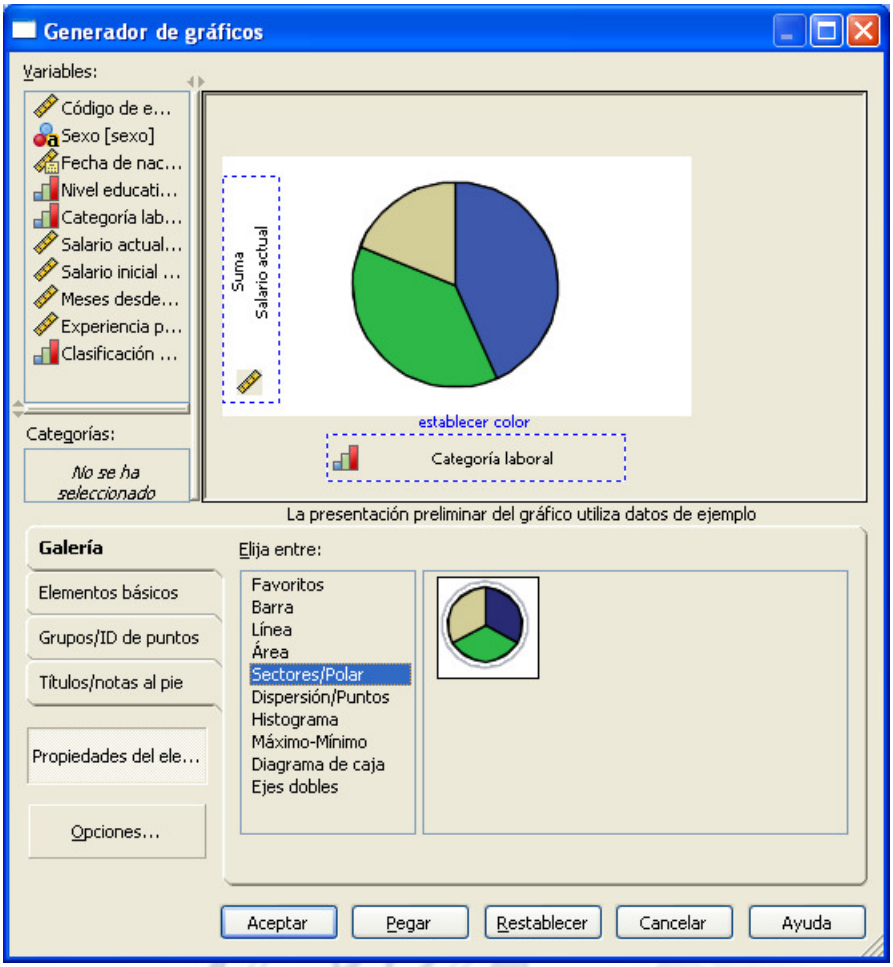
En ocasiones, nos interesará utilizar una **variable de agrupación**.

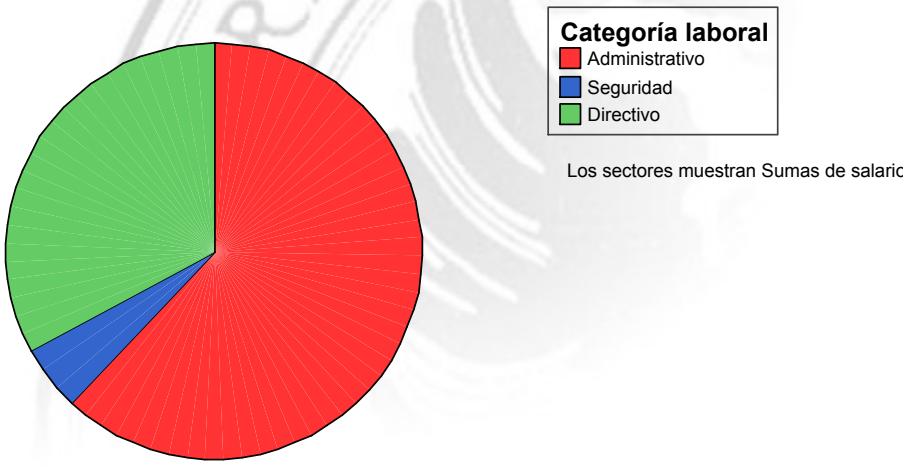
Por ejemplo, podría interesarnos ver la distribución del *Salario* para los administrativos en comparación con el de los directivos:



En los anteriores ejemplos, hemos trabajado siempre con el recuento de individuos (frecuencias o número de apariciones). A veces, sin embargo, puede que nos interese utilizar otra **medida de recuento**.

Esto nos permitirá, por ejemplo, responder preguntas del tipo: ¿cómo se reparte la nómina total de la empresa (la suma de las nóminas de todos los empleados) entre las distintas categorías laborales?





Preprocesamiento

Selección de datos

Datos / Segmentar Archivo

Permite seleccionar grupos de registros para que cualquier procedimiento de análisis de datos que se realice posteriormente se aplique de forma separada sobre cada uno de esos grupos.

Por ejemplo, si segmentamos el archivo en función del *Sexo*, cada vez que lancemos un análisis, éste se ejecutará sólo sobre los hombres y luego sólo sobre las mujeres.

Ejercicio: Sobre los *Datos de Empleados*, seleccione *Sexo* para segmentar los datos y construya un diagrama de cajas sobre el *Salario Actual*.

Datos / Seleccionar Casos

Permite excluir de los análisis conjuntos de registros. Podemos incluir una variable numérica de filtro (excluiría los registros que en dicha variable tengan un valor perdido o igual a cero) o establecer una condición más compleja

Ejercicio: Sobre los *Datos de Empleados*, seleccione sólo aquellos individuos con categoría laboral 1 ó 2, y además con un valor de minoría distinto de cero: (catlab = 1 | catlab = 2) & (minoría ~= 0). Mantenga la segmentación según *Sexo* y construya un histograma del *Salario Actual* sobre estos individuos. Analice los resultados obtenidos.

IMPORTANTE
Una vez completado el ejercicio anterior, elimine la segmentación realizada (seleccione la opción "*Analizar todos los casos. No crear los grupos*").

NOTA:

Algunos modelos estadísticos y de minería de datos son sensibles a los valores desconocidos. Si vamos a realizar un análisis que involucre a una variable sobre la que hay registros con valores desconocidos, podemos excluirlos seleccionando la función MISSING(nombre de variable) y usando el conectivo lógico NOT (~)

Transformación de datos

Creación de nuevas variables

Transformar / Calcular

Construya una nueva variable, *incr_salario* que represente el incremento porcentual del salario de un empleado. Este incremento se calculará realizando la transformación siguiente:

100 * (Salario actual – Salario inicial) / (Salario inicial)

Ejercicio: Construya un diagrama de cajas sobre el incremento porcentual, agrupando por sexo, para ver si el incremento salarial se aplica de la misma forma a hombres y mujeres. Analice el resultado.

Discretización de variables

A veces, una variable presenta un nivel de detalle innecesario y complica la generación de modelos. Por ejemplo, en un estudio de la influencia del sexo en el *Salario Actual*, no nos importa demasiado precisar este último hasta el último céntimo.

También hemos de asumir que los datos están sujetos a variaciones por el propio método de recopilación de datos (ruido, errores de medida, etc.).

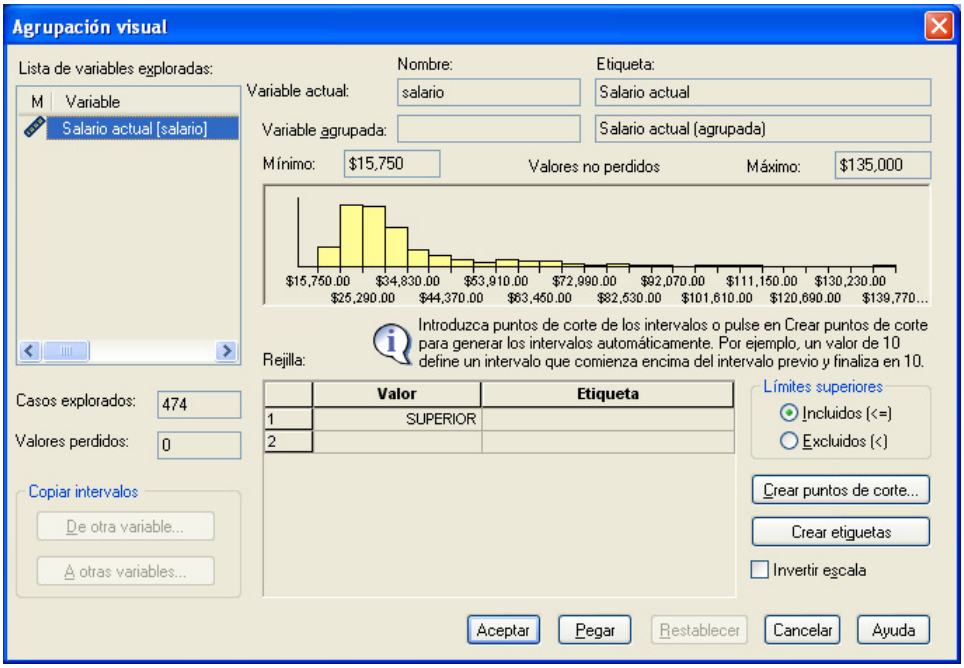
En casos así, podríamos estar interesados en crear otra variable con una discretización de ésta (incluso podemos eliminar los datos originales y quedarnos únicamente con los datos discretizados).

NOTA: Este proceso de discretización será necesario como paso previo para las variables continuas involucradas en el análisis de datos utilizando técnicas como las reglas de asociación.

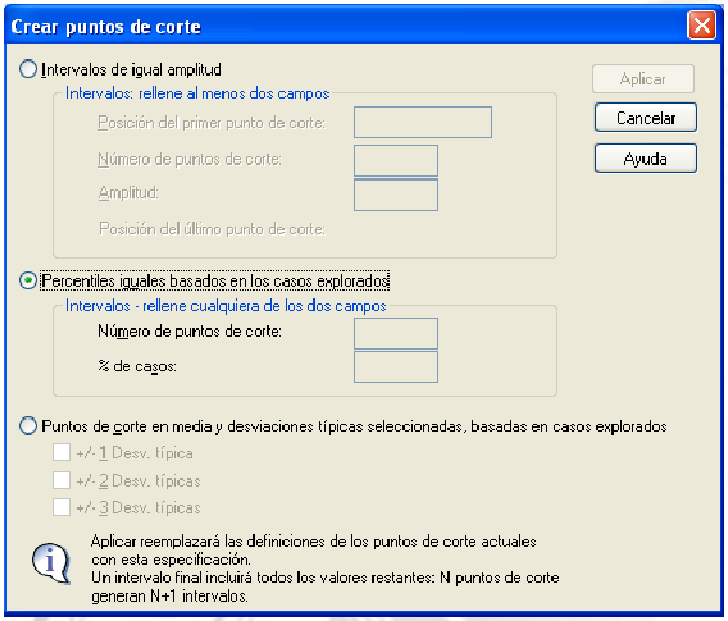
Transformar / Recodificar en distintas variables (manual)

Pinchando en "*Valores antiguos y nuevos*", posteriormente se irían indicando manualmente las transformaciones de los datos

Transformar / Agrupación visual



Pinchamos en *Crear puntos de corte…*



- *Intervalos de igual amplitud* (discretización *equi-width*): No es demasiado usual elegirlo, ya que algunos intervalos pueden tener muchos datos y otros no.
- *Percentiles iguales basados en los casos explorados* (discretización *equi-depth*): Cada intervalo contendrá el mismo número de datos, por lo que la amplitud de losintervalos será distinta.

NOTA: Posteriormente, los puntos de corte se pueden desplazar manualmente en el histograma que aparece en la ventana de *Agrupación Visual*. Finalizado el proceso, pinchamos en *Crear Etiquetas* y *Aceptar*.

Ejercicios de discretización

Cree la variable nueva *Salario Agrupado*, construida a partir del *Salario Actual* utilizando el método de discretización equi-depth con 5 intervalos.

Cree un diagrama de puntos (a través del menú de Gráficos Interactivos) en el que la abscisa represente la variable nominal *Salario Agrupado* y seleccione para el eje de ordenadas la variable *Sexo*. Lo que se mostrará será la moda de dicha variable. Interprete los resultados.

Seleccione ahora la *Experiencia previa* en el eje de ordenadas. Lo que se mostrará será la media aritmética de dicha variable. Interprete los resultados.

Ampliación

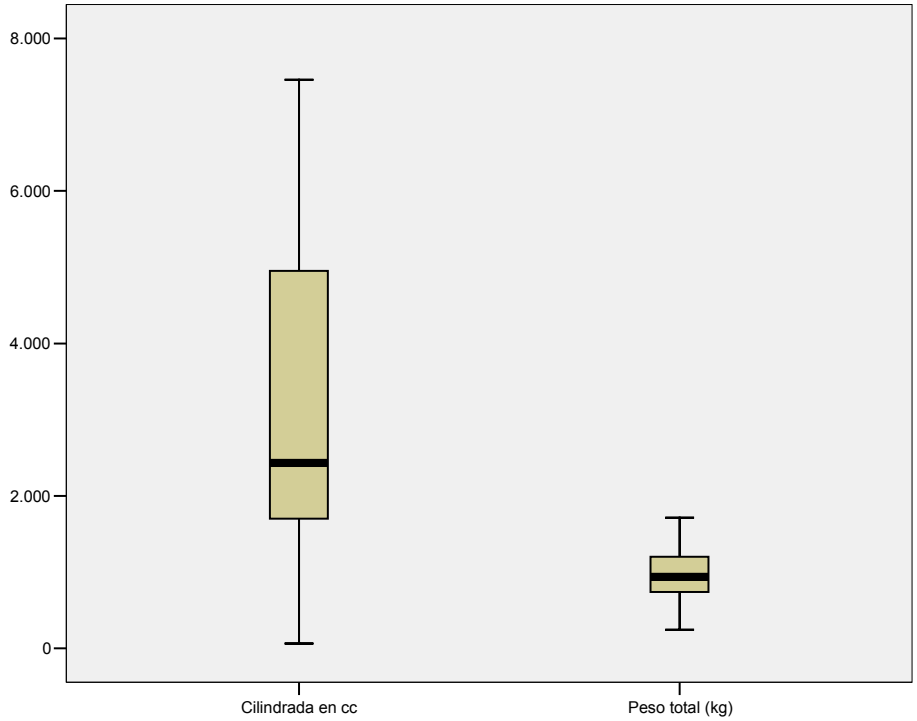
- Usando discretización, podemos detectar asociaciones que no sean lineales (las asumidas por las técnicas tradicionales de regresión) entre variables cuantitativas.

- En problemas de clasificación, se pueden usar métodos de discretización que utilizan la información de la variable que indica la clase para establecer los intervalos de discretización. Estos métodos se conocen como métodos de discretización supervisada. SPSS no ofrece ninguna opción que permita realizar este tipo de discretización.

Normalización de variables

Cargue el fichero `coches.sav`

En *Gráficos / Cuadros de diálogo antiguos / Diagramas de Caja*, seleccione *Simple*, *Resúmenes para distintas variables*, *Definir*, y elija como variables la la cilindrada y el peso total del vehículo:



¿Cuál de las dos variables presenta mayor dispersión?

Parece que la cilindrada. Sin embargo, el problema es que la escala usada en el eje de las ordenadas es la misma para las dos variables pero éstas tienen rangos de valores muy diferentes (se mueven en escalas distintas).

Para suprimir el efecto de la escala (algo muy importante en algunos análisis estadísticos y de minería de datos como, por ejemplo, en los modelos de clustering) es necesario transformar (normalizar) la variable en otra similar que guarde las mismas proporciones, pero en una escala estándar.

¿Qué significa guardar las mismas proporciones?

Que sea una transformación lineal, es decir, del tipo $Y=a+bX$

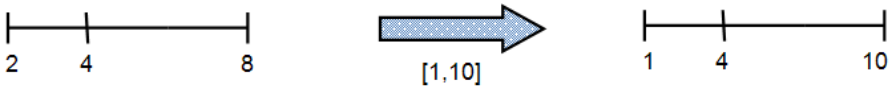
Algunas normalizaciones usuales:

- **Normalización [0,1]**

$$Y = \frac{X - mínimo\_original}{máximo\_original - mínimo\_original}$$

- **Normalización [min,max]**

$$Y = \min + \frac{X - mínimo\_original}{máximo\_original - mínimo\_original}(\max - \min)$$



- **Normalización z-score (tipificación),** muy utilizada en Estadística:

$$X \rightarrow Z = \frac{X - \overline{X}}{S}$$

Anteriormente, ya mencionamos que la mayor parte de las distribuciones toman valores comprendidos entre 2 desviaciones de la media (media ± 2 S). Si tipificamos la variable (transformación Z-score), tendremos que la nueva variable Z casi siempre tomará valores en el intervalo [-2,2].

Recordemos que la probabilidad de que una distribución normal N(0,1) tome valores en el intervalo [-1.96,1.96] es de 0.95, la variable tipificada utilizará como rango de referencia es el rango de valores de la distribución N(0,1).

Esta transformación permite expresar una variable en función de cuántas veces – en unidades dadas por la desviación típica – un valor dado está por encima o por debajo de la media.

Al realizar operaciones lineales con constantes (media y desviación típica), el resultado es otra distribución que guarda las mismas proporciones.

En SPSS habrá que seleccionar *Transformar / Calcular Variable* e introducir la fórmula de normalización (lo que me creará una variable nueva).

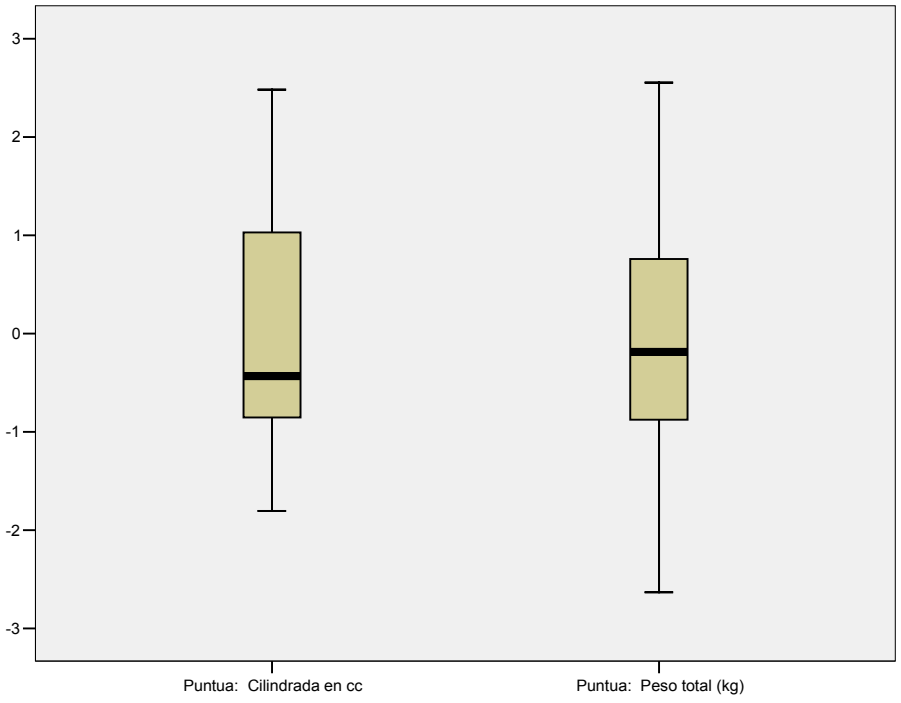
Para tipificar, podemos seleccionar directamente *Analizar / Estadísticos Descriptivos / Descriptivos / Guardar valores tipificados como variables.*

Ejercicios de normalización/tipificación

Sobre los datos de los coches, tipifique las variables cilindrada y peso.

Construya los diagramas de cajas con las variables tipificadas y analice los resultados en comparación con el diagrama que obtuvimos anteriormente.

Construya también los histogramas de las variables sin tipificar y tipificadas. Compare el histogramas de cada variable sin tipificar con el histograma de la misma variable tipificada. Analice el resultado obtenido.



OBSERVACIÓN:

Un caso donde es imprescindible la tipificación es en el cómputo de distancias entre puntos con métricas como la distancia euclídea. Dicha métrica tiene en cuenta la escala de medida, por lo que si queremos comparar dos coches en función de su peso y cilindrada, la primera variable influirá menos que la segunda, ya que el rango de la cilindrada es mucho mayor.

Al utilizar técnicas de clustering, será imprescindible trabajar con variables tipificadas.

Detección de outliers (valores atípicos)

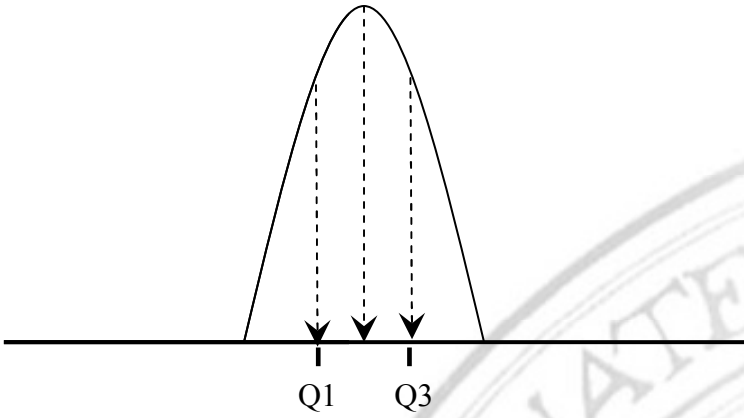
Un outlier es un valor anormalmente distante del resto de valores.

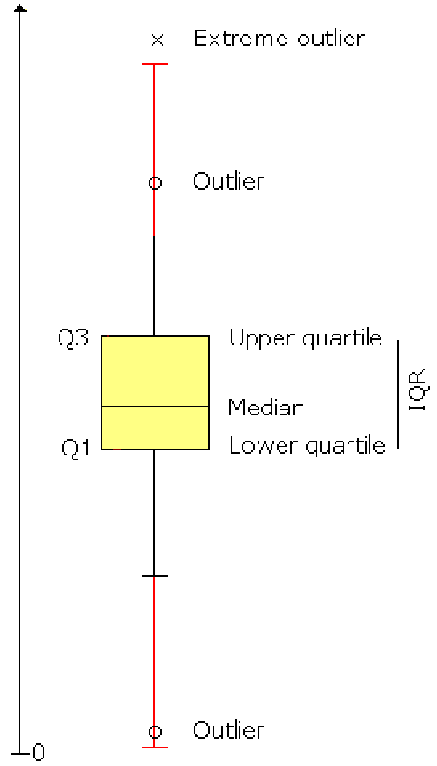
El filósofo Francis Bacon sentenció en 1620: *"Errors of Nature, Sports and Monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."* [*Novum organum*,o *Indicaciones relativas a la interpretación de la naturaleza*]

¿Qué criterio se utiliza para definir un valor anómalo?

- Si los datos se ajustan a una distribución estadística, serían los datos que hay más alejados de los valores centrales (los que hay en las colas). Se puede usar un test estadístico (por ejemplo, el de Grubb para la distribución normal)

- Si consideramos una única dimensión (un único atributo), sea cual sea la distribución, se considera que los valores anormales son los que están más alejados de la mediana:





IQR = Q3 - Q1

P es un Outlier si P > Q3 + 1.5 IQR
P es un Outlier si P < Q1 - 1.5 IQR

P es un Outlier Extremo si P > Q3 + 3 IQR
P es un Outlier Extremo si P < Q1 - 3 IQR

NOTA:
SPSS usa por defecto un círculo para los outliers y una estrella para los outliers extremos.

Si consideramos varias dimensiones, existen distintas aproximaciones:

- "Local Outlier Factor" da una puntuación de hasta qué punto un valor es un outlier (este tipo de técnicas se estudian en el Máster de Sistemas Inteligentes).

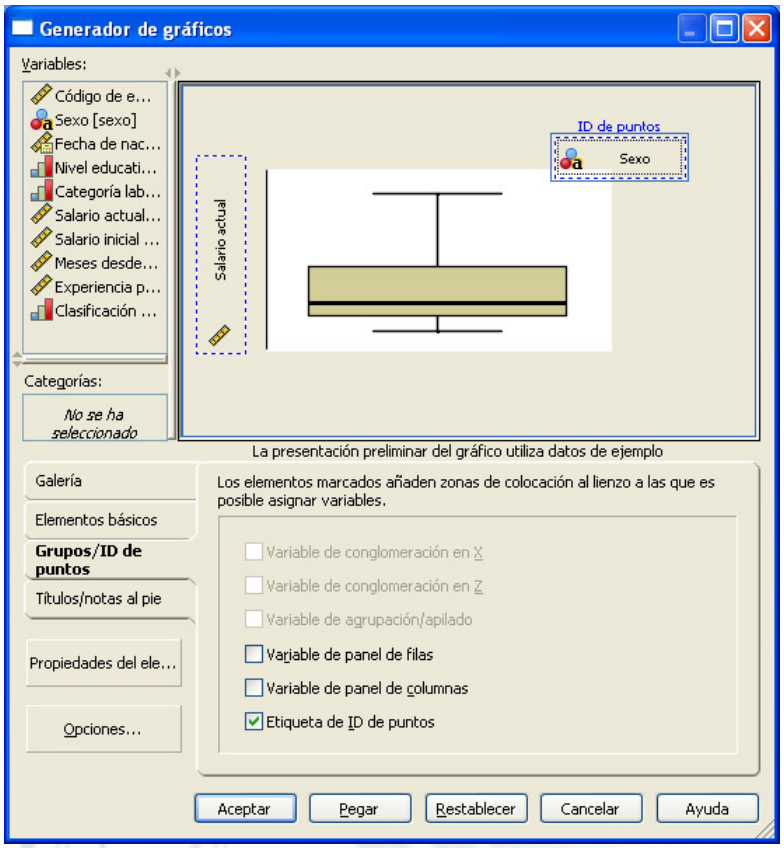
- Métodos de clustering (se verán posteriormente)

¿Qué hacer con los registros que presentan un outlier en alguno de sus atributos?

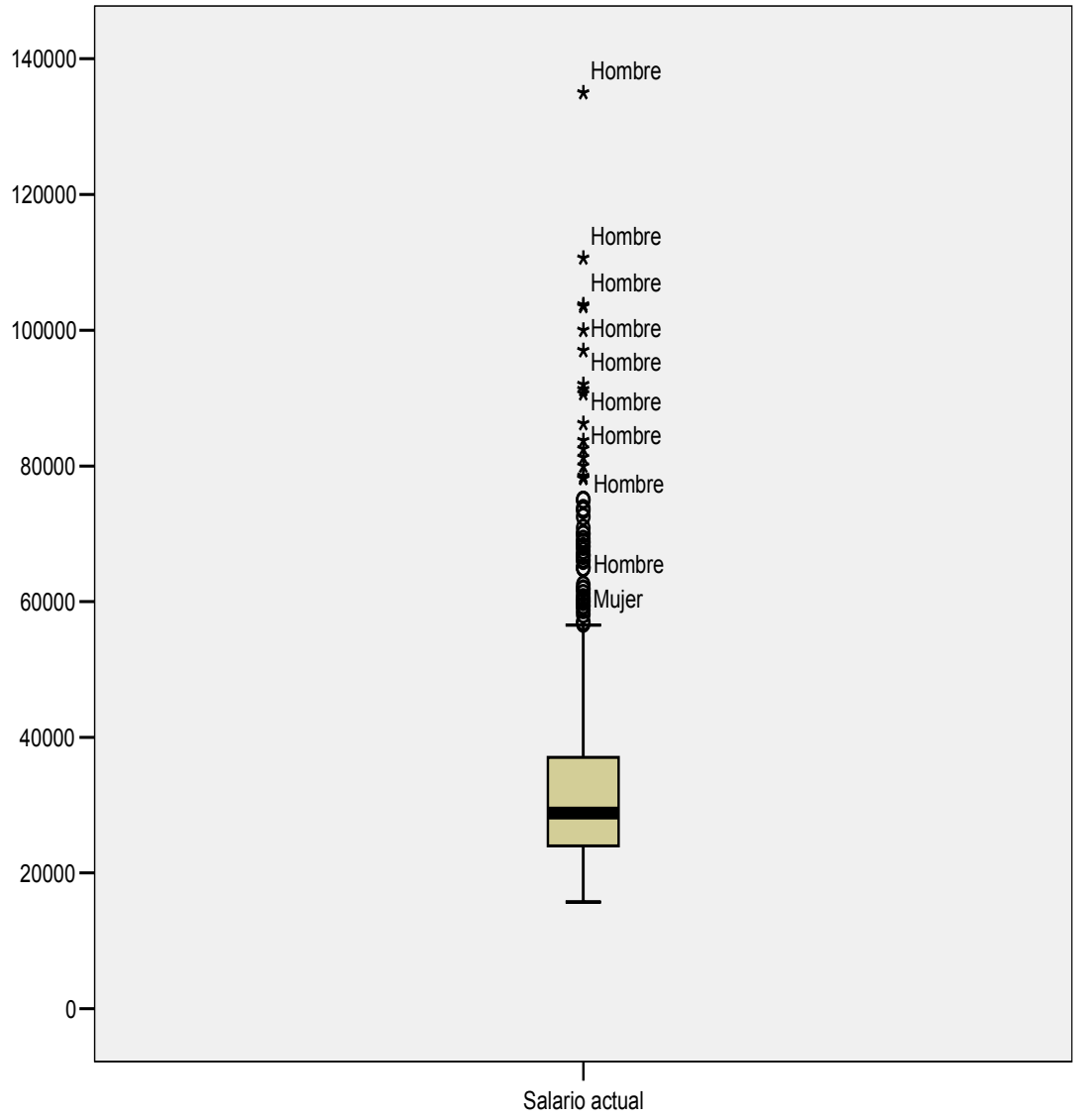
- En primer lugar, analizar si son registros que se pueden excluir del estudio. A veces, representan información interesante y otras veces no son más que errores de medida.

- Si la técnica estadística o de minería de datos que utilicemos lo permite, se pueden dejar dichos registros para que los procese la propia técnica. Si no es así, pueden excluirse del estudio correspondiente (utilizando técnicas de selección de datos)

Creemos de nuevo un BoxPlot sobre el Salario Actual:

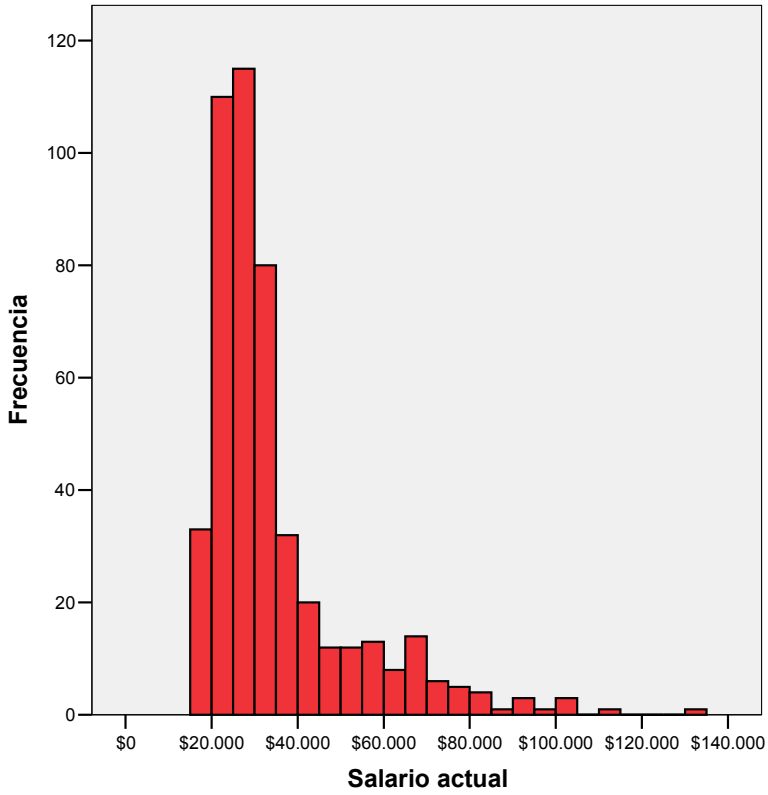


Etiquetando los valores atípicos en función del *Sexo*, se puede apreciar que los valores atípicos suelen corresponder a hombres (y siempre en sueldos altos).



Calcule los cuartiles del *Salario Actual* y seleccione aquellos casos correspondientes a registros que no tienen un valor atípico en este atributo (*Datos > Seleccionar casos > Si se satisface la condición*). En la ventana habrá que especificar la condición lógica de que el salario **no** sea un outlier (aplicando la fórmula con los valores de los cuartiles calculados anteriormente). Una vez hecho esto, genere el diagrama de caja correspondiente al *Salario Actual* y analice el resultado.

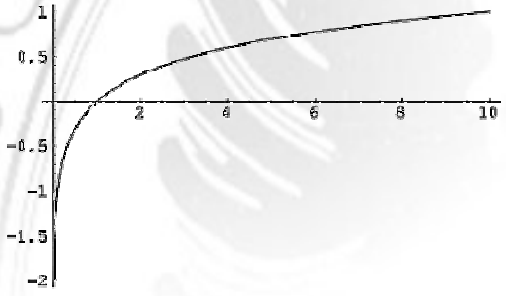
El problema de la definición de outliers aplicando la fórmula que utiliza la distancia intercuartil es que no funciona demasiado bien con distribuciones asimétricas, tal y como sucede con la del salario:



Ejercicio tipo B

Elimine la exclusión de los casos atípicos realizada en el ejercicio anterior (mediante *Datos > Seleccionar Casos > Todos los casos*). Añada ahora, manualmente, una copia de alguna tupla del conjunto de datos e indique en el *Salario Actual* el valor 11.000 (valor atípico por bajo). Dibuje el diagrama de cajas asociado y observe que no se detecta como outlier.

Una posible solución: Se aplica una transformación artificial a la variable *Salario Actual* para que marque más las diferencias en las zonas de alta densidad (la cola izquierda) y suavice las diferencias en las de baja densidad (la cola derecha). Una forma de hacerlo es usando la función logaritmo, que tiene la gráfica siguiente:



¡OJO! Esta transformación cambia la forma de la distribución de la variable, por lo que no debemos realizar inferencias a partir de esta nueva variable.

Ejercicio tipo A

Cree una nueva variable a partir del *Salario* (*Transformar* > *Calcular Variable*) llamada *logSalario*. Esta nueva variable se definirá usando la función logaritmo (por ejemplo, en base 10, que en SPSS se llama `log`) sobre una normalización previa de rango.

La idea es que los valores entre 0 y 12.500 caigan en la zona de transformación brusca para acentuar las diferencias (el intervalo [0,1] en el caso del logarirmo), mientras que el resto caiga en la zona de transformación suave que disminuye las diferencias relativas (el intervalo $[1,\infty)$). Esto hará que se acentúen las diferencias en salarios bajos, a la vez que se suavizan en los salarios medios y, sobre todo, en los altos.

Construya el histograma de la nueva variable para comprobar que es distinto al de la variable original. Cree también un diagrama de cajas en el que ahora deberán aparecer menos outliers en la zona alta y deberá detectarse el registro añadido anteriormente con un salario 11.000 como un outlier.

Observaciones:

- En el caso de que la asimetría fuese a la inversa, sería necesario darle la vuelta a la distribución antes de aplicar el logaritmo.

- Para suavizar los valores de forma un poco menos agresiva podríamos utilizar el logaritmo en base 2 o logaritmo natural (en base e).

NOTA FINAL:

Existen otras técnicas de preprocesamiento muy importantes, como por ejemplo:

- Selección de registros relevantes (métodos de edición y condensado):
  Estos métodos consisten en la selección de una muestra del total de registros, de tal forma que la selección sea lo suficientemente representativa.

- Selección de características:
  Las técnicas de selección de características escogen un subconjunto de los atributos de nuestro conjunto de datos, para evitar la presencia de muchos atributos correlados que no aportan información útil (p.ej. Análisis Factorial).