

# **Máster en Ciencia de Datos e Ingeniería de Computadores:**

## **Introducción a la Ciencia de Datos**

### **Fundamentos de Estadística: Análisis Exploratorio de Datos y Tests Estadísticos.**

Juan Carlos Cubero

Universidad de Granada

<http://decsai.ugr.es/jccubero>

Octubre 2017

1	Motivación	4
2	Introducción a SPSS: Datos y Variables	5
3	Análisis Exploratorio de Datos (AED). Gráficos y Estadística Descriptiva sobre una variable nominal.	8
4	Análisis Exploratorio de Datos (AED). Gráficos y Estadística Descriptiva sobre una variable de escala.	11
4.1	Histograma	11
4.2	Estadísticos de localización y dispersión	13
4.3	Cuantiles y diagramas de cajas	20
4.4	Outliers o valores atípicos	27
5	Normalización	30
6	Inferencia estadística	33
6.1	Introducción	33
6.2	Gráficos Q-Q	38
6.3	Estimación Puntual	41
6.4	Estimación por intervalos de confianza	46
7	Contraste o Test de Hipótesis	47
7.1	Construcción del test	47
7.2	T (t-Student)-tests	52
7.3	Tipos de errores. Potencia del test	56
8	AED: Informes y gráficos sobre varias variables (con agrupaciones) Nominal-Nominal	59
9	Análisis Estadísticos de dependencia. Nominal-Nominal	61
10	AED: Informes y gráficos sobre varias variables. Numérica-Numérica.	67
11	Análisis Estadísticos de dependencia. Numérica-Numérica. Regresión.	75
12	AED: Informes y gráficos sobre varias variables Numérica-Nominal	81
13	Análisis Estadísticos de dependencia. Numérica-Nominal	85
13.1	Prueba para múltiples muestras: ANOVA	85
13.2	Post-hoc	89
13.3	ANOVA de medidas repetidas	97
13.4	Anova de varias vías (Multi-way ANOVA)	98
13.5	Requisitos ANOVA	103
14	Tests no paramétricos	105
14.1	Ajuste de la distribución	105
14.2	Comparación de Medianas. Variables Independientes	110

14.3	Comparación de Medianas. Variables Pareadas	114
15	Comparación de Clasificadores	126
16	Bibliografía	130

# 1 Motivación

La Estadística es fundamental en la minería de datos:

- a) En sí misma, proporciona técnicas de minería de datos, como por ejemplo, el análisis de componentes principales, regresión, análisis factorial, etc.
- b) Sirve de filtro previo a 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 pudiesen suprimirse antes de proceder a dicho estudio?
- c) Se usa como parte de técnicas propias de minería de datos.  
Usar el test de la Chi cuadrado como medida de implicación entre dos items de una regla de asociación.

En cualquier caso, las técnicas estadísticas:

- a) Suelen requerir que el experto diga exactamente lo que quiere comprobar.
- b) 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 estadísticas "no paramétricas"

Usaremos SPSS para ilustrar algunos de los conceptos teóricos que veremos en esta Introducción a la Estadística.

También se puede usar PSPP, que es libre, aunque carece de bastantes opciones, sobre todo gráficos. Para descargar el fichero de instalación para Windows: <http://pspp.awardspace.com/>

## 2 Introducción a SPSS: Datos y Variables

### ¿Qué es SPSS?

- SPSS es un software informático desarrollado para realizar análisis estadísticos y gestión de datos.
- Utiliza menús descriptivos y cuadros de diálogo simples para ejecutar las funciones solicitadas por el usuario.
- También ofrece la posibilidad de ejecutar una serie de comandos especificados en lo que se denomina fichero de sintaxis.
- SPSS posee una estructura tipo modular. Las distintas funcionalidades que incorpora se corresponden con módulos, cada uno de los cuales ha sido realizado por alguna institución.
- El módulo base forma el núcleo del sistema, y contiene los comandos de lectura y transformación de datos y ficheros, así como procedimientos estadísticos básicos.
- Estudiaremos la versión 20.0.

### Ejecución de SPSS:

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

⇒ Cargad "Datos de Empleados"

C:\Program Files\IBM\SPSS\Statistics\20\Samples\Spanish\Employee data.sav

También disponible en: <http://decsai.ugr.es/jccubero/EmployeeData.sav>

Hay una vista de datos y vista de variables (tipo hoja de cálculo)

La extensión de los ficheros con los que trabaja SPSS es .sav. Se pueden crear ficheros de datos nuevos, importar bases de datos Excel, Oracle, etc, o importar ficheros de texto (datos con algún tipo de separaciones) PSPP también puede abrir directamente los ficheros sav.

## Definición de variables:

Los nombres de variables no pueden tener más de 8 letras, pero se les puede poner una etiqueta 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:* Es la más importante. Establece qué mide la variable.

- Nominal: Una variable que toma valores no ordenados entre sí. Por ejemplo, el color de pelo.
- Ordinal: Una variable que toma valores ordenados entre sí. Por ejemplo, el nivel de satisfacción, medido de 0 a 5
- Escala: Una variable que toma valores numéricos usuales, para los que tiene sentido la operación de resta. Por ejemplo, la edad.

*Tipos:* Establece cómo codificamos en la BD lo que la variable mide

- Numérico con una anchura determinada.
- Cadena: La típica cadena de caracteres
- Otros: Dólar, fecha, etc.

*Ejemplos:*

- Color de Pelo de una persona.  
Lo lógico sería una medida nominal y un tipo de cadena. Pero también podríamos usar una medida nominal y un tipo numérico con anchura de 1 dígito (0 para el rojo, 1 para el negro, etc)
- Ingresos de una persona  
Lo lógico sería una medida de escala y un tipo numérico
- Grado de satisfacción del usuario.  
Deberíamos usar una medida ordinal. El tipo podría ser de 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 es más ser Directivo que Administrativo, usaríamos una medida ordinal, y un tipo de 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 el orden lexicográfico (no es usual)

Puede que queramos restringir los posibles valores que pueda tomar una variable. Por 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, se usa la columna de **valores**. Observad que por una parte están los valores ("h", "m") que deben corresponderse con el tipo de la variable y por otra parte están las **etiquetas de los valores** que son una cadena de caracteres que luego aparecerá en los resúmenes que SPSS haga.

Observad la categoría laboral. Lo lógico sería una medida ordinal con un tipo de cadena con valores "d", "a", "s" (o incluso "directivo", "administrativo", "seguridad"). Sin embargo es un tipo numérico. La razón es que algunos tests estadísticos necesitan que la variable sea numérica (aunque tenga una medida ordinal y no de escala) para poder trabajar con ella. Observad que como posibles valores tiene 1, 2, 3. Pero luego, como etiquetas de valores tiene "Administrativo", "Seguridad", "Directivo"

### 3 Análisis Exploratorio de Datos (AED).

#### Gráficos y Estadística Descriptiva sobre una variable nominal.

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

Queremos responder a la pregunta:

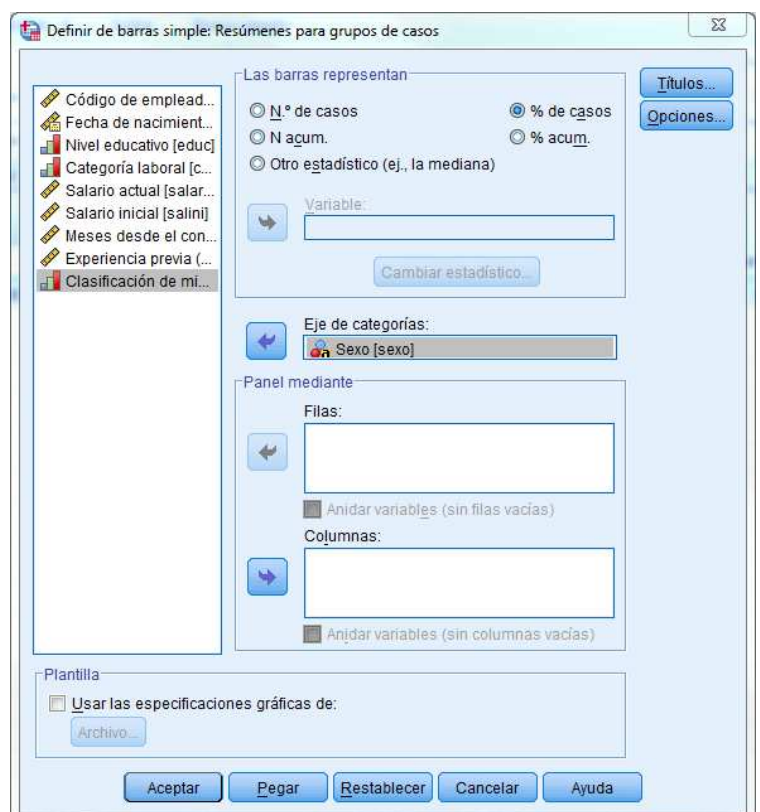
**¿Cómo se distribuye una variable nominal?**

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

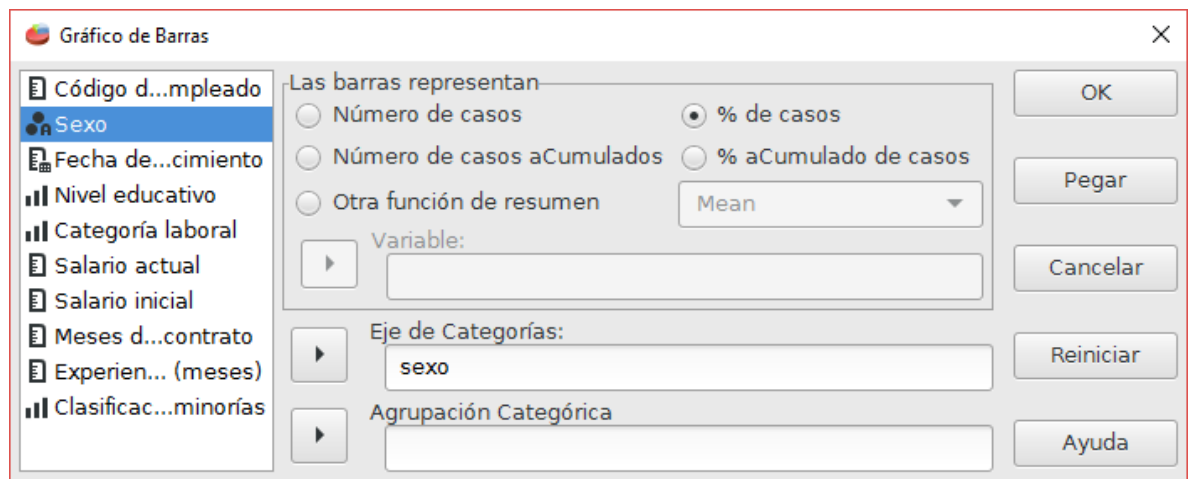
⇒ SPSS: Gráficos/Cuadros de diálogo antiguos/Barras/ → Simple.

PSPP: Gráficos/Gráficos de Barras

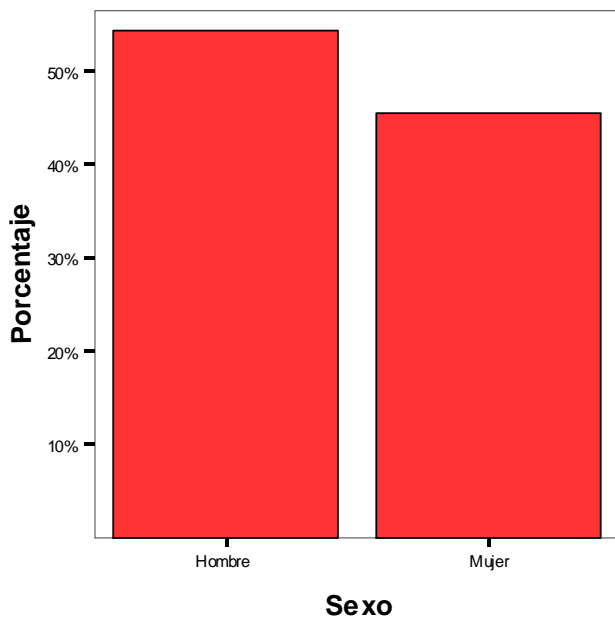
Arrastramos con el ratón la variable Sexo al cuadro del eje de categorías y seleccionamos % de casos.



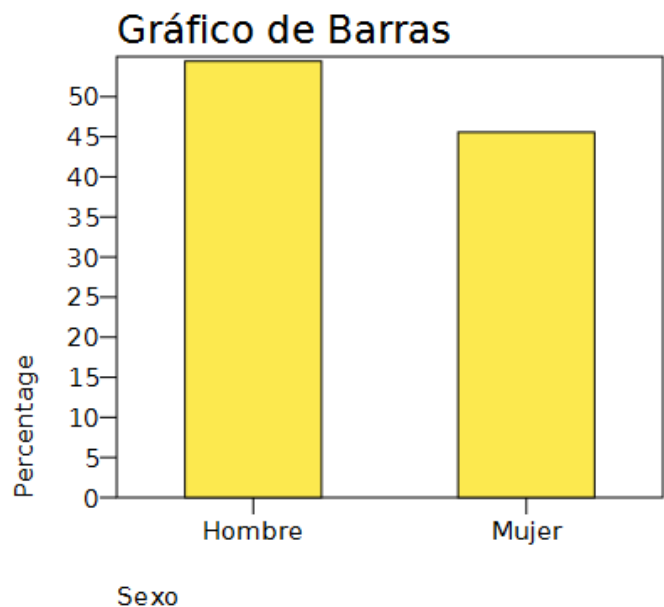




Aparece el visor de resultados. Todos los análisis que se vayan realizando se van guardando en el mismo sitio. Si quisiésemos suprimir cualquier elemento del visor de resultados, basta con borrarlo del panel izquierdo. El contenido del visor se guarda en un fichero con extensión spv.



Las barras muestran porcentajes



Al hacer doble click sobre el gráfico anterior (sólo SPSS), se abre un marco interactivo en el que podemos editar y cambiar algunos de los elementos del gráfico, o bien podemos cambiar las variables indicadas, o incluso añadir cajas de texto con nuestros propios comentarios. Los gráficos obtenidos en el visor de resultados, pueden cambiarse, modificarse, copiar a Word por ejemplo, etc.

Una vez que tenemos una aproximación gráfica, podemos ver algunos estadísticos que nos informen de cómo es la muestra. Para una variable de tipo de escala no hay mucha información que ofrecer: la moda, frecuencias relativas, y poco más (obviamente, la media no tiene sentido, por ejemplo)

⇒ **Analizar/Estadísticos Descriptivos/Frecuencias/Seleccionad Sexo**

*Gráficos -> Gráficos de barras / Porcentajes.* (nos muestra el mismo gráfico anterior)

*Estadísticos -> Aunque puede marcarse, ningún estadístico aparece en el resultado (ni siquiera la moda).* Esto ocurre porque se eligió un tipo de cadena de caracteres, pero 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 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

No sale la etiqueta  
"Administrativo" sino el  
valor 1. ☹

## 4 Análisis Exploratorio de Datos (AED).

### Gráficos y Estadística Descriptiva sobre una variable de escala.

#### 4.1 Histograma

Queremos responder a la pregunta:

**¿Cómo se distribuye una variable numérica?**

Para las variables de escala, usaremos un **histograma**.

Los histogramas de frecuencias son una forma habitual de visualización de distribuciones para una sola variable. Para construirlo, se divide el rango entre el mayor y menor valor en intervalos de un mismo tamaño, y se representa en ordenadas mediante una barra el número de casos cuyo valor de la variable está contenido en el intervalo correspondiente.

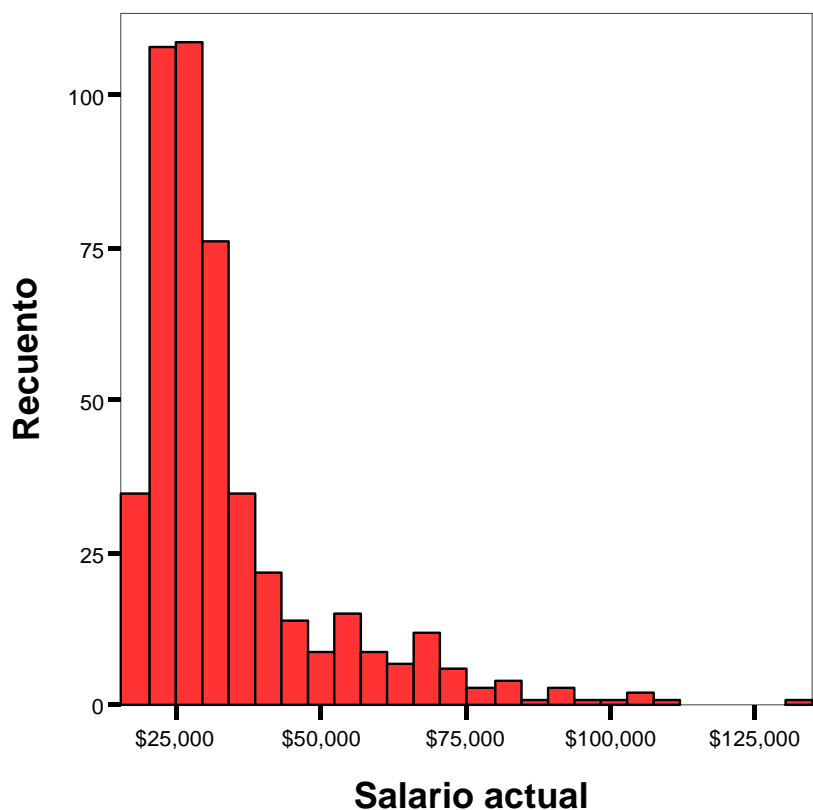
**¿Cómo se distribuye el salario entre los empleados?**

⇒ SPSS: Gráficos/Cuadros de diálogo antiguos/Histograma/

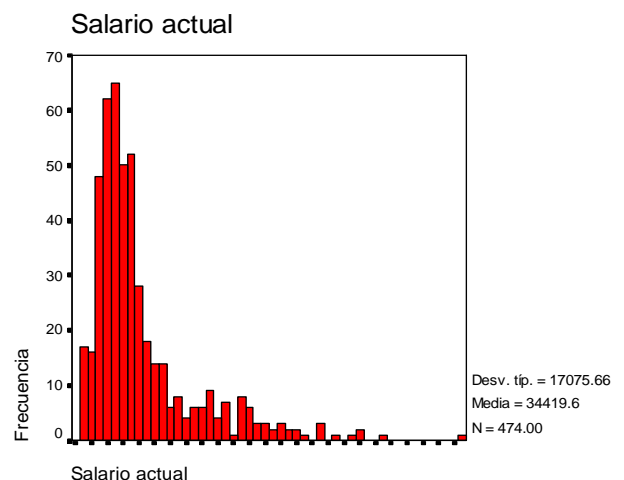
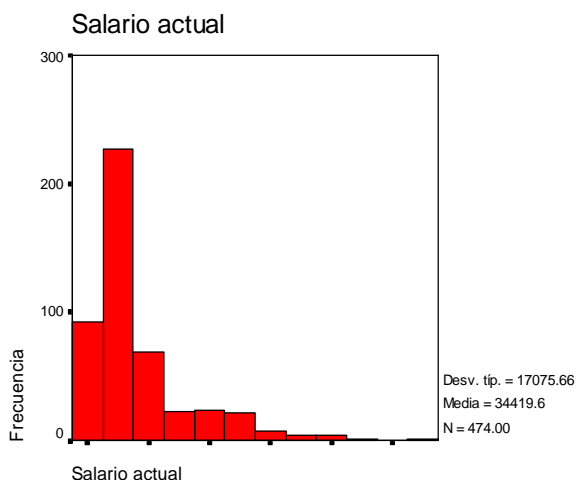
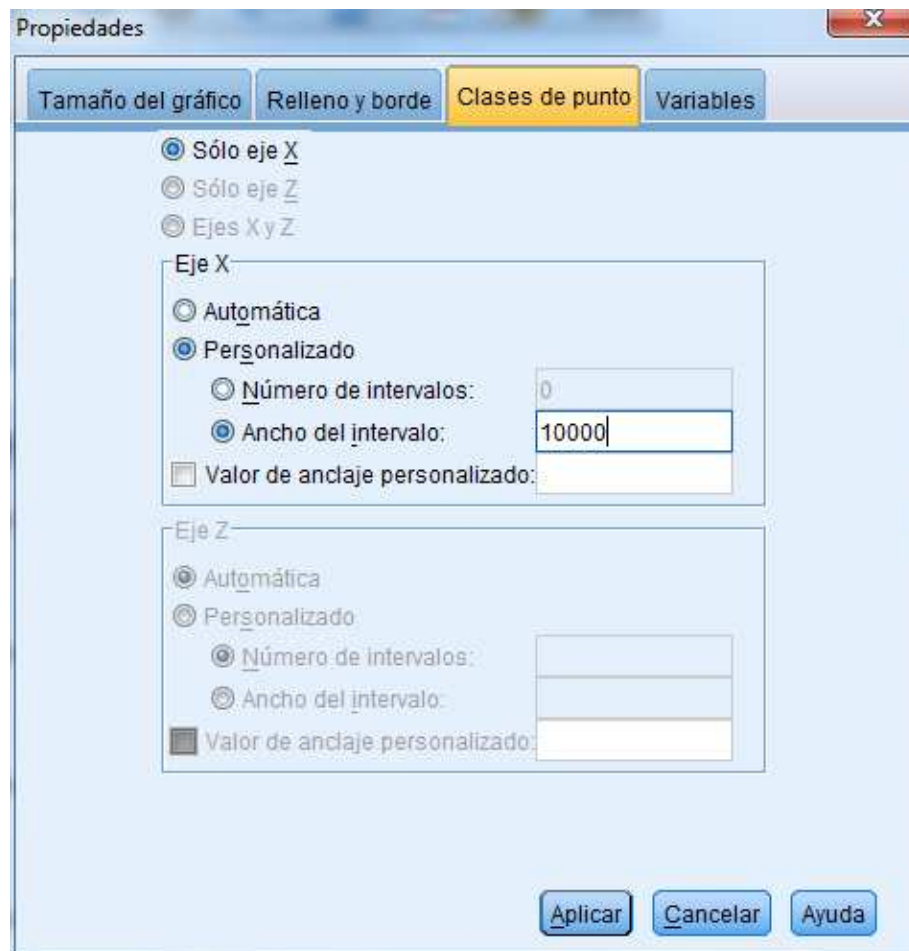
PSPP: Gráficos/Histograma/

Seleccionad

Salario Actual en "Variables"



El ancho de los rectángulos (que determina el número de ellos) afecta a la información que muestra el histograma y, específicamente, puede afectar a su forma. Cambiando esta característica podemos obtener información más precisa y detallada. Por ejemplo, podemos partir de pocos rectángulos e ir detallando progresivamente si es necesario. Para ello, haced doble click (sólo SPSS) sobre el gráfico que hay en el visor de resultados. Aparecerá el Editor de Gráficos. Haced doble click sobre cualquiera de las barras:



## 4.2 Estadísticos de localización y dispersión

Queremos tener una visión global de la distribución, utilizando *medidas de resumen*. Es decir, queremos resumir la información contenida en la distribución, usando un par de valores numéricos. Dichos valores se construyen 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:**  $N$  ( $n$ ), es el número de casos en la muestra.

Existen dos tipos importantes de estadísticos:

- ✓ **De localización.** Dan una idea de cuáles son los valores habituales de la distribución. Tratan de decirnos dónde es más densa la distribución.

- **Media muestral.**

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

Problema: sensible a casos aislados (outliers).

Media (40, 50, 50, 60, 40, 60) = 50

Media (40, 50, 50, 60, 40, 130) = **61.6...**

- **Mediana** muestral: valor central de la lista ordenada de valores. Menos sensible a outliers. El 50% de los valores están a su derecha y el otro 50% a la izda.

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.

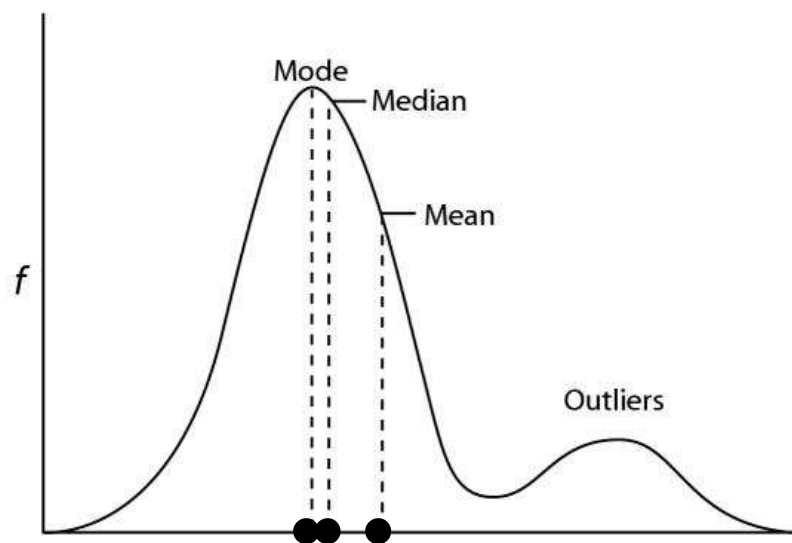
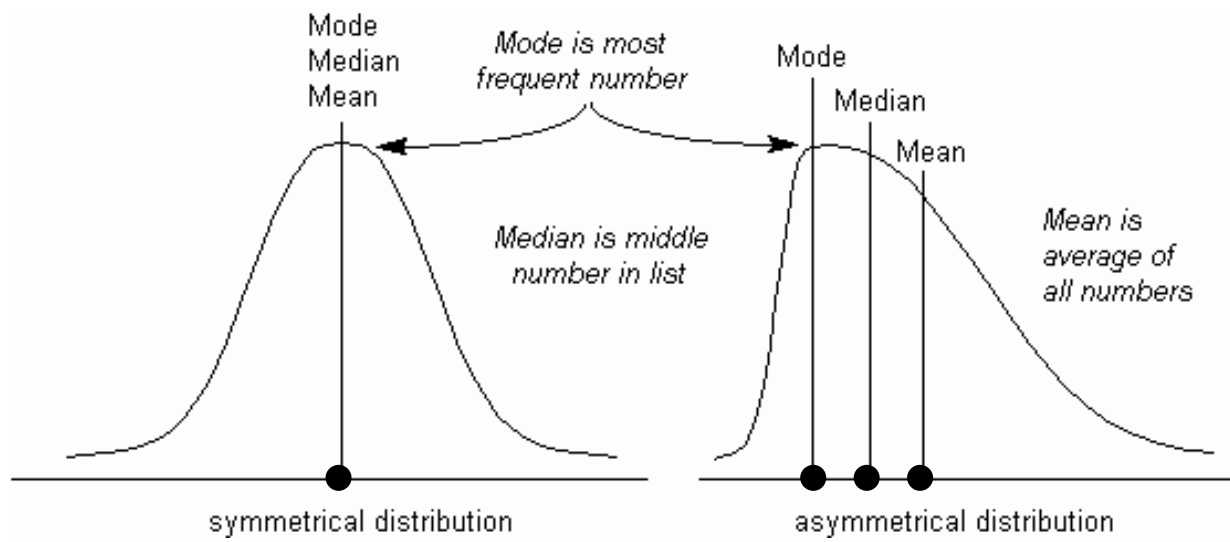
Mediana(40, 50, 50, 60, 40, 60) = Mediana(40, 40, 50, 50, 60, 60) = 50

Mediana (40, 50, 50, 60, 40, 130) = Mediana (40, 40, 50, 50, 60, 130) = **50**

- **Moda** muestral: valor más común. En ocasiones, una distribución tiene más de una moda. En caso contrario, se dice unimodal.

Moda(40, 50, 50, 60, 40, 60) = {40, 50, 60}

Moda(40, 50, 50, 60, 40, 130) = {40, 50}



✓ **De dispersión.** Dan una idea de cual es la variabilidad en los datos.

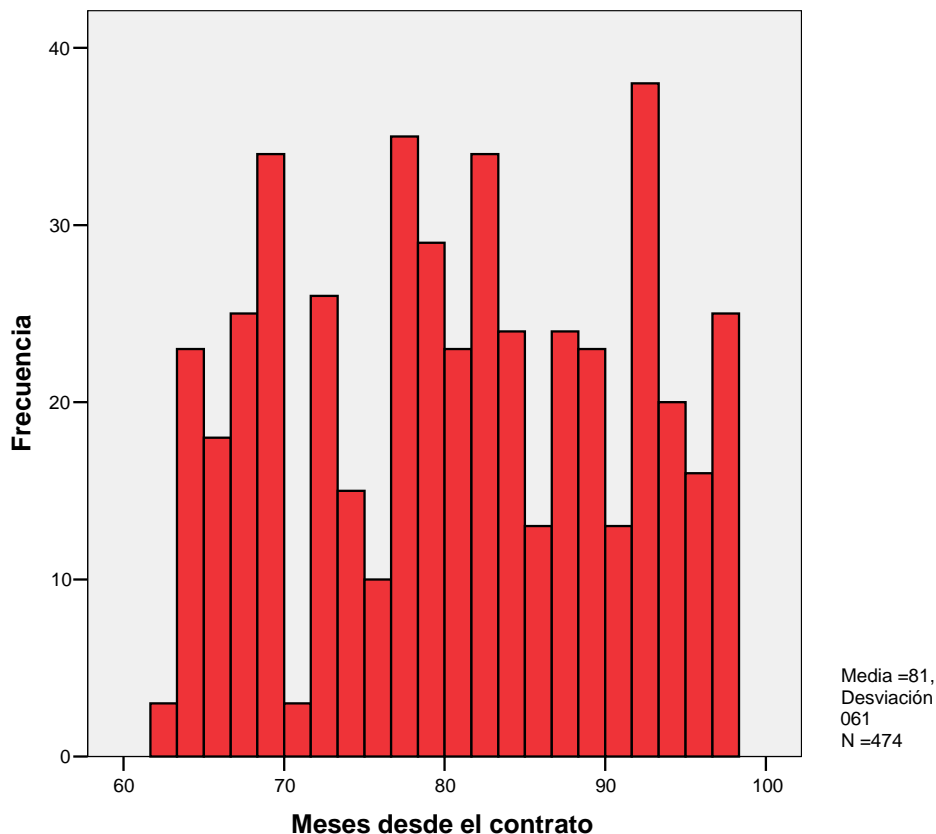
- **Desviación típica muestral.** Representa cómo de dispersos están los datos con respecto a la media aritmética. Es una medida global.

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

Para una amplia mayoría de distribuciones, la mayor parte de los valores (~95%) están comprendidos entre 2 desviaciones de la media (media  $\pm 2 S$ ) y ~70% de los casos están a una distancia de la media no mayor de 1 desviación.

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

Mucha varianza



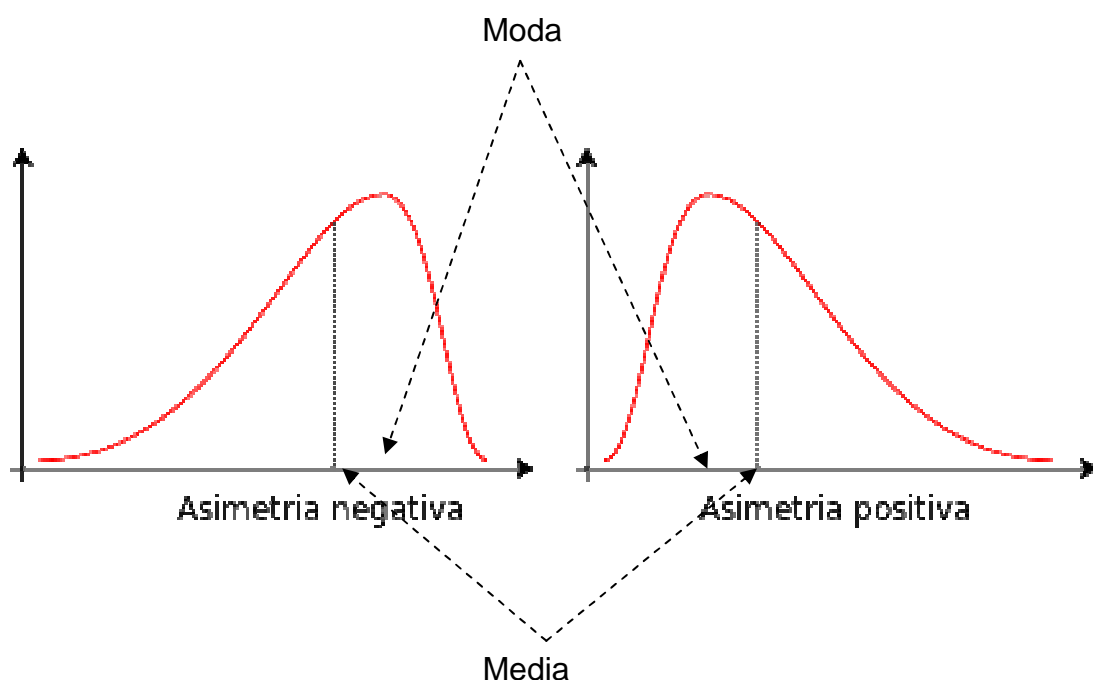
Ampliación: Consultad en Wikipedia/Google: “Chebyshev's inequality” y “Vysochanskij-Petunin inequality”. Por ejemplo, la desigualdad de Chebyshev nos dice que un total de  $1 - 1/k^2$  valores de cualquier distribución, pertenecen al intervalo Esperanza  $\pm k \cdot$  desviación

- ✓ **De forma.** Dan una idea de cual es la forma de la distribución.



- **Skew.** (Asimetría) Representa si los datos están más presentes a la derecha o izquierda (en distribuciones unimodales)

$$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$



Su definición exacta es el tercer momento tipificado de la distribución, aunque Pearson dió una forma aproximada de calcularlo en distribuciones unimodales como  $(\text{media}-\text{moda})/\text{desv.tip}$

La distribución normal es simétrica por lo que tiene un valor de asimetría 0. Un valor de asimetría mayor que 1, en valor absoluto, indica generalmente una distribución que difiere de manera significativa de la distribución normal. Usualmente toma valores entre -2 y 2 y distribuciones muy asimétricas entre -4 y 4.



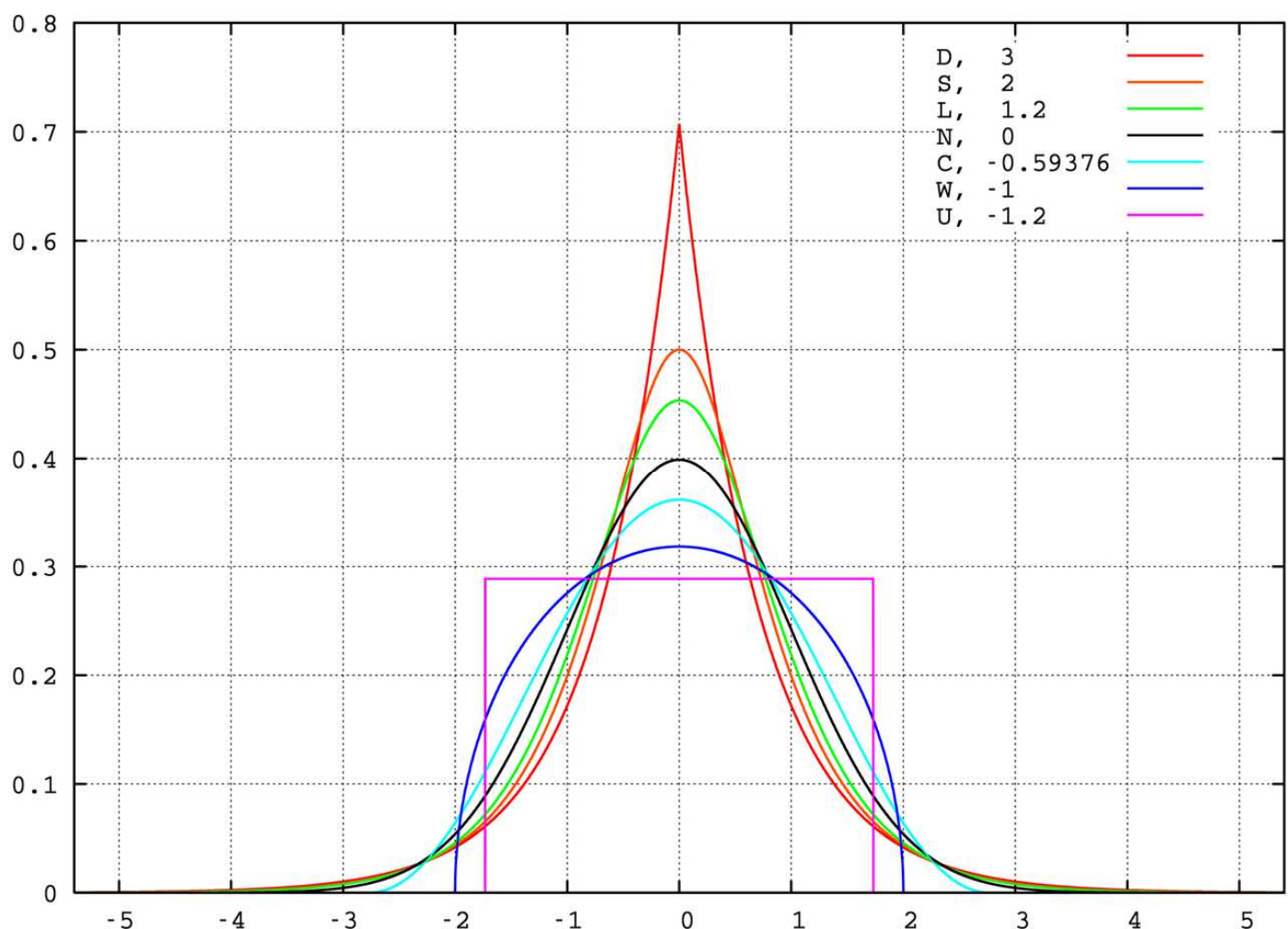
- **Kurtosis (curtosis).** Medida del grado en que las observaciones están agrupadas en torno al punto central.

$$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$$

Tomando la distribución normal como referencia, una distribución puede ser:

(vea <https://www.riskprep.com/all-tutorials/36-exam-22/145-understanding-kurtosis>)

- kurtosis cero (mesocúrtica): como la distribución normal.
- kurtosis positiva (leptocúrtica): más apuntada (las observaciones se concentran más) y con colas más largas que la normal
- kurtosis negativa (platicúrtica): menos apuntada (las observaciones se agrupan menos) y con colas más cortas que la normal



**¿Cual es la media aritmética del salario de los empleados?**

**¿Qué dispersión ó varianza presenta el salario entre los empleados?**

⇒ **Analizar/Estadísticos Descriptivos/Frecuencias/Seleccional Salario Actual**, quitad "Mostrar tablas de frecuencias", en gráficos seleccionad Histograma y en Estadísticos la media, desviación típica, mínimo y máximo. Si aparecen \*\*\*\*\*, tendremos que agrandar convenientemente la tabla de resultados.

### Estadísticos

Salario actual

N	Válidos	474
	Perdidos	0
Media		\$34,419.57
Mediana		\$28,875.00
Desv. típ.		\$17,075.661
Asimetría		2,125
Error típ. de asimetría		,112
Curtosis		5,378
Error típ. de curtosis		,224

Con apenas cuatro valores de resumen, nos hacemos una idea muy aproximada de cual es la distribución de los datos. La media es \$34.419. La mediana es \$28.875, por lo que la mitad de los trabajadores ganan menos de \$28.875 y la otra mitad gana más.

En cuanto a la variabilidad, el 70% de los individuos tienen un salario en el intervalo  $[34.419 - 17.075, 34.419 + 17.075] \approx [17.5, 51.5]$

y la mayor parte (95%) de los individuos tienen un salario en el intervalo

$$[34.419 - 2 * 17.075, 34.419 + 2 * 17.075] = [0.419, 68.4]$$

Hay una asimetría positiva, por lo que los salarios se concentran en la zona de la izquierda (salarios bajos) y una curtosis también positiva alta por lo que hay una gran concentración en torno a un intervalo estrecho.

**Ejercicio:** Realizad el mismo análisis (descriptivo y gráfico) con el Salario Inicial y Meses desde el Contrato

**Ejercicio:** Realizad el mismo análisis (descriptivo y gráfico) sobre algunas variables del fichero de datos Mundo 95

**Opcional:**



Desde SPSS podremos sacar los mismos estadísticos desde distintos sitios. De hecho, en SPSS podemos hacer distintos análisis desde distintos menús (la verdad es que confunde un poco esta forma de trabajar)

⇒ **Analizar/Informes/Resúmenes de casos**/Seleccionamos "Salario Actual" y quitamos "Mostrar los casos". Como estadísticos, seleccionamos los mismos:

Haced lo mismo incluyendo también la variable Experiencia Previa.

**Resúmenes de casos**

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

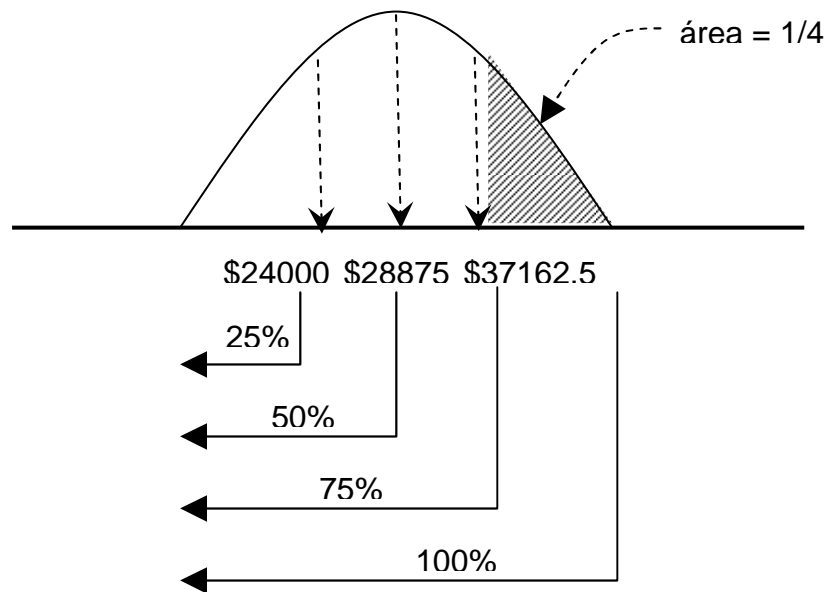
⇒ **Analizar/Estadísticos Descriptivos/Descriptivos**/Seleccionamos "Salario Actual" y los mismos estadísticos dentro de Opciones.

Podemos incluir varias variables en la misma tabla de resumen:

⇒ **Analizar/ Informes/Resúmenes de Casos/** Seleccionad Salario Actual y Experiencia Previa. No mostrad los casos y como estadísticos, seleccionad la media y desviación típica

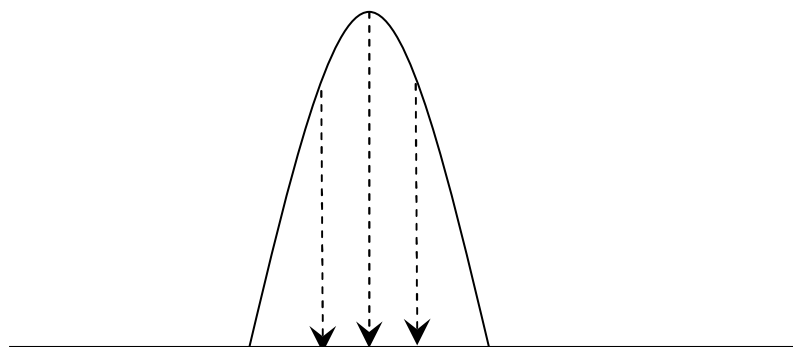
### 4.3 Cuantiles y diagramas de cajas

Los **cuantiles** son unos estadísticos de tendencia central, pero que también ofrecen información sobre la dispersión de los datos. El cuantil 0.25 es un valor tal que el 25% de los valores de la muestra son menores que él. Y así con el resto (obviamente, el cuantil 0.50 es la mediana)



Los cuantiles que dividen la distribución en 4 partes (0.25-0.50-0.75) se denominan **cuartiles**. Los cuantiles que dividen la distribución en 100 partes se denominan **percentiles**.

Si dos cuantiles están muy próximos (imaginemos \$27500 y \$27900), significa que un 25% de la muestra tiene salarios muy parecidos, por lo que hay una concentración en ese intervalo.



⇒ **Analizar/Estadísticos Descriptivos/Frecuencias/Seleccionad Salario Actual**, y en Estadísticos añadir los cuartiles.

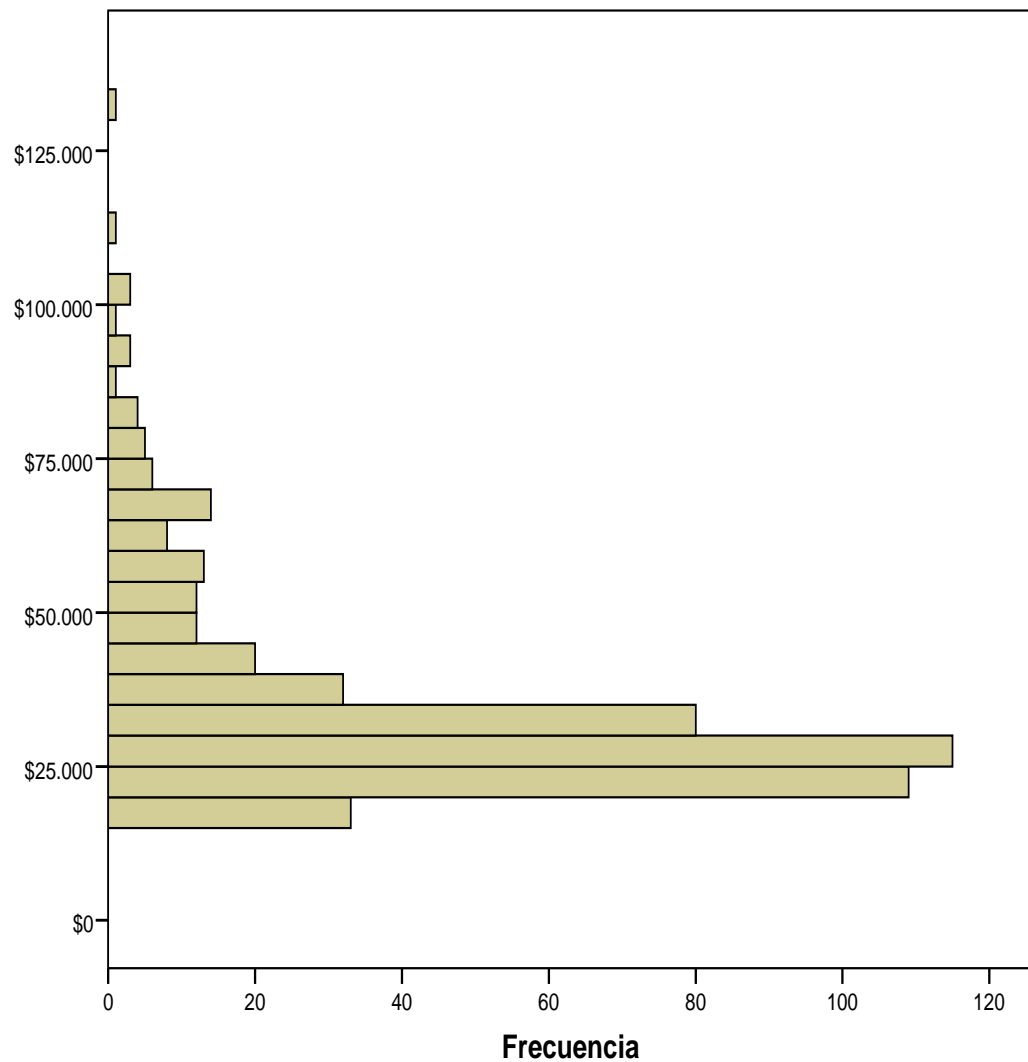
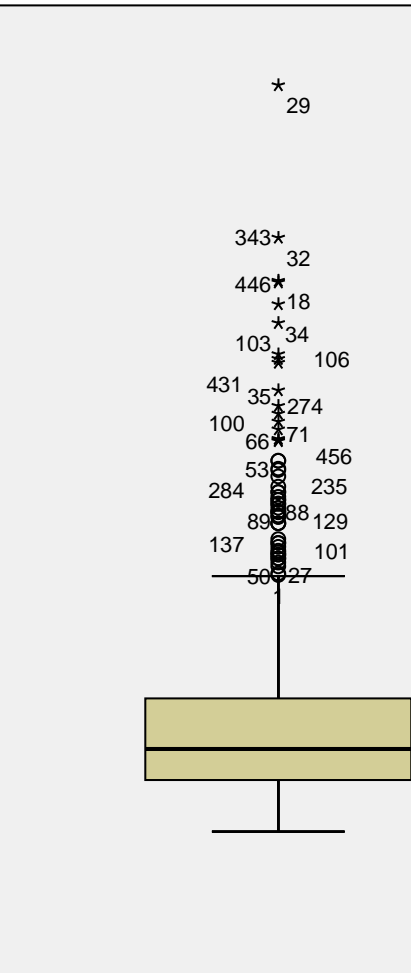
Desde PSPP hay que hacerlo desde línea de comando

### Estadísticos

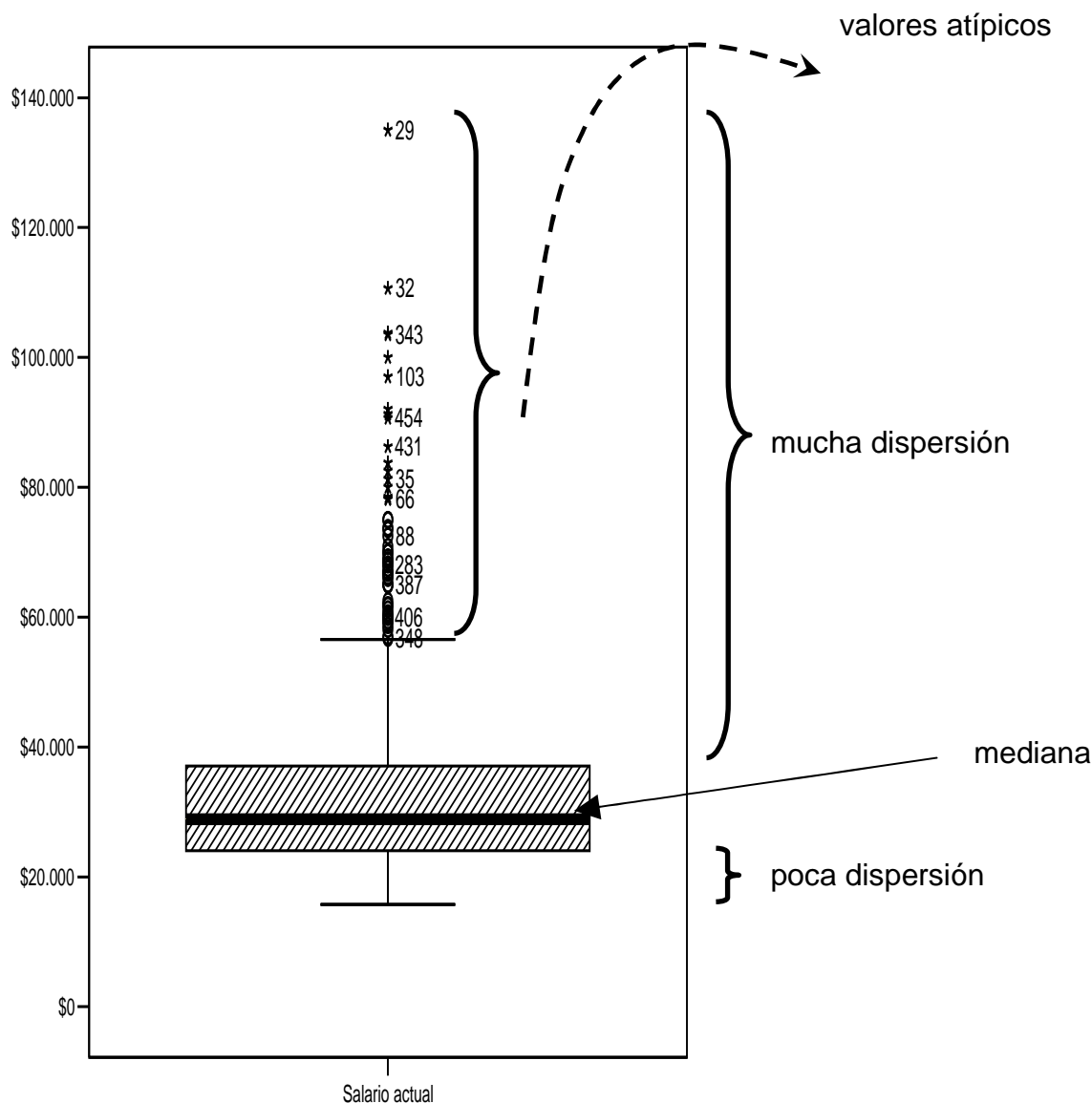
Salario actual

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

⇒ Gráficos/Generador de gráficos/



El diagrama de caja contiene la información de los cuantiles. Son 4 bloques correspondientes a los cuantiles 0.25,0.50,0.75,1.00. Se representan como una línea vertical, bloque, bloque y otra línea vertical. La anchura (en horizontal) de las cajas no representa nada. Lo importante es la altura de cada segmento de la caja. Cuanto mayor sea, mayor es la dispersión.

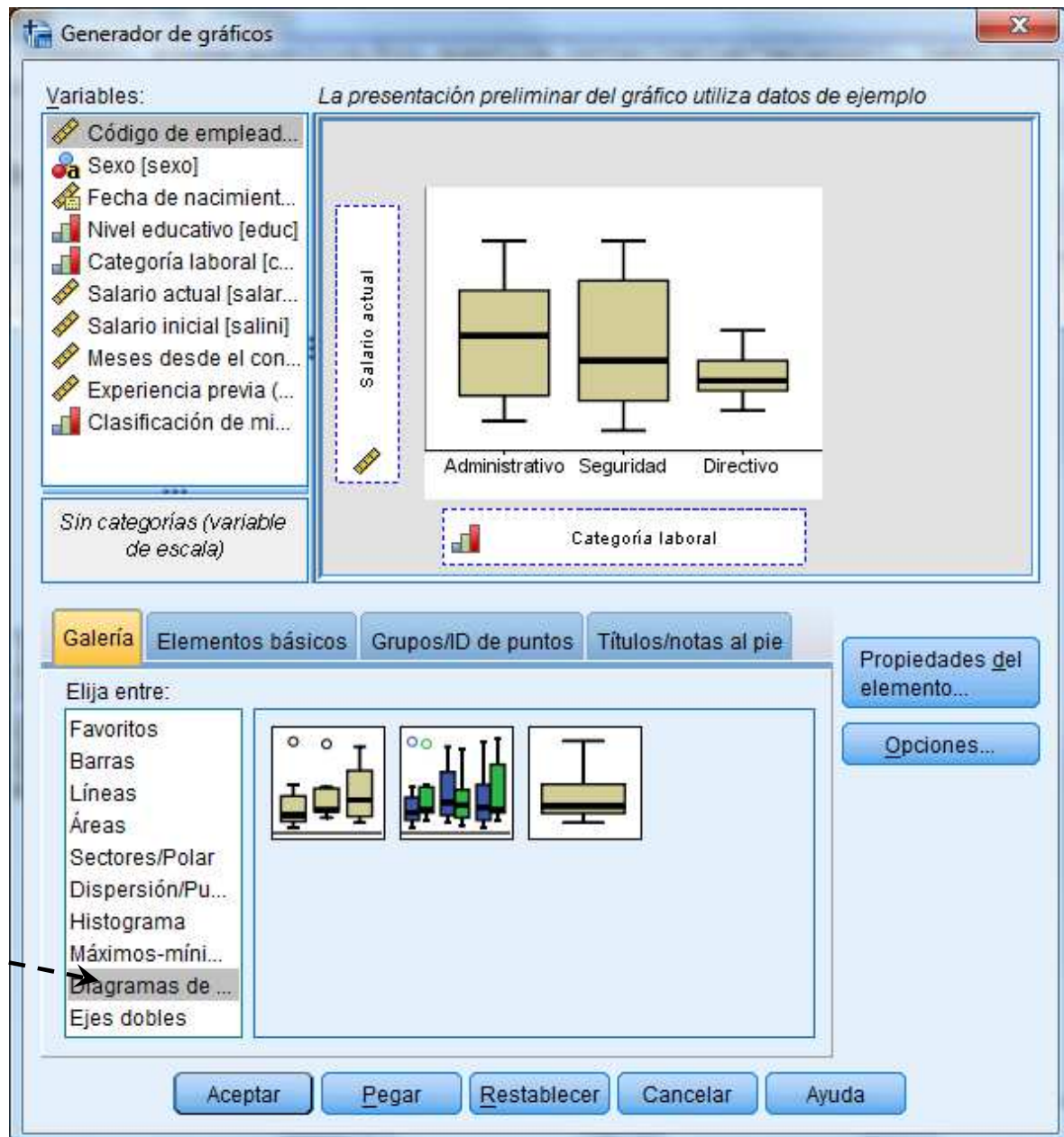


Podemos apreciar que la mitad de los empleados ganan entre unas 15000 y 30000 mientras que en la otra mitad hay mucha más variación de salarios (entre unas 30000 y 140000).



**Ejercicio:** Incluid otro gráfico con la variable "Experiencia Previa". Se ve que la mitad de los datos están agolpados en un intervalo de valores muy pequeño, mientras que la otra mitad están mucho más dispersos. Sin embargo, con la variable "Meses desde el contrato", no hay apenas dispersión (ved también el histograma correspondiente)

Podríamos estar interesados en usar una variable de agrupación. Por ejemplo, queremos ver la distribución del Salario para los Administrativos, de forma separada de los Directivos.



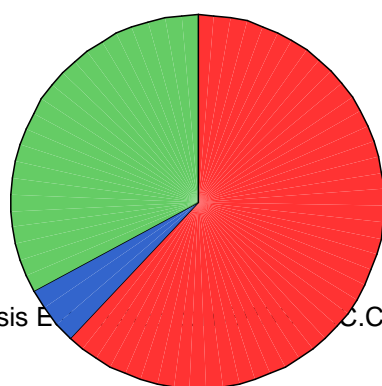
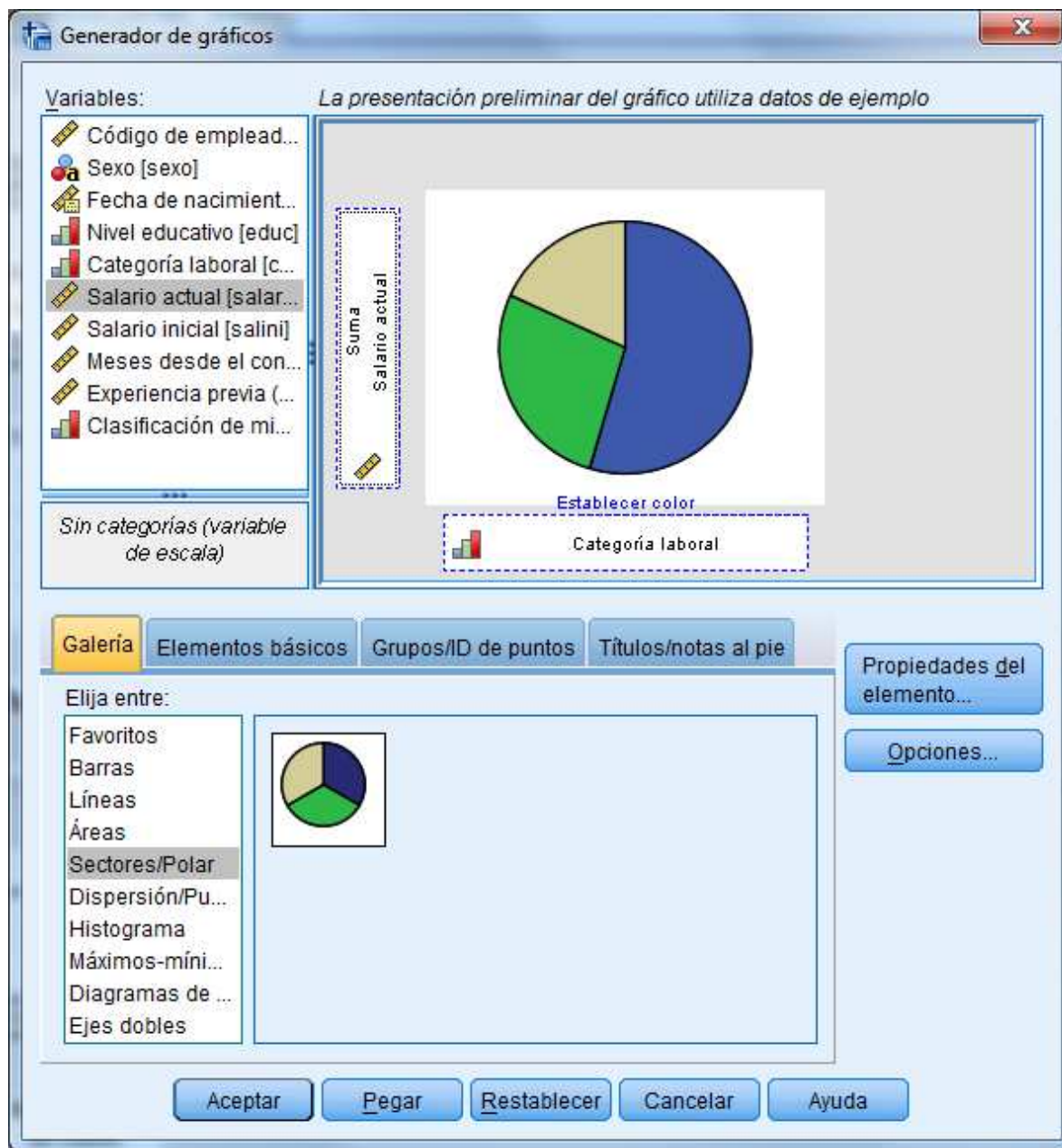
### Opcional:

Ampliación

En los anteriores ejemplos, hemos trabajado con el recuento de individuos (frecuencias o número de apariciones). A veces, podemos estar interesados en usar otra medida de recuento. Por ejemplo, para 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?**

Seleccionaríamos:

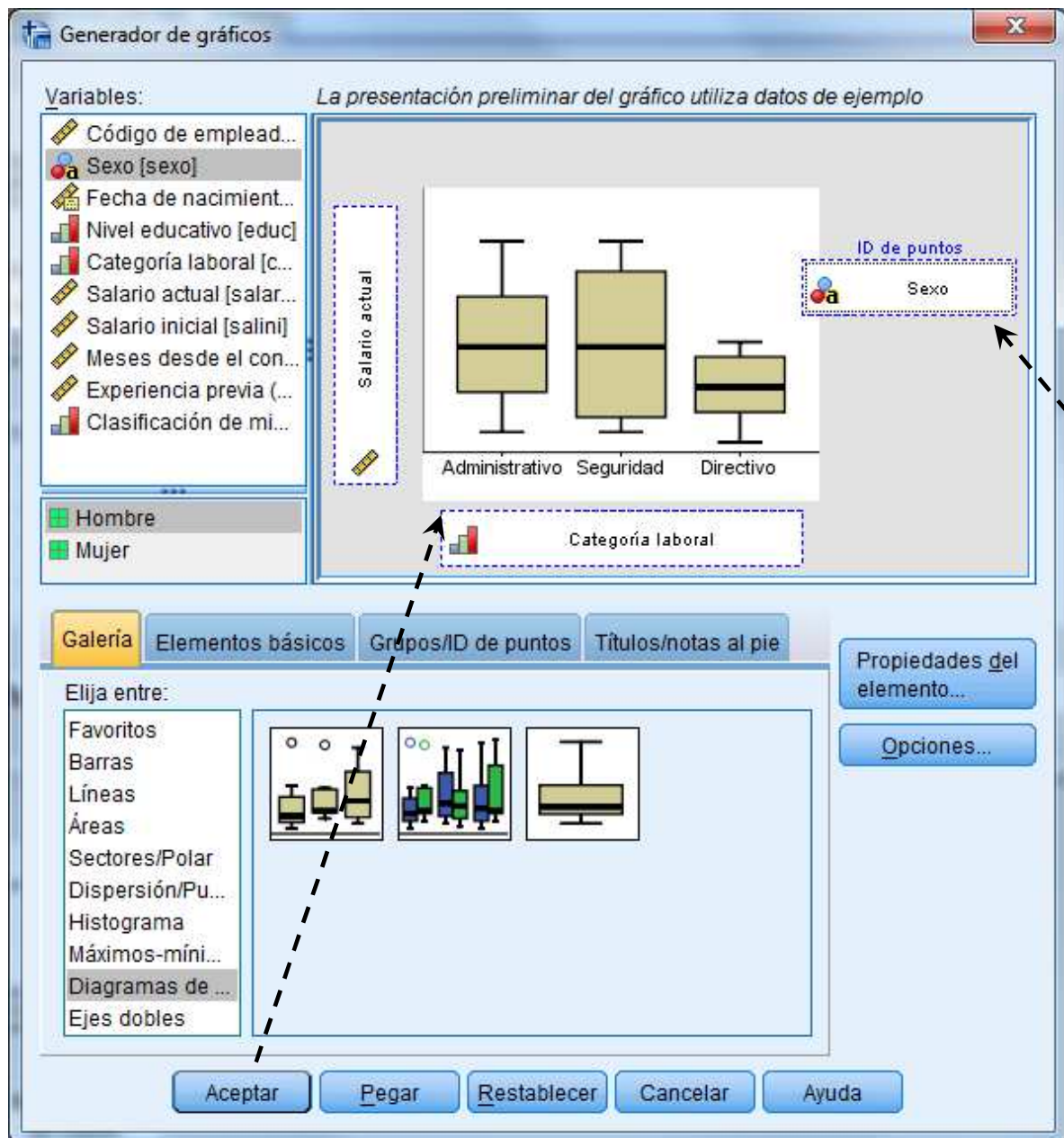


**Categoría laboral**  
■ Administrativo  
■ Seguridad  
■ Directivo

Los sectores muestran Sumas de salario

## 4.4 Outliers o valores atípicos

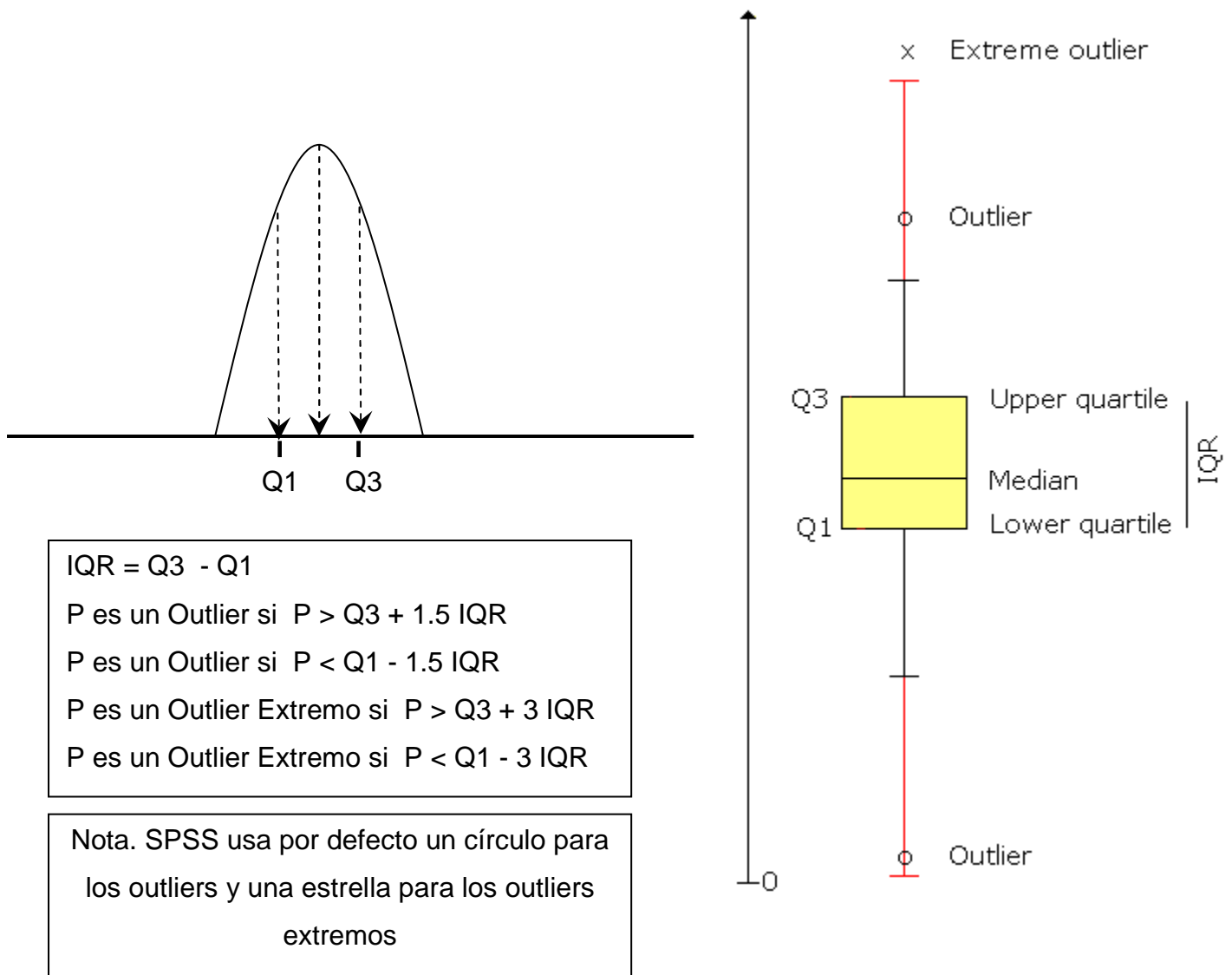
Volved a generar al mismo gráfico (diagrama de cajas con Salario Actual), pero con las opciones siguientes:



Ahora, los valores atípicos se etiquetan en función del Sexo. Se aprecia que los valores atípicos corresponden a los hombres (y siempre en sueldos altos)

¿Qué criterio se usa para decir que un valor determinado es 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, pero no está disponible en SPSS)
- Si consideramos una única dimensión (un único atributo), sea cual sea la distribución, se considera que los valores anormales son los que están *alejados* de la mediana:



- Si consideramos varias dimensiones, existen distintas aproximaciones (se ven en el curso "Minería de datos: Aprendizaje no supervisado y detección de anomalías"):
  - = "Local Outlier Factor" Da una puntuación de hasta qué punto un valor es un outlier
  - = Métodos de clustering.

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

1. En primer lugar, analizar si son registros que se pueden excluir del estudio. A veces, pueden representar una información real interesante y otras veces son errores de medida.
2. Si la técnica estadística o de minería de datos lo permite, se pueden dejar dichos registros para que los procese la propia técnica
3. Si la técnica estadística o de minería de datos no maneja outliers, pueden excluirse del estudio correspondiente

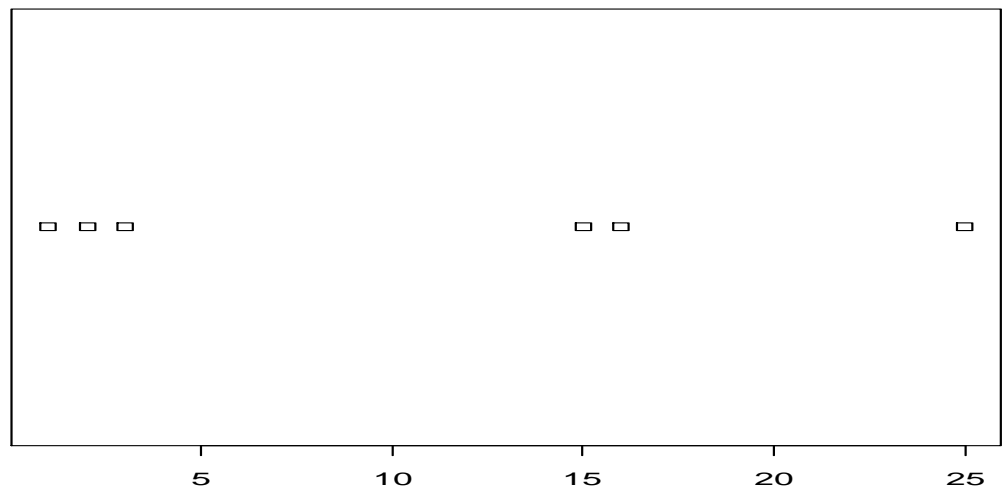
## 5 Normalización

En la fase de preprocesamiento en un proceso de Minería de datos será necesario seleccionar las variables adecuadas, eliminar los valores atípicos, muestrear un subconjunto de los registros disponibles, etc. En el curso Minería de Datos I: Preprocesamiento y Clasificación se ven distintas técnicas.

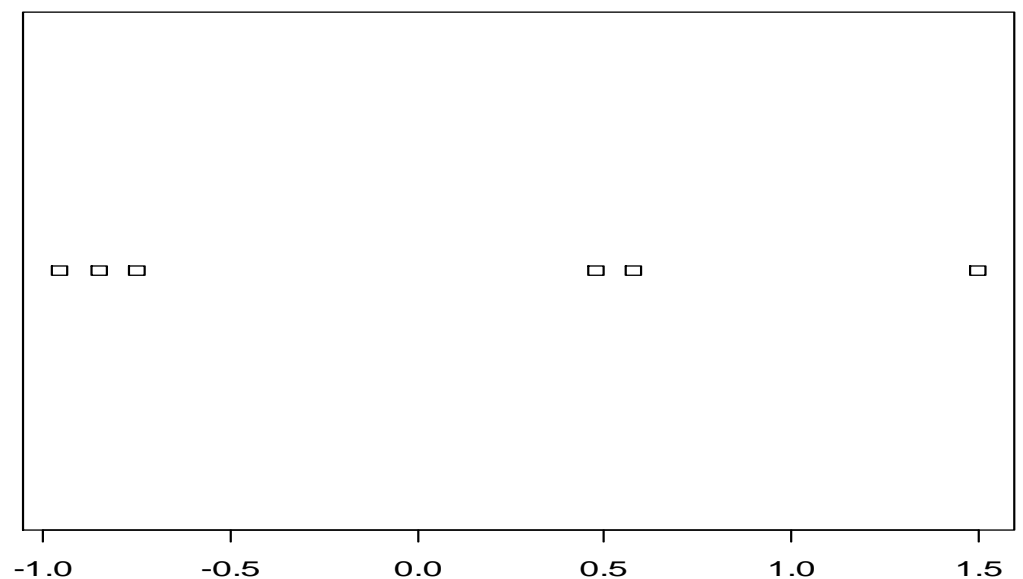
Cuando se van a trabajar con distancias (clustering, KNN, etc) es necesario que todas las variables estén especificadas en el mismo rango, para que unas no dominen sobre otras. Una técnica muy usada en Estadística es la normalización por el estadístico t-Student.

$$X_i \rightarrow \frac{X_i - \bar{X}}{S}$$

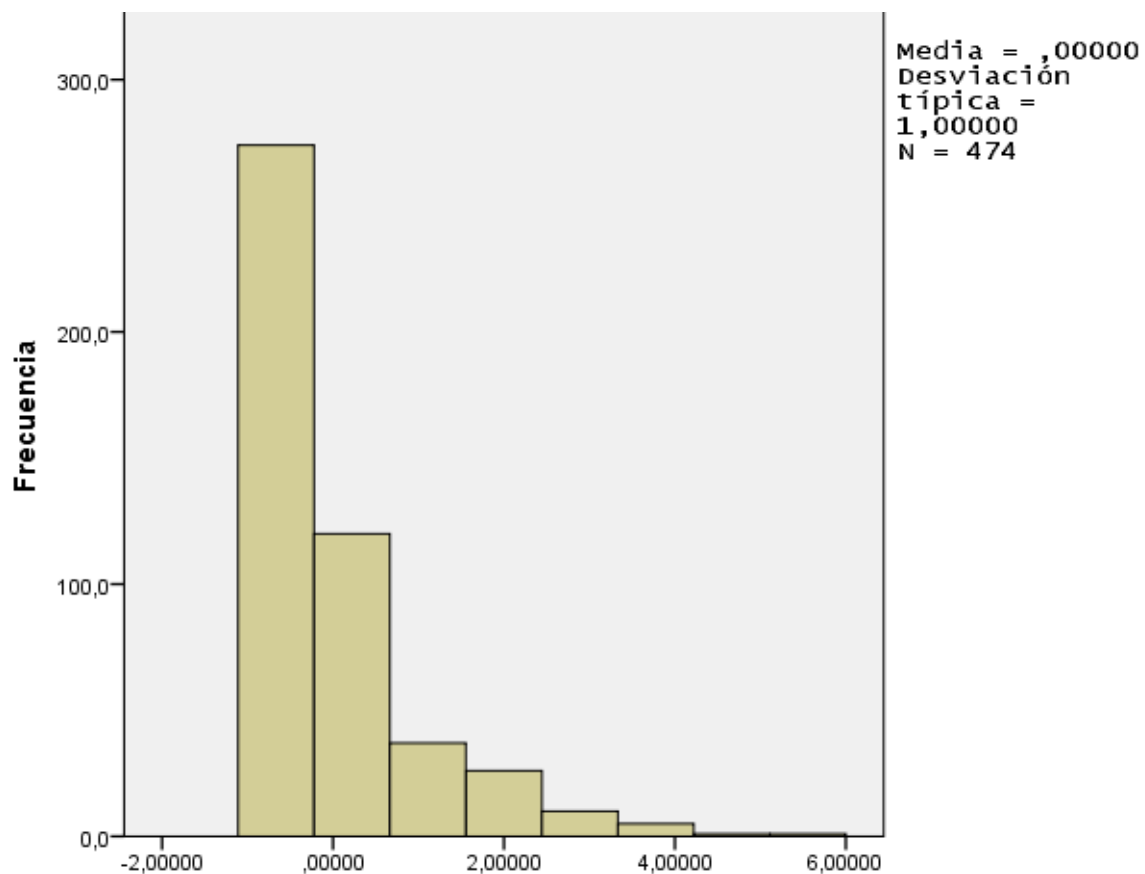
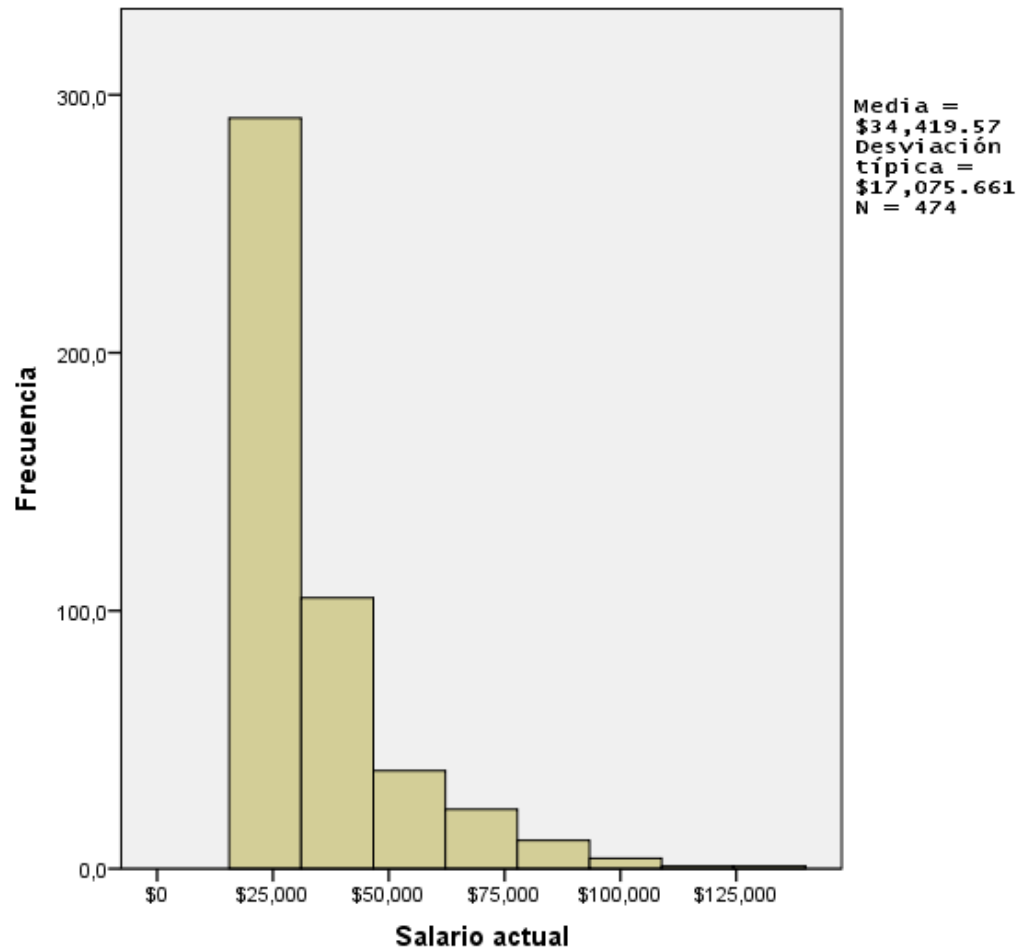
Supongamos  
los datos 1,2,3,15,16,25



Apliquemos una  
normalización  
t-Student



Si representamos los datos con un histograma, éstos son idénticos (antes y después de la normalización)



La normalización tiene las siguientes ventajas:

- Expresa cualquier variable en el mismo rango, que es el rango de valores en los que se mueve la  $N(0,1)$ . Esto permite mezclar las variables en cálculos como distancias.
- Un valor negativo indica que el correspondiente original estaba por debajo de su media
- Un valor positivo indica que el correspondiente original estaba por encima de su media
- El valor absoluto normalizado nos indica cuántas veces, el valor original, estaba por encima de su media, en términos de la desviación muestral.

Nota: Si en vez de usar la media y desviación muestral, se usasen la media y desviación de la población, el estadístico es el Z-score. El problema es que dichos valores suelen ser desconocidos. En cualquier caso, en la literatura, suele llamarse Z-score al estadístico t.



## 6 Inferencia estadística

### 6.1 Introducción

En el apartado anterior hemos visto qué forma tiene la muestra (los datos de la BD) y los estadísticos que pretenden resumir dicha información. De esto se ocupa la **Estadística Descriptiva**.

Ahora bien, si queremos extrapolar resultados a toda la población de la que se supone que se ha extraído la muestra, debemos usar técnicas de **Inferencia Estadística**.

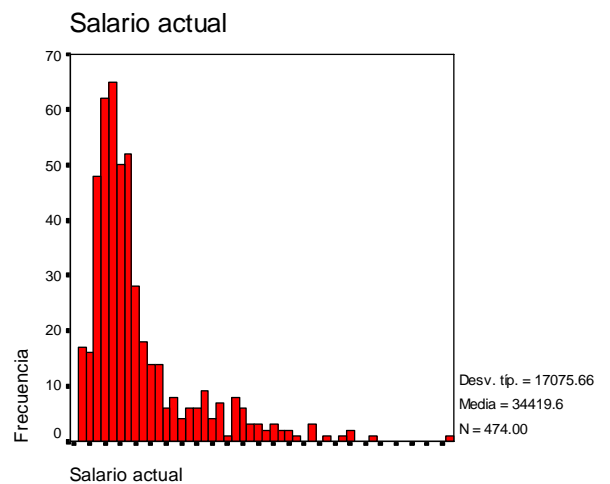
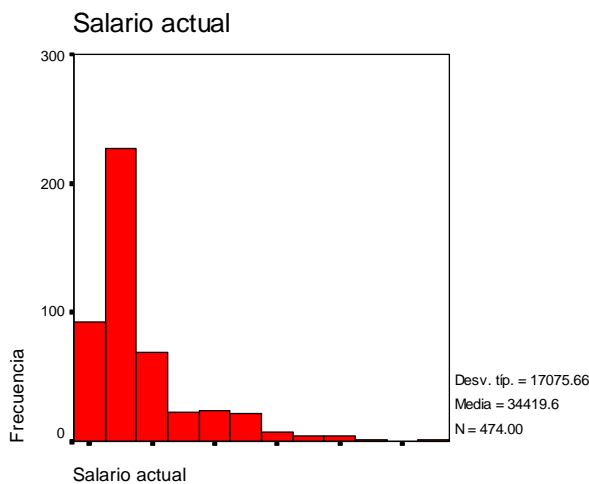
La inferencia estadística es el proceso de obtener conclusiones sobre una población a partir del análisis de una muestra. En la medida en que la muestra sea representativa de la población, los resultados podrán generalizarse.

Por tanto, asumimos que los datos con los que trabajamos podrían haber sido otros que difiriesen algo de los actuales. La Estadística ofrece métodos para “garantizar” que las medidas que construyamos son suficientemente fiables.

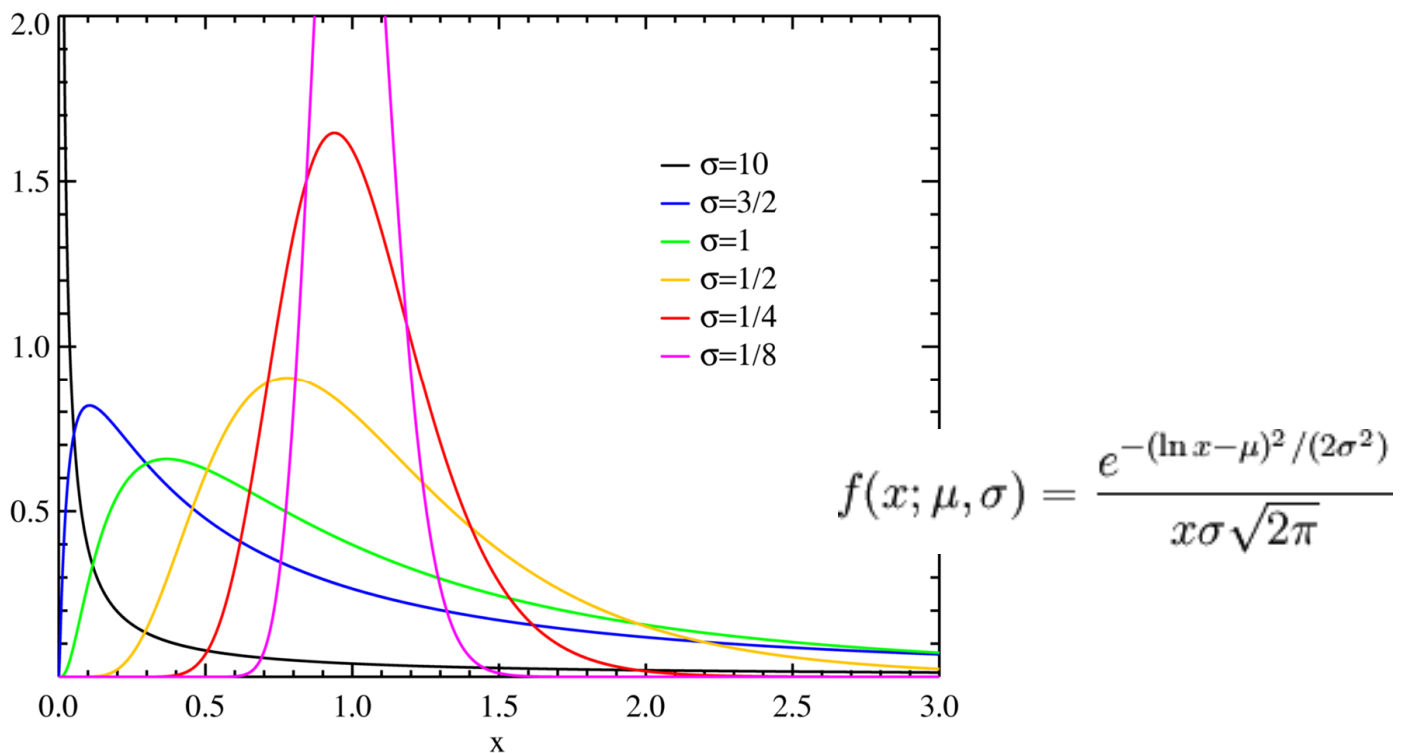
En Estadística clásica (paramétrica) se hace una suposición fundamental. Se supone que los valores que toma una variable están determinados por una probabilidad que se puede describir a través de una función (función de densidad de probabilidad)

Normalmente, dicha distribución no se conoce a priori, pero *echando un vistazo* a la distribución muestral, el experto puede suponer que se ajusta a una función determinada, o mejor dicho, a una familia de funciones. Una vez fijada la *forma* de la distribución, el procedimiento estadístico tratará de inferir cuál de todas ellas es la que mejor se ajusta a los datos.

Recordemos el histograma del Salario Actual:



Si extrapolamos el histograma, obtendríamos una función matemática. En Estadística se han estudiado muchas funciones. Por ejemplo, la anterior se asemeja a una distribución log-normal:

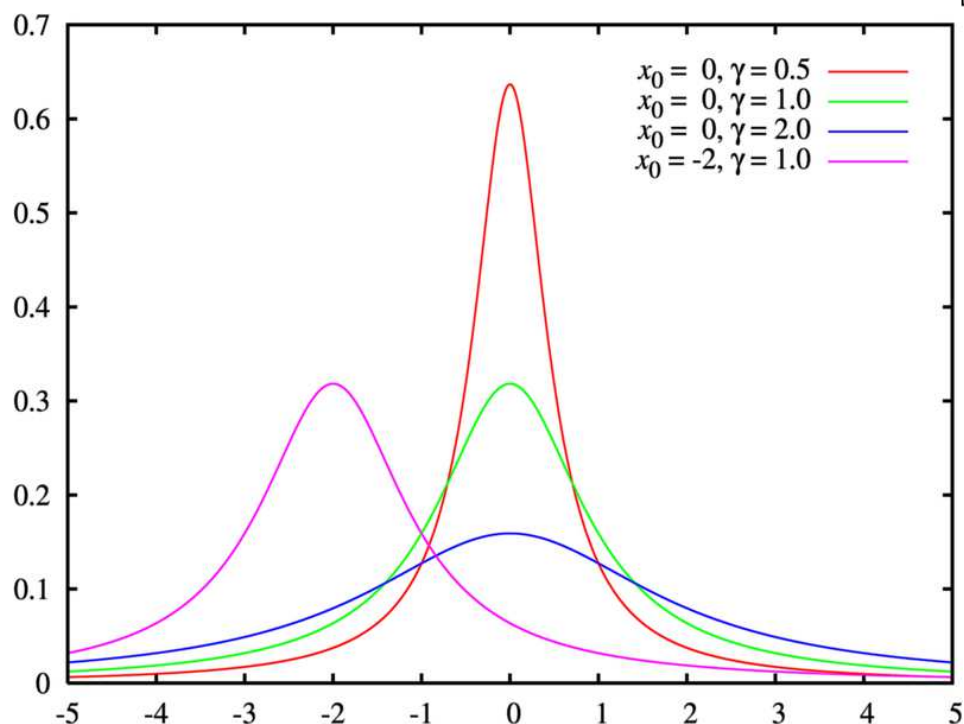


En el eje de las abscisas se representan los valores que puede tomar la variable (por ejemplo la edad)

En el eje de las ordenadas se representan los valores de probabilidad (entre 0 y 1). La función (*función de densidad*) determina la probabilidad con la que se da cada valor. El área debe ser igual a 1. Esta es una de las restricciones de la teoría de la probabilidad. Su relajación da lugar a distintas teorías que se estudiarán en el Master.

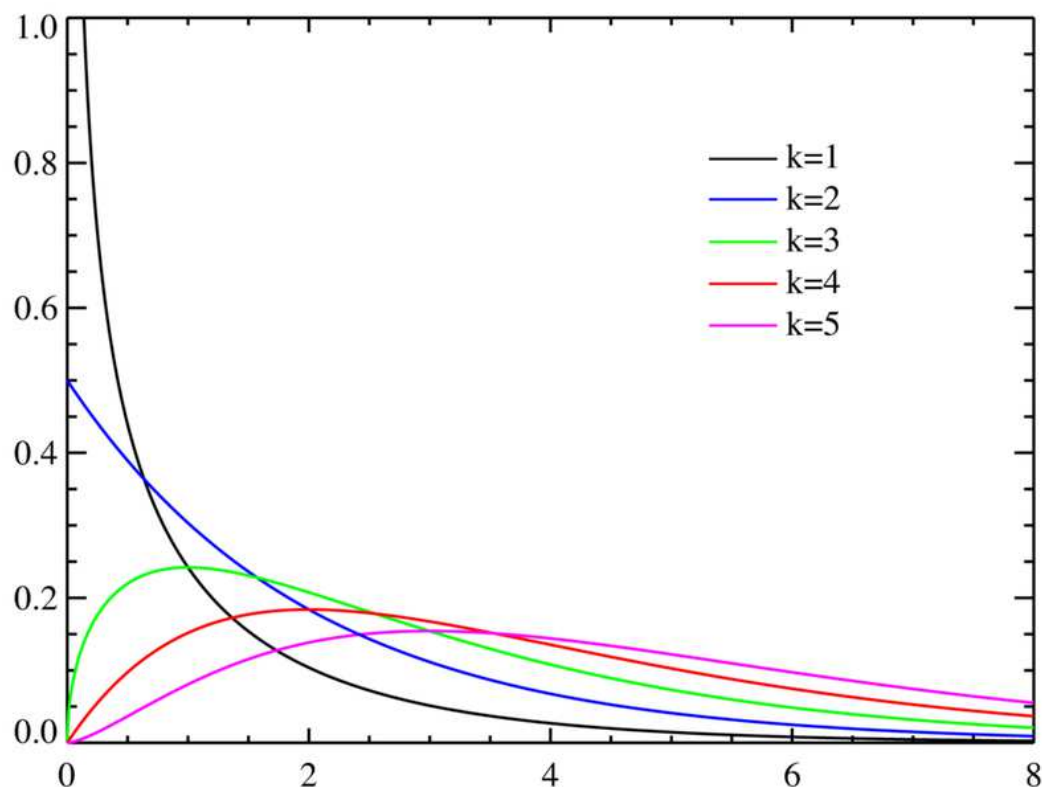
Distribución de Cauchy (parámetros:  $x_0$  y  $\gamma$ ):

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[ 1 + \left( \frac{x-x_0}{\gamma} \right)^2 \right]}$$



Distribución Chi Cuadrado (parámetro k):

$$f(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & \text{para } x \geq 0, \\ 0 & \text{para } x < 0 \end{cases}$$

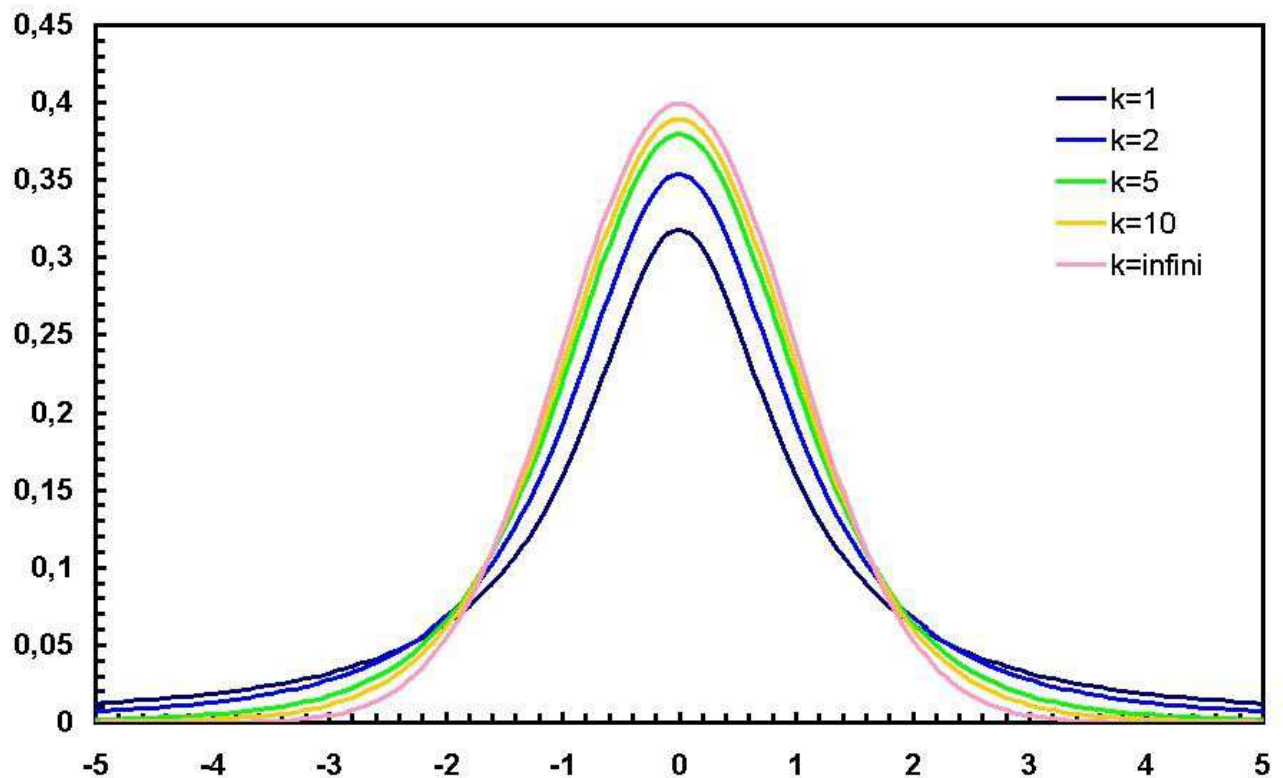


$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

La distribución **t-Student** depende de un único parámetro  $k$ , denominado **grado de libertad**.

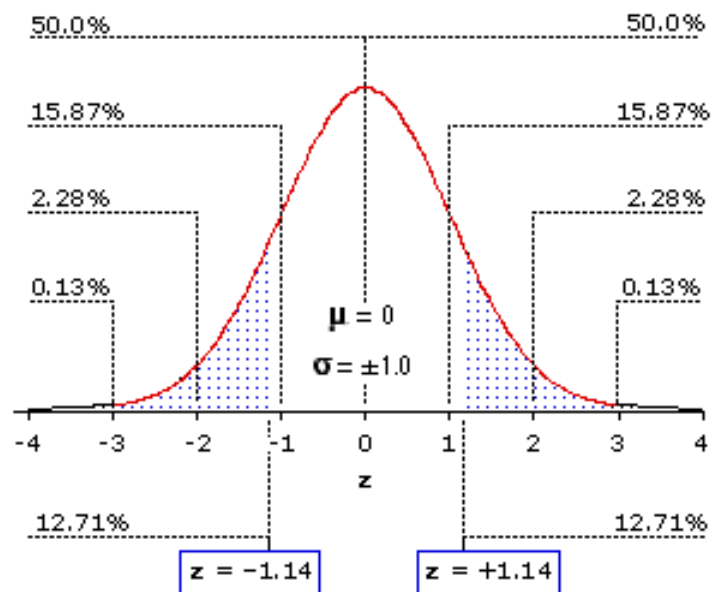
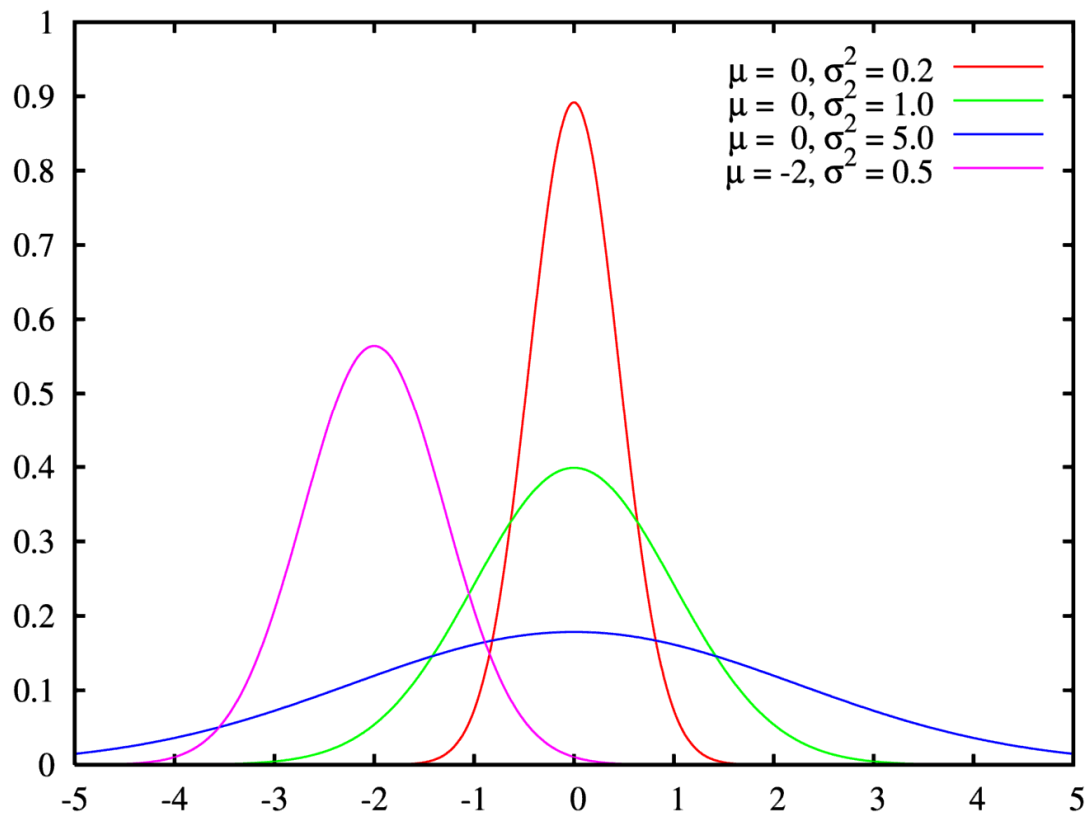
$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\Gamma(k/2)} (1 + t^2/k)^{-(k+1)/2}$$

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$



La famosa distribución normal para una variable numérica (de medida de escala). Parámetros:  $\mu$  y  $\sigma$

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



Dependiendo de lo queramos inferir tenemos distintos tipos de procedimientos estadísticos:

## 6.2 Gráficos Q-Q

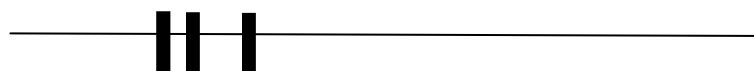
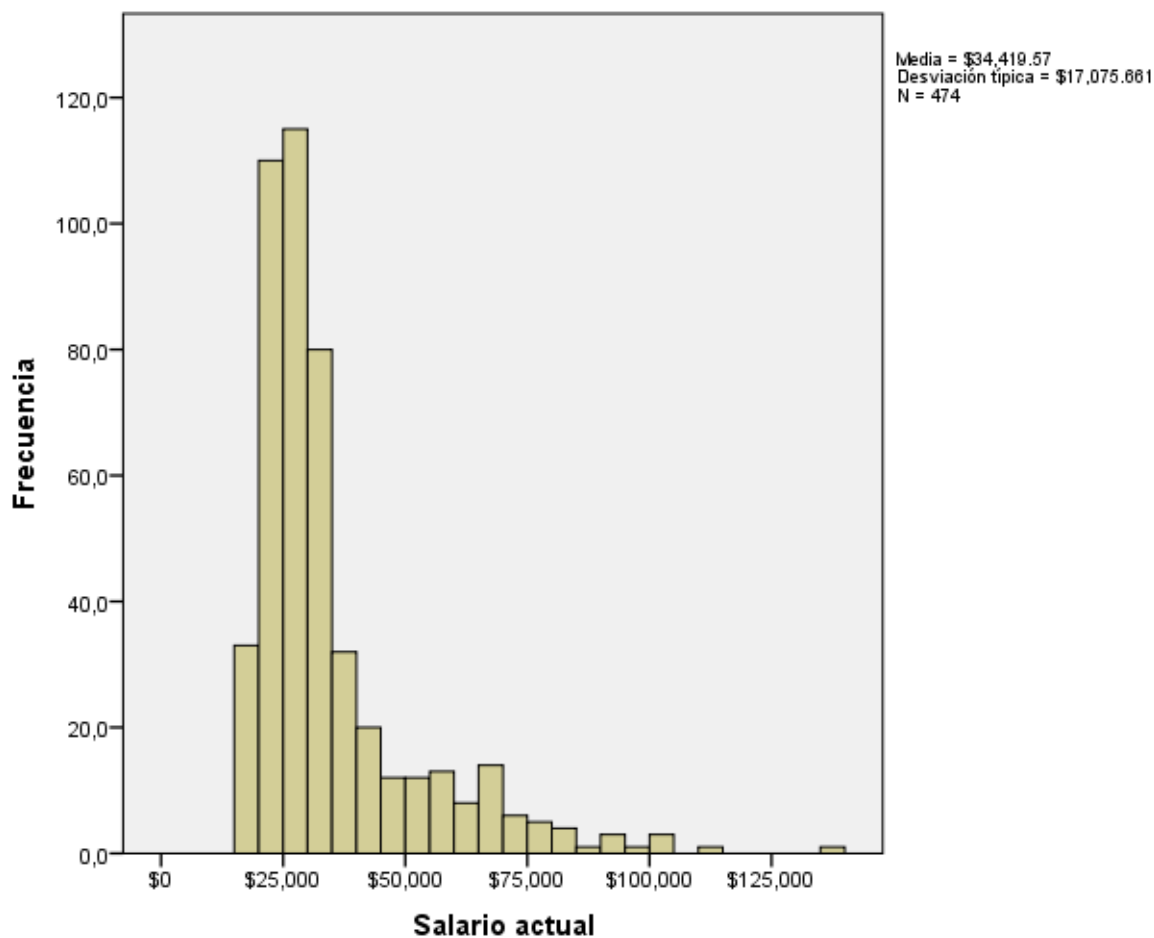


¿Cómo podemos comprobar si unos datos se ajustan a una distribución concreta?

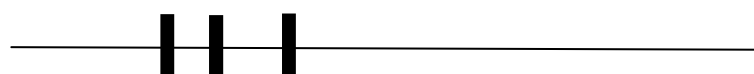
- Análisis Exploratorio: Gráfico Q-Q
- Test de Hipótesis: Lo vemos posteriormente

Podríamos comparar el histograma de los datos con la curva de la función de densidad, pero la forma de los histogramas puede variar ligeramente según sean los intervalos escogidos. Una técnica mejor la proporcionan los gráficos Q-Q.

En un gráfico Q-Q se comparan los cuantiles de la muestra (datos) y se comparan con los cuantiles de la distribución de contraste. Si están en una línea recta, puede asumirse que los datos se ajustan a dicha distribución.



Se comparan estos cuantiles con los cuantiles de una distribución lognormal similar



Se dibujan los puntos (cuantil(i) de los datos, cuantil(i) de la distribución).  
Si se ajusta bien, los puntos estarán en una línea recta

## ⇒ Analizar/Estadísticos Descriptivos/Gráficos Q-Q

Gráficos Q-Q

Variables:

Salario actual [salario]

Distribución de contraste:

Lognormal

gt:

Parámetros de distribución:

☒ Estimar a partir de los datos

Escala: 1

Forma: 1

Fórmula de estimación de la proporción:

☒ De Blom ☐ Rankit ☐ De Tukey

☐ De Van der Waerden

Rango asignado a empates:

☒ Media ☐ Mayor ☐ Menor

☐ Romper los empates arbitrariamente

Transformar:

☐ Transformación log natural

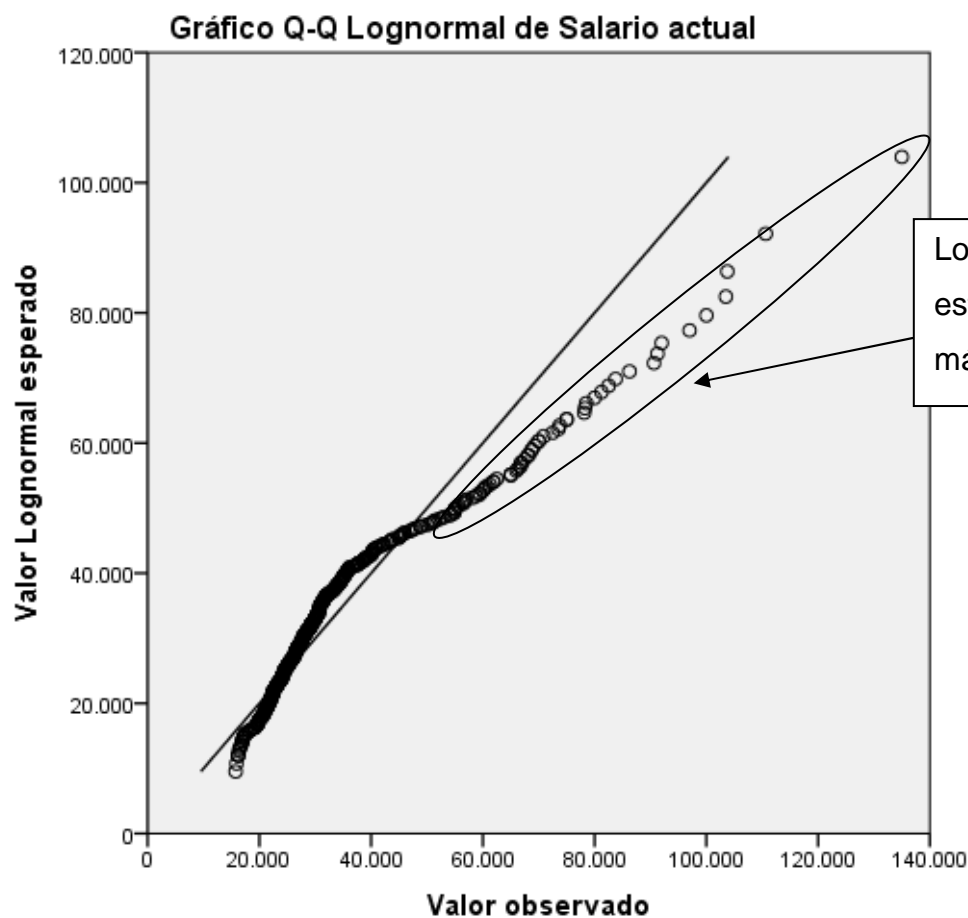
☐ Valores tipificados

☐ Diferencia: 1

☐ Diferenciar ciclo: 1

Periodicidad actual: Ninguna

Aceptar Pegar Restablecer Cancelar Ayuda





### 6.3 Estimación Puntual

Las funciones anteriores dependen de ciertos parámetros. Conociendo dichos parámetros conoceríamos toda la función y tendríamos descrita la probabilidad con la que dicha variable puede tomar valores en el eje de las abscisas.

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

Depende de  $\mu$  y  $\sigma$ .

$$f(x) = \frac{\Gamma((x+1)/2)}{\sqrt{k\pi}\Gamma(k/2)} (1 + t^2/k)^{-(k+1)/2}$$

Depende de  $k$ .

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[ 1 + \left( \frac{x-x_0}{\gamma} \right)^2 \right]}$$

Depende de  $x_0$  y  $\gamma$ .

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

Depende de  $\mu$  y  $\sigma$ .

Realmente, no llegaremos a *conocer* dichos parámetros, sino que los estimaremos con un estadístico, es decir, con un valor obtenido a partir de los datos muestrales.

Por ejemplo, en el caso de la Normal, un buen estimador de  $\mu$  es la media aritmética de la muestra y un buen estimador de  $\sigma$  es la desviación típica  $S$  de la muestra.

¿Qué significa "buen estimador"? Que cumple ciertas propiedades como por ejemplo que sea consistente (cuantos más datos tengamos, más se aproxima al parámetro), que sea invariante frente a cambios de localización o escala, que sea suficiente, insesgado, etc. En Estadística se analiza distintas propiedades que un estimador debería cumplir y para cada distribución se proponen "buenos estimadores" (estimadores que cumplen un buen número de éstas propiedades) de los parámetros de dicha distribución.

Pero se presentan dos problemas:

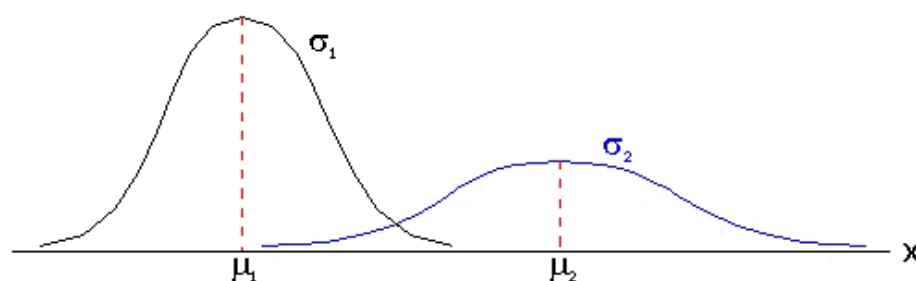
1. A veces no es posible encontrar un buen estimador
2. A veces no es posible ajustar una distribución concreta

En estos casos, no podemos hacer una estimación de los parámetros de la función de densidad de probabilidad pero sí podemos realizar una buena estimación de algunas "características" de la distribución.

Dada una función de densidad  $f(x)$ :

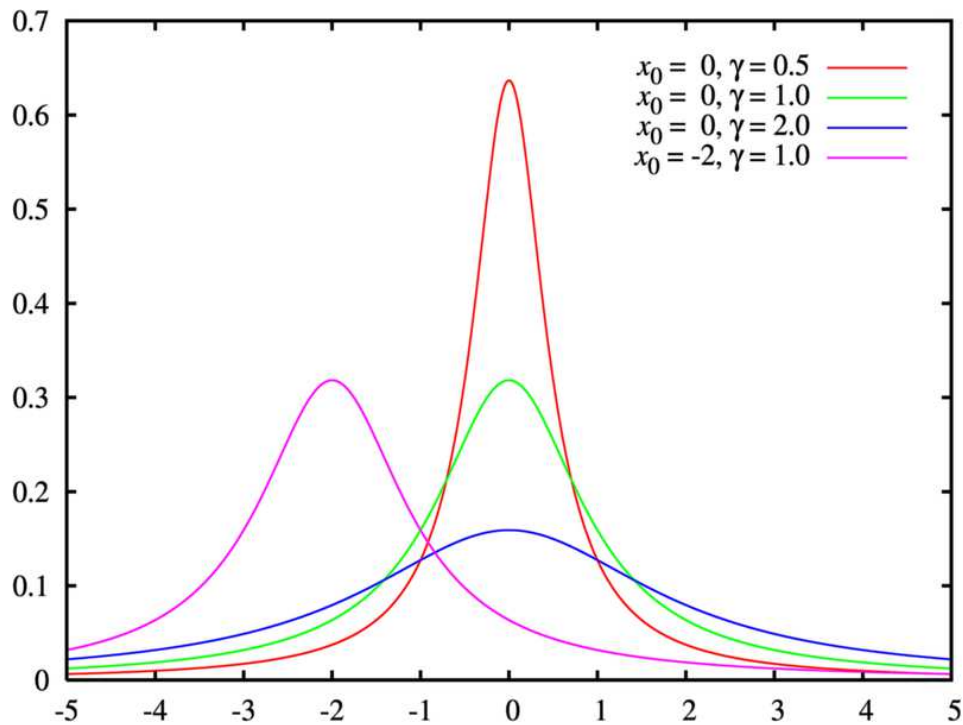
- El número calculado en la forma  $\int x f(x) dx$  es conocido como la **Esperanza (poblacional) de X**,  $E(X)$ , y se denota por  $\mu$ . Es la conocida medida de tendencia central que mide "el valor medio" de la población.
- El número calculado como  $\int (x-E(X))^2 f(x) dx$  es conocido como la **Varianza (poblacional) de X**. Observad que es igual a la esperanza de  $(X-E(X))^2$ , es decir,  $E((X-E(X))^2)$ . La raíz cuadrada positiva de este valor se conoce como la Desviación Típica de X. Se denota por  $\sigma$ . Es la conocida medida de dispersión que mide cómo de dispersos están los valores de la población.

Cuando  $f(x)$  es la densidad de una Normal, resulta que la esperanza es el propio parámetro  $\mu$  (por eso se le llama así) y la desviación típica es el propio parámetro  $\sigma$  (por eso se le llama así)



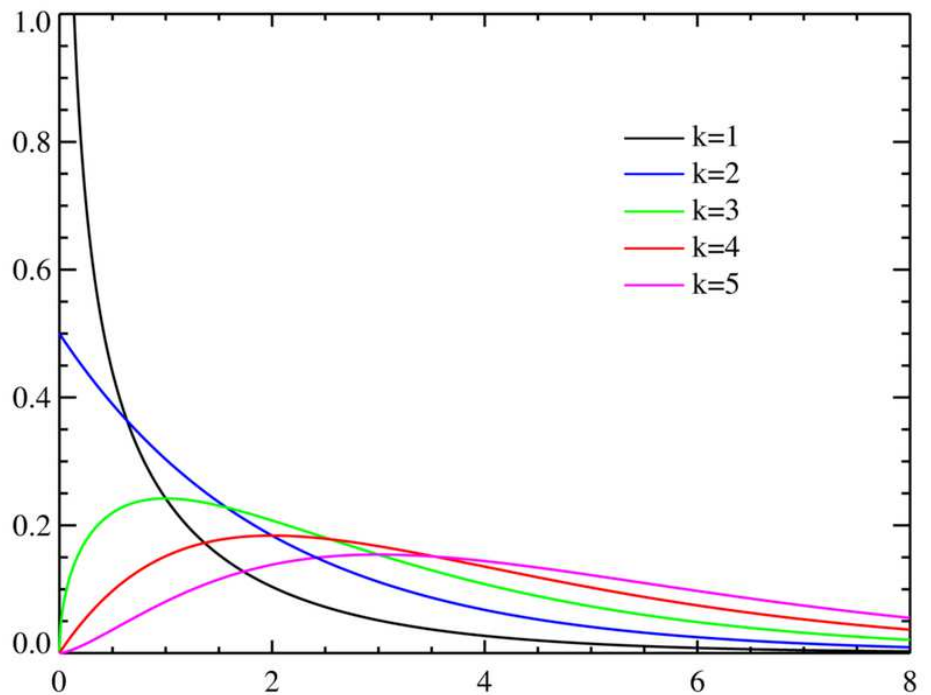
The normal curves with  $\mu_1 \neq \mu_2$  and  $\sigma_1 < \sigma_2$ .

Por ejemplo, en la distribución de Cauchy:



La curva roja ( $\gamma = 0.5$ ) corresponde a una distribución con una esperanza o valor medio superior a la de la curva morada ( $\gamma = 1.0$ )

En la distribución de Cauchy y en casi todas las distribuciones distintas de la normal, la desviación típica no es ningún parámetro de la distribución. El parámetro de la distribución de Cauchy es  $k$ :



La curva azul ( $k=2$ ) corresponde a una distribución con menos desviación que la de la curva morada ( $k=5$ ). Los valores están más dispersos.

Hay un resultado importantísimo (**Teorema Central del Límite**) y es que la media es un buen estimador de la esperanza de la población, sea cual sea la distribución subyacente de la población. Además, también se verifica que la varianza muestral  $S$  es una buena estimación de la varianza poblacional. En el caso de la varianza, la estimación no es "tan buena" como en el caso de la media (en el sentido de las propiedades que comentábamos anteriormente) y formalmente, el factor de ponderación es  $1/(n-1)$  en vez de  $1/n$ , siendo  $n$  el tamaño de la muestra. Para valores altos de  $n$ , también se cumple que la desviación típica muestral (raíz cuadrada de la varianza) es un buen estimador de la desviación muestral.

¿Cómo se formaliza matemáticamente este importante resultado? Lo vemos para la media:

Dadas  $n$  v.a.i.i.d (sea cual sea su función de densidad), la v.a.  $(X_1 + \dots + X_n)/n$  se distribuye **aproximadamente** según una Normal( $\mu$ ,  $\sigma/\sqrt{n}$ )

Es más, se puede demostrar también que:

**Dadas  $n$  v.a.i.i.d (sea cual sea su función de densidad), la v.a.  $(X_1 + \dots + X_n)/n$  se distribuye **aproximadamente** según una Normal( $\mu$ ,  $S/\sqrt{n}$ )**

Claro está, la primera aproximación es *mejor* pero requiere el conocimiento de  $\sigma$ , lo que en la realidad es inviable.

Así pues, aunque no seamos capaces de estimar cuál es la densidad de la población, al menos estamos estimando características importantes suyas. Recordad que en el caso de la distribución Normal, dichos valores (media y desviación) sí son estimaciones de los dos parámetros de los que depende la distribución.

En resumen:

- Para cualquier distribución estadística, la media y la desviación típica muestral, son unos buenos estimadores de la esperanza y la desviación típica de la población, lo que nos dará una idea resumen del comportamiento de la distribución.
- En el caso de la Normal, el parámetro  $\mu$  es la esperanza de la distribución y el parámetro  $\sigma$  es la desviación de la distribución. Al estimar dichos parámetros con la media y desviación muestral, estamos estimando los parámetros que determinan la función de densidad.

## 6.4 Estimación por intervalos de confianza



La idea es sencilla. En vez de decir que la estimación del valor medio de la aceleración es 34419.56, queremos dar un intervalo que contenga al *verdadero* valor con una alta probabilidad. Para ello, se utiliza la distribución del estadístico correspondiente que estima al parámetro que queremos estudiar.

Por ejemplo, la media muestral tiene una distribución aproximada  $\text{Normal}(\mu, S/\sqrt{n})$  según el TCL. Cuanto mayor sea  $S$  (la dispersión de los datos) y menor sea  $n$  (el número de datos), peor será la precisión de la estimación y mayor será la anchura del intervalo.

Como ocurre en SPSS, debemos irnos a otro menú distinto de los vistos anteriormente.

⇒ **Analizar/Estadísticos Descriptivos/Explorar/** En variables "dependientes" seleccionamos Salario Actual y en estadísticos seleccionamos Intervalos al 95%.

Descriptivos			Estadístico	Error típ.
Categoría laboral	Media		1.41	3.55E-02
	Intervalo de confianza para la media al 95%	Límite inferior	1.34	
		Límite superior	1.48	
	*****		*****	

Como puede apreciarse, SPSS sólo muestra IC para la media. Si quisiéramos un IC para otro parámetro de la población (por ejemplo la desviación típica) tendríamos que calcularlo a mano.

## 7 Contraste o Test de Hipótesis

### 7.1 Construcción del test

En numerosas ocasiones, el experto estadístico está interesado en comprobar si se puede aceptar o rechazar una hipótesis que él plantea **a priori**, de forma **explícita**. Dicha hipótesis se suele denotar por  $H_0$  y se conoce como hipótesis nula. Los test de hipótesis son un mecanismo estadístico que permiten rechazar una hipótesis nula planteada explícitamente.

**La idea es comprobar si los datos de la muestra concuerdan o no con  $H_0$ . Siempre hay que contrastar frente a una hipótesis alternativa llamada  $H_1$ .**

**Por ejemplo**, podemos estar interesados en comprobar si la media (esperanza de la población) es igual o distinta a un valor fijado de antemano.

$$H_0. \mu = 25000$$

$$H_1. \mu \neq 25000$$

Claro está, podríamos haber construido un IC al 95% (por ejemplo) y comprobar si 25.000 está en dicho intervalo. Hacemos lo mismo pero con otro mecanismo: los TH.

¿Cómo se construye matemáticamente el procedimiento estadístico de los TH? Vamos a ver los pasos necesarios, y para ilustrarlo, seguiremos el ejemplo anterior.

$$H_0. \mu = 34000$$

$$H_1. \mu \neq 34000$$

- Como siempre, se supone que la muestra corresponde a una función de densidad conocida con parámetros desconocidos
- Se usa un estadístico T. Como es una función de la muestra, lo llamamos  $T(X_1, \dots, X_n)$   
En nuestro ejemplo, utilizamos el estadístico media muestral:

$$T(X_1, \dots, X_n) = (X_1 + \dots + X_n)/n = \bar{X}$$

- **El secreto está en escoger un estadístico tal que cuando se calcule su distribución  $f_0$  suponiendo que  $H_0$  sea cierta, no aparezcan parámetros desconocidos.**

Por ejemplo, si la muestra corresponde a una  $N(\mu, \sigma)$ , recordemos que habíamos dicho que la media muestral se distribuía según una  $N(\mu, \sigma/\sqrt{n})$ , o lo que es lo mismo,  $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$  se distribuye según una  $N(0,1)$ . Cuando  $H_0$  es cierta, en nuestro caso,  $\mu = 34000$ , resulta que el estadístico se nos queda  $\frac{\sqrt{n}(\bar{X} - 34000)}{\sigma}$ . Sin embargo, el valor de  $\sigma$  sigue siendo desconocido.

Pero se puede demostrar que  $\frac{\sqrt{n}(\bar{X} - \mu)}{S}$  se distribuye según la función de densidad de una t-Student con  $n-1$  grados de libertad (los grados de libertad representan un parámetro en esta función). Se denota por  $t_{(n-1)}$ . Y cuando  $H_0$  es cierta, resulta que el estadístico se nos queda así:

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim t_{n-1}$$

y todos los valores son conocidos!

Nota: En aquellos casos en los que no podamos conocer la distribución exacta, podremos utilizar aproximaciones asintóticas parecidas a la del Teorema Central del Límite.

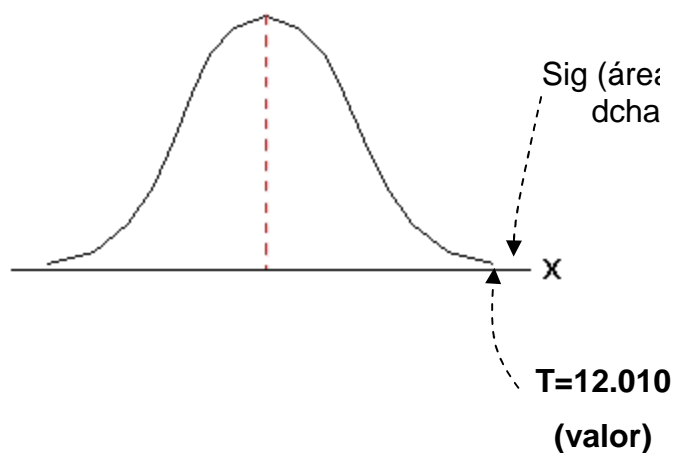


- Se calcula el valor numérico concreto que T toma en la muestra  $X_1, \dots, X_n$ . Lo llamamos **T**. (es el valor llamado t en SPSS)

En nuestro caso, sería  $\frac{\sqrt{n}(\bar{X} - 34000)}{S} = 12.010$

- Ahora, basta comprobar si dicho valor es poco probable que se de en la distribución definida por  $f_0$ . Para ello, el experto fija un **nivel de significación**  $\alpha$  como por ejemplo 95% (0.05) o 99% (0.01). Si el valor que sale de **T** está en la zona de valores que se dan con menos probabilidad que  $\alpha$ , entonces, por reducción al absurdo, deberíamos rechazar la hipótesis nula  $H_0$ , ya que si la aceptamos, **T** no debería tener una probabilidad baja. En caso contrario no se rechaza.

En vez de fijar  $\alpha$  podemos ver el valor de probabilidad que le corresponde a **T**. Este sería el valor denominado **p-value** que en SPSS se muestra como **Sig**. En nuestro caso es prácticamente cero y por tanto rechazamos la hipótesis nula.



La distribución de la **t- Student** es parecida a la de la normal (esta es la distribución de T cuando  $H_0$  es verdad). Esta distribución depende de un único parámetro, al que se le denomina **grado de libertad**.

Hay que destacar que la fiabilidad en los test de hipótesis la conseguimos cuando logramos rechazar la hipótesis nula. Por eso, cuando no rechazamos, debemos seguir haciendo estudios para poder llegar a aceptar la hipótesis nula (obteniendo más datos, usando otros procedimientos estadísticos, etc)

Hemos visto un ejemplo cuando la hipótesis nula es sobre un parámetro específico como la esperanza de una normal. Existen multitud de contrastes para distintos parámetros de un sin fin de distribuciones. El secreto está en conseguir el estadístico adecuado en cada caso.

$$g.l = 474 - 1 = 473$$

**TABLE B.3 (cont.)** Critical Values of the *t* Distribution

<i>v</i>	$\alpha(2): 0.50$ $\alpha(1): 0.25$	0.20 0.10	0.10 0.05	0.05 0.025	0.02 0.01	0.01 0.005	0.005 0.0025	0.002 0.001	0.001 0.0005
62	0.678	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454
64	0.678	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
66	0.678	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
68	0.678	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
72	0.678	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
74	0.678	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
76	0.678	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
78	0.678	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
82	0.677	1.292	1.664	1.989	2.373	2.637	2.885	3.193	3.413
84	0.677	1.292	1.663	1.989	2.372	2.636	2.883	3.190	3.410
86	0.677	1.291	1.663	1.988	2.370	2.634	2.881	3.188	3.407
88	0.677	1.291	1.662	1.987	2.369	2.633	2.880	3.185	3.405
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
92	0.677	1.291	1.662	1.986	2.368	2.630	2.876	3.181	3.399
94	0.677	1.291	1.661	1.986	2.367	2.629	2.875	3.179	3.397
96	0.677	1.290	1.661	1.985	2.366	2.628	2.873	3.177	3.395
98	0.677	1.290	1.661	1.984	2.365	2.627	2.872	3.175	3.393
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
105	0.677	1.290	1.659	1.983	2.362	2.623	2.868	3.170	3.386
110	0.677	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381
115	0.677	1.289	1.658	1.981	2.359	2.619	2.862	3.163	3.377
120	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
125	0.676	1.288	1.657	1.979	2.357	2.616	2.858	3.157	3.370
130	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
135	0.676	1.288	1.656	1.978	2.354	2.613	2.854	3.152	3.364
140	0.676	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361
145	0.676	1.287	1.655	1.976	2.352	2.610	2.851	3.147	3.359
150	0.676	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357
160	0.676	1.287	1.654	1.975	2.350	2.607	2.846	3.142	3.352
170	0.676	1.287	1.654	1.974	2.348	2.605	2.844	3.139	3.349
180	0.676	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.345
190	0.676	1.286	1.653	1.973	2.346	2.602	2.840	3.134	3.342
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
250	0.675	1.285	1.651	1.969	2.341	2.596	2.832	3.123	3.330
300	0.675	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323
350	0.675	1.284	1.649	1.967	2.337	2.590	2.825	3.114	3.319
400	0.675	1.284	1.649	1.966	2.336	2.588	2.823	3.111	3.315
450	0.675	1.283	1.648	1.965	2.335	2.587	2.821	3.108	3.312
500	0.675	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310
600	0.675	1.283	1.647	1.964	2.333	2.584	2.817	3.104	3.307
700	0.675	1.283	1.647	1.963	2.332	2.583	2.816	3.102	3.304
800	0.675	1.283	1.647	1.963	2.331	2.582	2.815	3.100	3.303
900	0.675	1.282	1.647	1.963	2.330	2.581	2.814	3.099	3.301
1000	0.675	1.282	1.646	1.962	2.330	2.581	2.813	3.098	3.300
$\infty$	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

..... 12

$T=12.010 \gg 1.965$ , por lo que se rechaza la hipótesis nula. El p-value correspondiente a 12.010 con 473 g.l es prácticamente cero.

⇒ Analizar/Comparar Medias/Prueba T para una muestra/Seleccionad como valor de prueba 25000.

gl = grado de libertad. Parámetro de la t-student

$H_0: \mu = 25000$

**Prueba para una muestra**

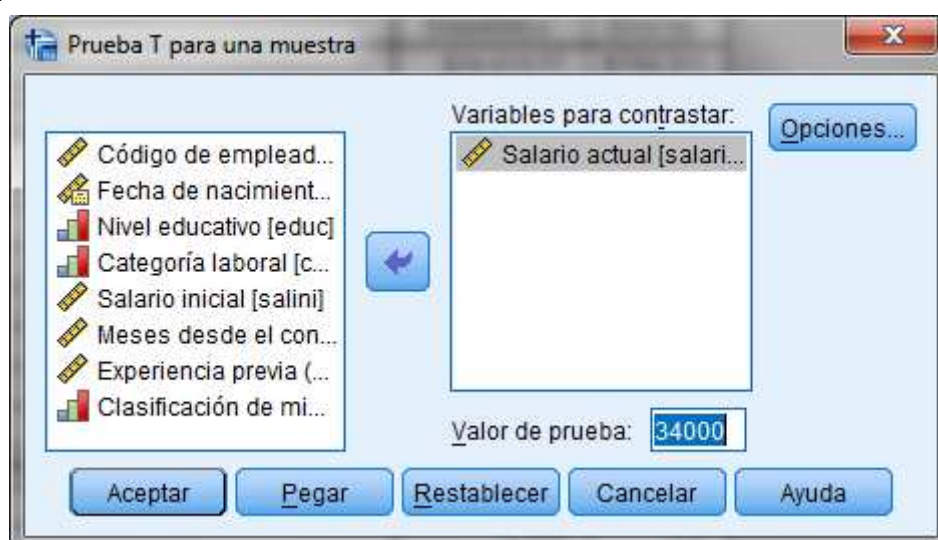
Valor de prueba = 25000						
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Salario actual	12.010	473	.000	\$9,419.57	\$7,878.40	\$10,960.73

Valor T del estadístico

p-value

Si la Sig. es un valor muy pequeño (usualmente se acepta como pequeño un valor por debajo de 0.05) rechazaremos la hipótesis nula. En caso contrario, no diríamos que la aceptamos sino que la muestra no contradice la hipótesis nula. En este caso rechazamos que la media pueda ser igual a 25000.

⇒ Analizar/Comparar Medias/Prueba T para una muestra/Seleccionad como valor de prueba 34000.



Comprobad que no se rechaza que la media pueda ser igual a 34000

### Prueba para una muestra

	Valor de prueba = 34000					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Salario actual	.535	473	.593	\$419.57	-\$1,121.60	\$1,960.73

valor muy alto > 0.05

En estas pruebas, se presupone que los datos se distribuyen según una normal, aunque el test es bastante robusto frente a desviaciones de esta hipótesis. Por eso, lo hemos aplicado al salario, aunque ésta se asemeje más a una log-normal. Para desviaciones importantes, habrá que aplicar un test no paramétrico (ver al final)

## 7.2 T (t-Student)-tests

Hay muchos tipos de test de hipótesis. Los que involucran la comparación de la media de un grupo con un valor concreto, o la comparación de dos medias de dos grupos o dos medias de dos variables, se denominan T-tests:

- Comparar si el valor medio de una variable es igual a un valor concreto.

SPSS: Prueba **T para una muestra**.

El visto anteriormente.

- Comparar si las medias de dos grupos de casos de una misma variable son iguales.

SPSS: Prueba **T para muestras independientes**.

- Comparar si las medias de dos variables "pareadas" son iguales **-paired test-**.

SPSS: Prueba **T para muestras relacionadas**

Si tenemos más de 2 grupos, debemos recurrir al ANOVA (se ve después)

## Prueba T (t-Student) para muestras independientes

Se tienen dos muestras de sendas variables, no necesariamente del mismo tamaño. Se quiere comprobar si las medias pueden considerarse iguales o no.

Por ejemplo, se aplica un tratamiento a un grupo de individuos y un placebo a otro. La asignación de cada individuo a cada grupo ha de ser aleatoria. Se ve el nivel de algún indicador después del tratamiento y se comparan las medias de dicho nivel en cada uno de los dos grupos.

¿Cual es la base intuitiva de este test?

Podríamos intentar medir las desviaciones de cada individuo con respecto a la media global de todos, pero entonces, no estaríamos considerando la formación de grupos. Así que deberíamos medir las desviaciones de las medias de cada grupo con respecto a la media global. Pero supongamos los siguientes ejemplos:

Ejemplo 1.

Grupo A: 3 3 3 3 3 Media = 3  
Grupo B: 3.5 3.5 3.5 3.5 3.5 Media = 3.5  
Media global: 3.25

Muy poca variabilidad intra-grupo

Ejemplo 2.

Grupo A: 1 3 4 4 Media = 3  
Grupo B: 2 3 4 5 Media = 3.5  
Media global: 3.25

Alta variabilidad intra-grupo

En el Ejemplo 1 el Grupo A difiere del grupo B, más de lo que lo hace en el Ejemplo 2, ya que la distribución subyacente del segundo sugiere que podríamos haber obtenido perfectamente una media de 3.25, de 3 o de 3.5. Sin embargo, en el primero parece que siempre obtenemos el mismo valor, por lo que no será fácil obtener una media distinta de 3.25 (3.5 respectivamente)

Estos ejemplos ponen de manifiesto que debemos tener en cuenta la variabilidad dentro de cada grupo, además de la variabilidad entre los grupos. En definitiva, el estadístico que se construya debería tener en cuenta la diferencia entre las medias de cada grupo y la dispersión en los grupos.

$$H_0. \mu_1 = \mu_2$$

$$H_1. \mu_1 \neq \mu_2$$

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

⇒ Analizar/Comparar Medias/Prueba T para muestras independientes



**Estadísticos de grupo**

	Sexo	N	Media	Desviación típ.	Error típ. de la media
Salario actual	Hombre	258	\$41,441.78	\$19,499.214	\$1,213.968
	Mujer	216	\$26,031.92	\$7,558.021	\$514.258

**Prueba de muestras independientes**

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia		Inferior	Superior
Salario actual	Se han asumido varianzas iguales	119,669	,000	10,945	472	,000	\$15,409.862	\$1,407.906		\$12,643.322	\$18,176.401
	No se han asumido varianzas iguales			11,688	344,262	,000	\$15,409.862	\$1,318.400		\$12,816.728	\$18,002.996



## Prueba T (t-Student) para muestras pareadas (paired-test)

Se tienen dos variables pareadas (por lo que las muestras no son independientes y además tienen el mismo tamaño). Se quiere comprobar si las medias pueden considerarse iguales o no. Por ejemplo, se aplica un tratamiento y se ve el nivel de un indicador antes y después del tratamiento. Los indicadores se miden sobre el mismo individuo. Tenemos tantas filas como individuos y tantas columnas como indicadores estemos midiendo, por lo que este es un caso de *medidas repetidas*.

Aunque sea análogo, en vez de poner  $H_0: \mu_1 = \mu_2$  estos tests suelen presentarse como un test sobre la diferencia de las medias:

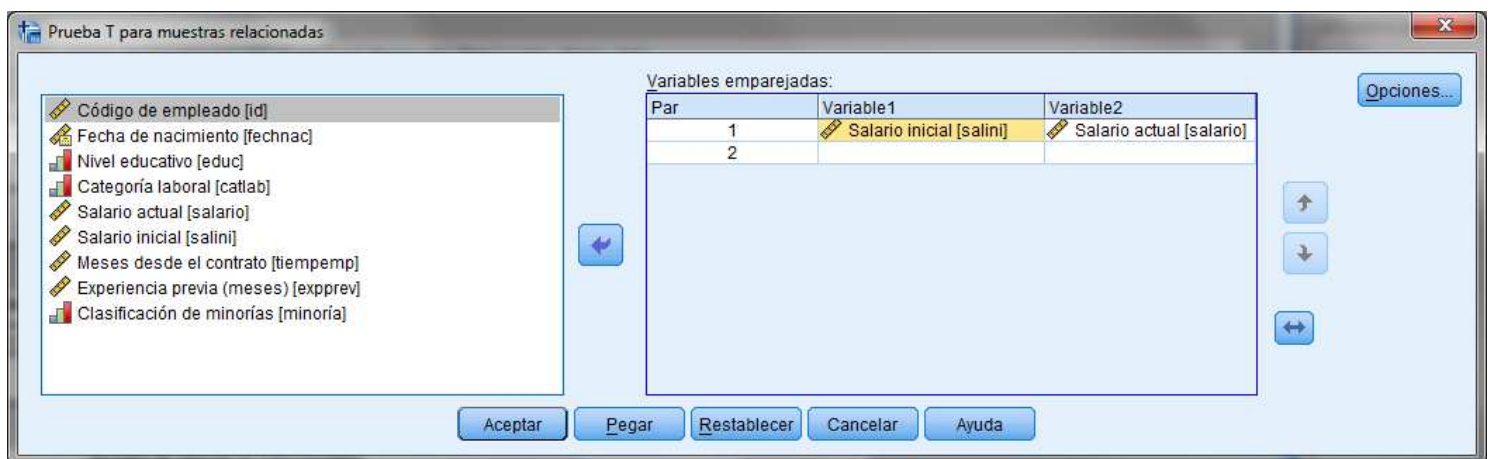
$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Se calculan las diferencias y a partir de ellas, la correspondiente media y desviación:

$$d_i = X_{1i} - X_{2i} \quad T = \frac{\bar{d}}{s_d} \sim t_{n-1}$$

⇒ Analizar/Comparar Medias/Prueba T para muestras relacionadas



Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	Salario inicial - Salario actual	-\$17,403.481	\$10,814.620	\$496.732	-\$18,379.555	-\$16,427.407	-35,036	473	,000

### 7.3 Tipos de errores. Potencia del test

El objetivo de un test es poder llegar a la conclusión de que se debe rechazar la hipótesis nula con las mayores garantías posibles.

	$H_0$ es cierta	$H_1$ es cierta
No se rechaza $H_0$	No hay error (TP)	<b>Error tipo II <math>\beta</math> (FN)</b>
Se rechaza $H_0$	<b>Error tipo I <math>\alpha</math> (FP)</b>	No hay error (TN)

¿Cuál es la probabilidad de que rechace erróneamente?

$$\text{Error tipo I} = P(\text{rechazar } H_0 \mid H_0 \text{ es cierta}) = \alpha$$

El test decide rechazar, pero se ha equivocado.

En investigación experimental, este tipo de error se conoce como Falso Positivo

Falso —————> El test se ha equivocado  
Positivo —————> El test ha decidido rechazar



Falso positivo: viene del inglés. En español sería más intuitivo decir Positivo Falso (el test ha dado positivo, es decir, ha decidido rechazar, pero se ha equivocado porque no había que rechazar)

Los test de hipótesis están diseñados para que el experto controle (fije a priori) este error y por tanto se asegure que hay pocos falsos positivos. Normalmente, se considera aceptable  $\alpha=0.05$ . Pero depende del tipo de experimento:

$H_0$ . El paciente tiene Ébola

$H_1$ . El paciente no tiene Ébola

Obviamente, habrá que imponer un valor de  $\alpha$  mucho menor que 0.05 ya que no nos podemos permitir falsos positivos (pacientes que tienen ébola pero el test decide rechazar que estén enfermos).



Pero un test podría aceptar sistemáticamente ("**no se moja**" y nunca rechaza) y por tanto no cometería nunca un error de tipo I (no hay falsos positivos porque nunca hay positivos) Obviamente, sería un test de poca utilidad. Así, pues, hay que tener en cuenta otro tipo de error.

Error no controlado: Probabilidad de que no detecte una situación en la que tiene que rechazar

$$\text{Error tipo II} = P(\text{No rechazar } H_0 \mid H_0 \text{ es falsa}) = \\ P(\text{rechazar } H_1 \mid H_1 \text{ es cierta}) = \beta$$

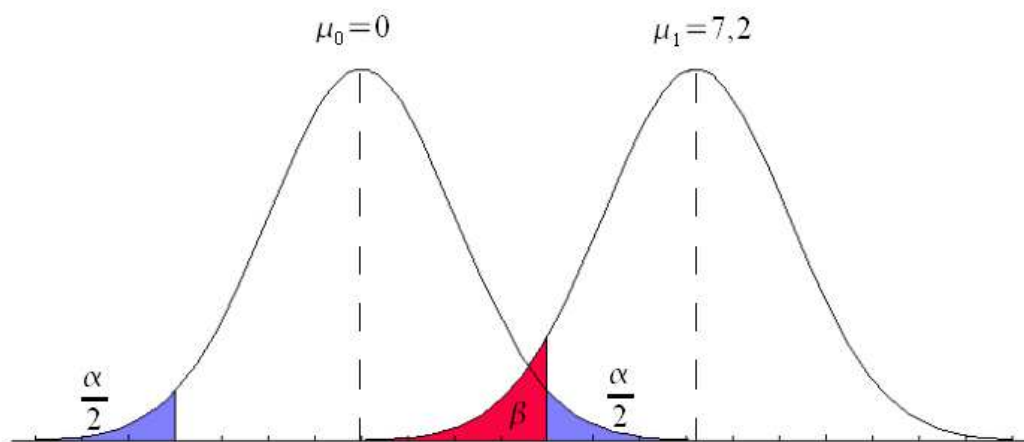
En investigación experimental, este tipo de error se conoce como Falso Negativo.

Falso —————> El test se ha equivocado  
Negativo —————> El test no ha "saltado"

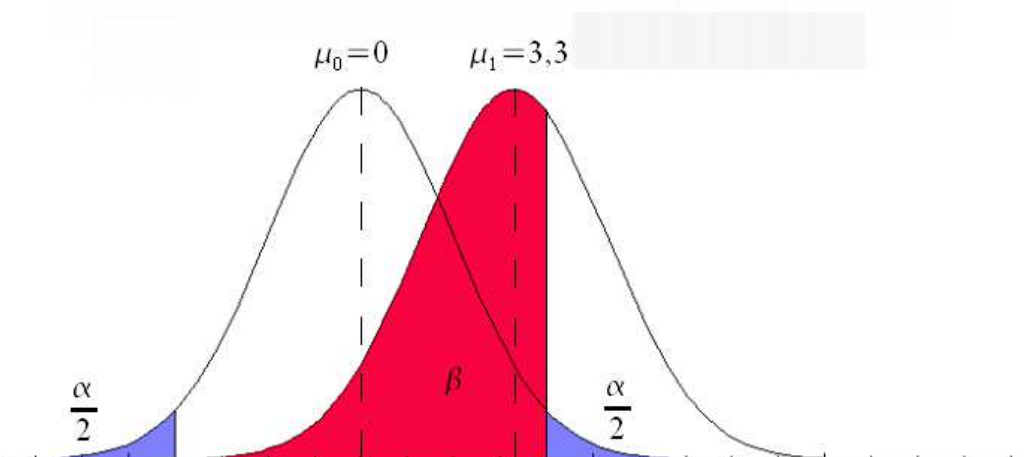


El test no ha *saltado* cuando tenía que haberlo hecho porque la hipótesis nula era falsa.

El error de tipo I y el error de tipo II están inversamente relacionados. Cuanto mayor sea uno, menor es el otro.



Ejemplo de un T-test  
comparando dos  
valores de la  
esperanza.



Ya hemos dicho que los test estadísticos están diseñados para controlar el error de tipo I. El error de tipo II cometido dependerá de:

- La propia hipótesis nula que se plantea

No es lo mismo comprobar si la media de una población próxima a 30 es igual a 40 que igual a 350.

- La *calidad* del estadístico

- El tamaño muestral. Cuanto mayor sea, menor será el error de tipo II

- La varianza de los datos. Cuanto mayor sea, mayor será el error de tipo II

Por tanto, podremos tener dos tests que, para un error tipo I fijado de antemano (0.05 por ejemplo) uno tenga un error tipo II menor que el otro. En este caso diremos que el que comete un mayor error tipo II es más **conservador** ya que logra una tasa de falsos positivos de 0.05 a costa de "no mojarse" y no decidir que hay que rechazar cuando sí se podía rechazar.

En Estadística, suele hablarse de la Potencia del Test en vez del error de tipo II, entendiendo la potencia como el complementario del error:

$$\text{Potencia del test} = P(\text{rechazar } H_0 \mid H_1 \text{ es cierta}) = 1 - \beta$$

Así pues, un test *conservador* tendrá *poca potencia*.

## 8 AED: Informes y gráficos sobre varias variables (con agrupaciones) Nominal-Nominal

SPSS no ofrece apenas información sobre variables nominales en los informes. Para ver posibles dependencias entre variables nominales utilizaremos sólo los gráficos.

Supongamos dos variables nominales. Estamos interesados en ver si existe alguna combinación de valores de variables que concentren más datos que otras. Es decir, queremos analizar cierta(s) variable(s) agrupando los valores según los que tome otra. Por ejemplo,

***¿Cómo se distribuyen las distintas categorías laborales, atendiendo al sexo?***

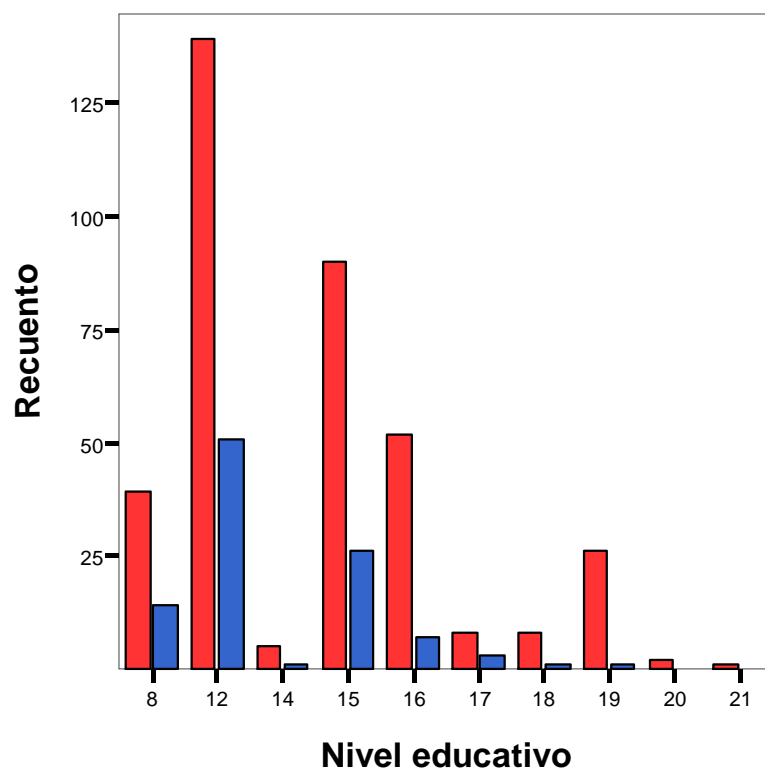
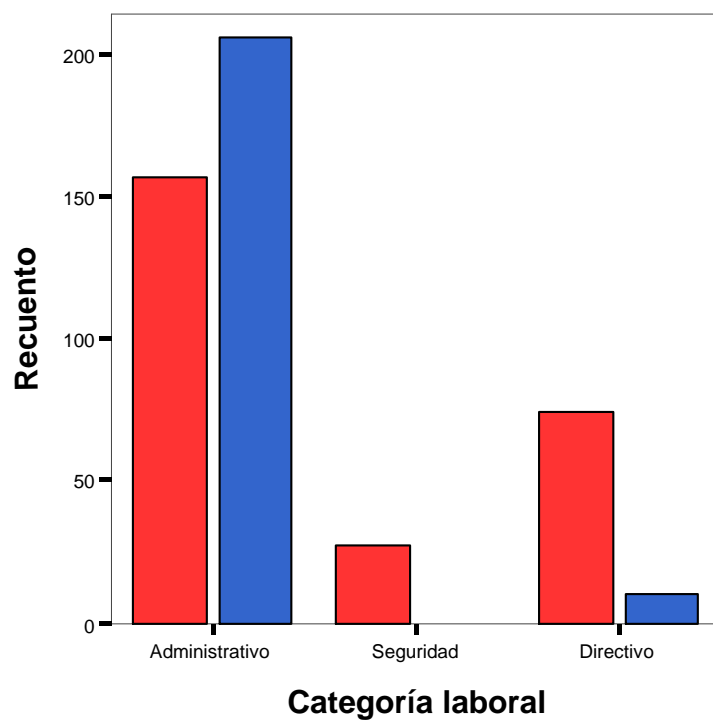
O más específicamente:

***¿Hay categorías laborales en las que predomine un sexo, en relación a las otras categorías?***

o dicho al revés,

***¿Son independientes las variables "Categoría laboral" y "Sexo"?***

⇒ **Gráficos/interactivos/barras** y seleccionamos \$count , "Categoría laboral" y como variable de leyenda "Sexo". Efectivamente, vemos que mientras que en los administrativos predomina el sexo femenino, en los otros dos predomina el masculino. En el anterior gráfico, seleccionad ahora "Nivel de estudios" y "Minoría étnica" como variable de leyenda. Vemos que no existe dependencia ninguna, ya que en todos los niveles de estudio siempre existe una **proporción** parecida entre los clasificados como minoría étnica y los que no.



## 9 Análisis Estadísticos de dependencia. Nominal-Nominal

Una vez que hemos hecho el AED y detectamos visualmente algún tipo de dependencia entre varias variables, vamos a cuantificarla, usando análisis estadísticos.

Cuando usamos variables de medida nominal (u ordinal), se utilizan las **tablas de contingencia**. La idea es simple. Se seleccionan varias variables. Por ahora dos. Se construye una tabla con tantas filas como valores distintos tenga la primera variable y tantas columnas como valores distintos tenga la segunda variable. Se cuenta el número de veces que se da cada casilla y se comprueba si la proporción de apariciones es la misma (da igual verlo por filas que por columnas)

**Tabla de contingencia Sexo \* Categoría laboral**

Recuento		Categoría laboral			Total
		Administrativo	Seguridad	Directivo	
Sexo	Hombre	157	27	74	258
	Mujer	206	0	10	216
Total		363	27	84	474

La proporción de sexos en cada categoría laboral es distinta de un sexo a otro:

Proporción de hombres en los Administrativos:  $157 / 363 \sim 0.43$

Proporción de hombres en los Directivos:  $74 / 84 \sim 0.88$

Es decir, en los Administrativos 4 de cada 10 empleados son hombres, mientras que en los directivos 9 de cada 10 empleados son hombres.

Para cuantificar estadísticamente esta diferencia en las proporciones, se aplica un test de hipótesis (conocido como test de la chi cuadrado). En este test, las hipótesis planteadas son:

¿Son las variables dependientes o independientes?

**¿Qué ponemos como hipótesis nula: son dependientes o son independientes?**

**Si lo que vamos buscando es demostrar que existe dependencia, tendremos que poner como Hipótesis nula que son independientes**, ya que, como dijimos anteriormente, lo mejor desde un punto de vista estadístico, es poder rechazar una hipótesis  $H_0$  (en este caso, rechazar independencia, para poder demostrar dependencia).

$H_0$ . Las variables son independientes (la proporción de valores de cada casilla con respecto al total de elementos de su columna es igual para todas las casillas de la misma fila)

$H_1$ . Las variables no son independientes (hay al menos dos casillas para las que no se verifica lo anterior)

Nota: Da igual considerar filas que columnas (intercambiando el papel de éstas en el anterior test)

Supongamos una tabla de contingencia de dos entradas con  $n$  filas y  $m$  columnas. Si llamamos  $O_{ij}$  al número de observaciones que caen en la casilla  $i$   $j$  de la tabla de contingencia, y si construimos los valores

$$E_{ij} = \frac{n_i}{m_j}$$

dónde

$n_i$  es el número total de elementos de la fila  $i$

$m_j$  es el número total de elementos de la columna  $j$

Entonces, se construye el estadístico siguiente:

$$T = \sum_{i=1}^n \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Cramer demostró que este estadístico, cuando la hipótesis nula es cierta (las variables son independientes) tiene una distribución ASINTÓTICA (es decir, cuando la muestra es grande) igual a la chi cuadrado (como siempre, gracias al Teorema Central del Límite). Esta distribución depende de un parámetro denominado "grado de libertad". En nuestro caso, la distribución del estadístico anterior es una chi-cuadrado con  $(n-1)*(m-1)$  grados de libertad.

El test **solo es válido** si se verifica:

- Ninguna  $E_{ij}$  es inferior a 1
- No mas del 20% de las  $E_{ij}$  son inferiores o iguales a 5

⇒ **Analizar/Estadísticos Descriptivos/Tablas de Contingencia.** En las filas seleccionamos "Sexo" y en las columnas "Categoría Laboral". Observad que, aunque el tipo de dato de Categoría es numérico, la medida es ordinal, por lo que podemos plantear una tabla de contingencia. Seleccionamos que muestre los gráficos de barras agrupadas (son los mismos que hemos generado anteriormente). En la pestaña Casillas que incluya las frecuencias observadas. En la pestaña Estadísticos, seleccionamos Chi Cuadrado (observad que cuando la variable es de medida nominal aparecen otros estadísticos aplicables).

En el ejemplo anterior, la tabla obtenida es:

**Pruebas de chi-cuadrado**

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	79,277 <sup>a</sup>	2	,000
Razón de verosimilitud	95,463	2	,000
N de casos válidos	474		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5.  
La frecuencia mínima esperada es 12,30.

Obtenemos un valor significativo de 0.000, es decir, que se rechaza que son independientes con mucha seguridad, por lo que aceptamos que existe cierta *dependencia*, es decir, que hay combinaciones de valores de las variables más frecuentes que otros, como por ejemplo Administrativo-mujer y Directivo-Hombre.

⇒ **Ejercicio:** Ejecutar la tabla de contingencia, con "Minoría étnica" y "Nivel de estudios" para comprobar que aquí sí se acepta la hipótesis de independencia. Da igual si en las filas consideramos una variable o la otra, aunque lo lógico será seleccionar como filas aquella variable con más valores; en este caso, "Nivel de estudios" (observad los gráficos de barras asociados)

Es muy importante que observemos que en la ventana de Tablas de Contingencia podemos elegir tanto las variables nominales como las numéricas. SPSS sólo debería mostrar aquellas variables de medida nominal u ordinal (aunque tuviesen un tipo numérico). Sin embargo, no es así y te permite que selecciones cualquier variable. Obviamente no tiene sentido para las

numéricas ya que crearía una fila o columna por cada valor numérico que tome la variable (eso sí, podríamos agrupar previamente en intervalos a través del menú Transformar/Recodificar de la ventana de Datos de SPSS)

Nota: En SPSS, se pueden introducir varias variables en las filas y varias variables en las columnas, pero no se construye una tabla de contingencia multidimensional, sino que se construye una tabla por cada cruce de cada variable incluida en el panel "Filas" con cada una de las variables incluidas en el panel "Columnas"

Una vez que rechazamos independencia, cabe realizar dos estudios adicionales:

1. Dar una medida que cuantifique el grado de *dependencia* entre las variables nominales. Para ello, basta seleccionar como estadístico, alguno de los proporcionados por SPSS como el coeficiente de contingencia o el coeficiente V de Cramer. Ver el tutorial de SPSS para una interpretación de dichas medidas.
2. Analizar cuáles son las casillas que más contribuyen a rechazar la independencia. SPSS no proporciona dichas medidas. En el curso de Modelos Avanzados de Data Mining se verá alguna.



Un problema importante que no da tiempo a verlo con detenimiento es el que se presenta cuando se encuentra una dependencia entre dos variables, pero causada de forma artificial por la presencia de otra variable no incluida en el estudio. Veamos un ejemplo.

La tabla siguiente corresponde a los resultados de los juicios por asesinato en Florida desde el 76 hasta el 87. Se quiere ver si influye la raza del acusado en el resultado del juicio, en el sentido de ver si se condenas a muerte a más negros que blancos.

Acusado	Pena de Muerte		% Si
	Si	No	
Blanco	53	430	10,97
Negro	15	176	7,8

Así pues, parece que hay un mayor porcentaje de condenados a muerte entre los blancos que entre los negros. Sin embargo, en un segundo estudio, consideramos otra variable más, a saber la raza de la víctima: es lo que se denomina una variable de *control*. Por cada valor de esta nueva variable, volvemos a ver los condenados a muerte blancos y negros. Los conteos son los siguientes:

Victima	Acusado	Pena de Muerte		% Si
		Si	No	
Blanco	Blanco	53	414	11.3
	Negro	11	37	22.9
Negro	Blanco	0	16	0
	Negro	4	139	2.8
Total		53	430	11.0
		15	176	7.9

Cuando la víctima es blanco:

Hay un 22.9% de condenados a muerte negros y un 11.3% de condenados a muerte blancos

Cuando la víctima es negro:

Hay un 2.8% de condenados a muerte negros y un 0% de condenados a muerte blancos

Así pues, en realidad, la proporción de condenados a muerte de blancos es menor que la de negros cuando la víctima es blanco; y lo mismo ocurre cuando la víctima es negro.

Esto es lo que se conoce como la paradoja de Simpson.

NOTA: En SPSS, las variables de control se seleccionan en el panel "Capas" del cuadro de diálogo de Tablas de Contingencia. Cuando se selecciona más de una, simplemente se realiza un cruce de la variable de fila y la de columna, con cada una de las variables especificadas en "Capas" por separado.

## 10 AED: Informes y gráficos sobre varias variables. Numérica-Numérica.

Ampliación

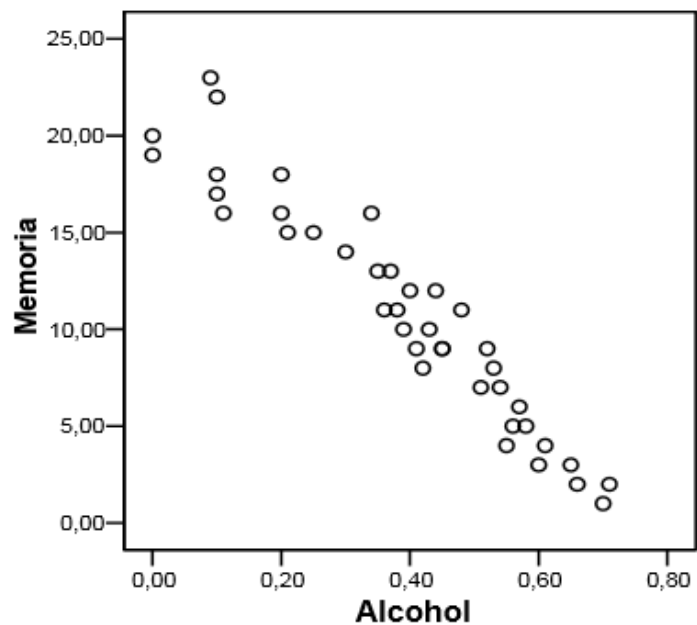
Suponemos que trabajamos con variables numéricas de escala y estamos interesados en ver si unas dependen de otras. Queremos responder a preguntas del tipo:

**¿Influye el Salario inicial en el Salario actual?**

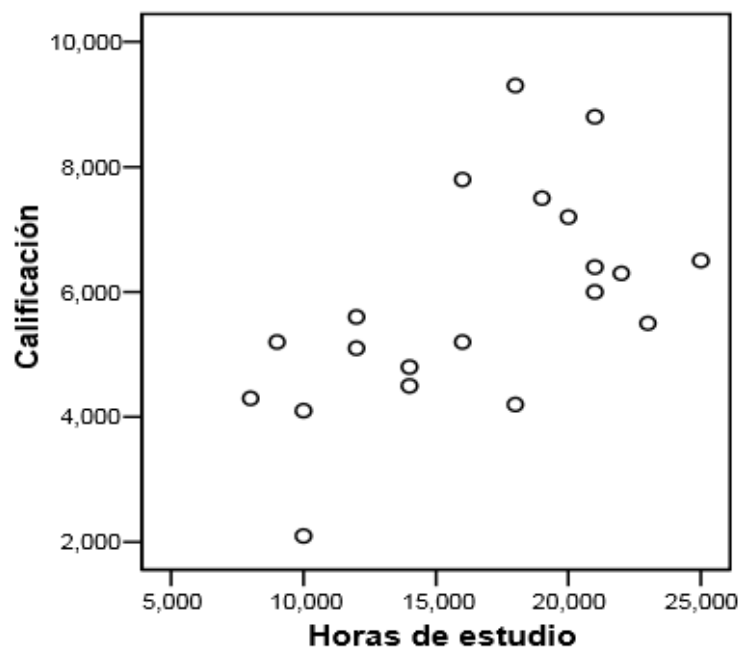
**¿Influye una variable de escala en los valores que toma otra variable de escala?**

Vamos a ver un diagrama de dispersión que muestra los valores de una variable en función de los que toma otra variable.

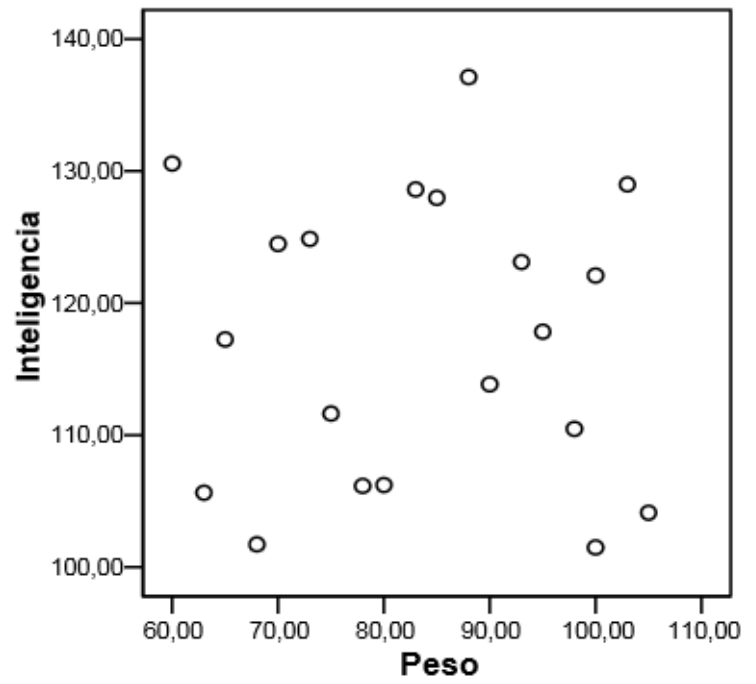
Tendencia inversa  
muy marcada



Tendencia positiva  
débil

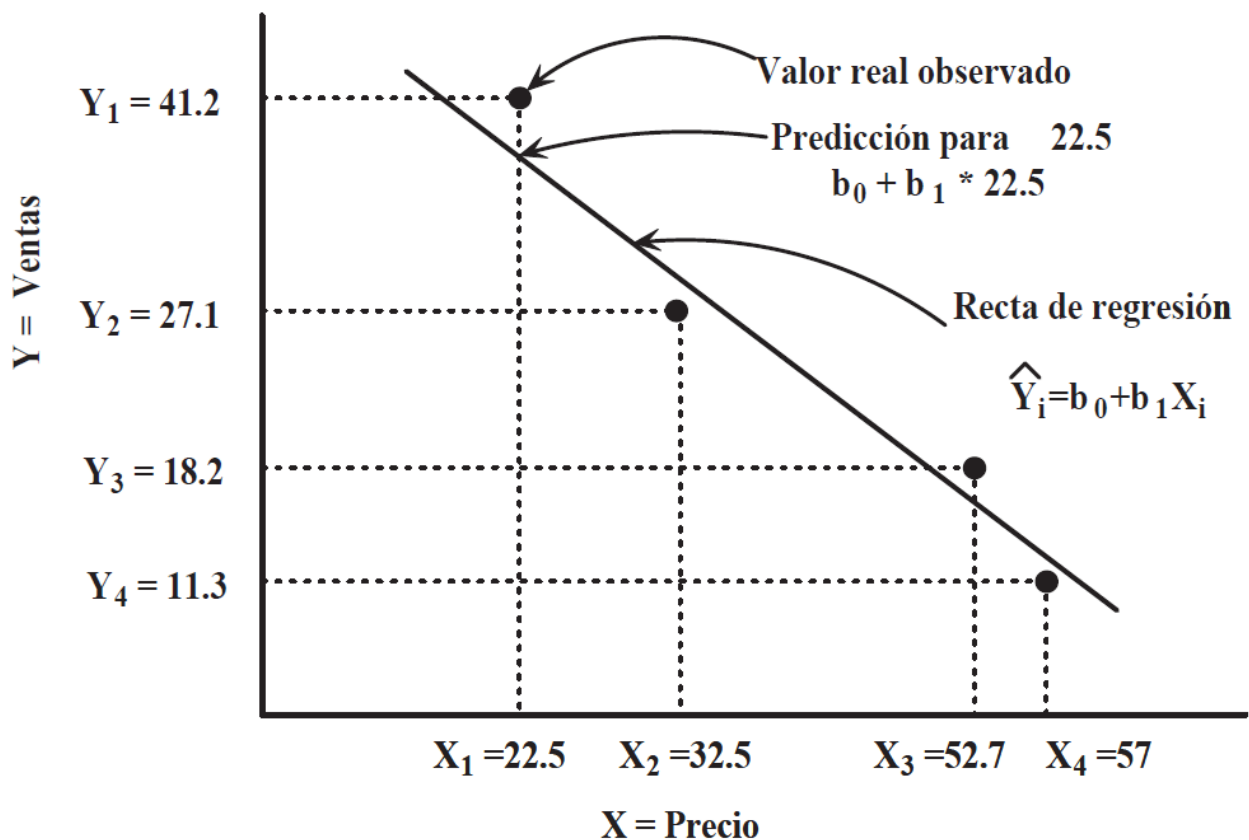


Ninguna relación



Podemos ajustar una recta de regresión a la nube de puntos.

$$\hat{Y}_i = b_0 + b_1 X_i$$



La recta se obtiene como aquella que minimiza las distancias de los puntos a la recta.

$$\min \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

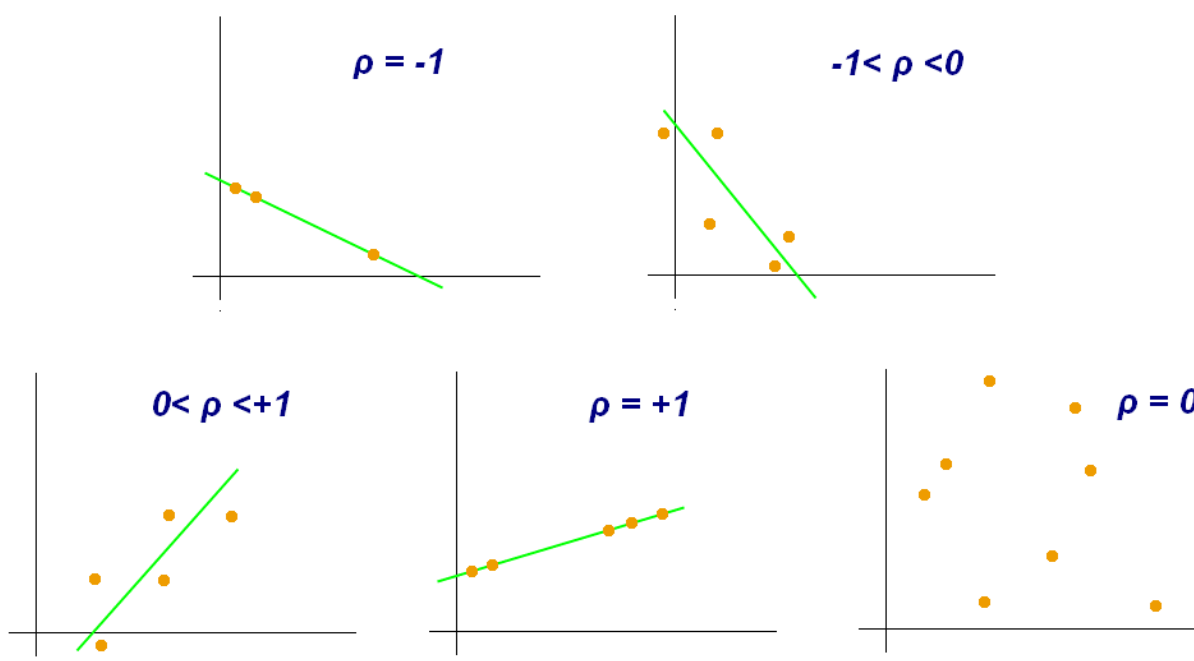
Es un problema de optimización con la siguiente solución:

$$b_1 = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

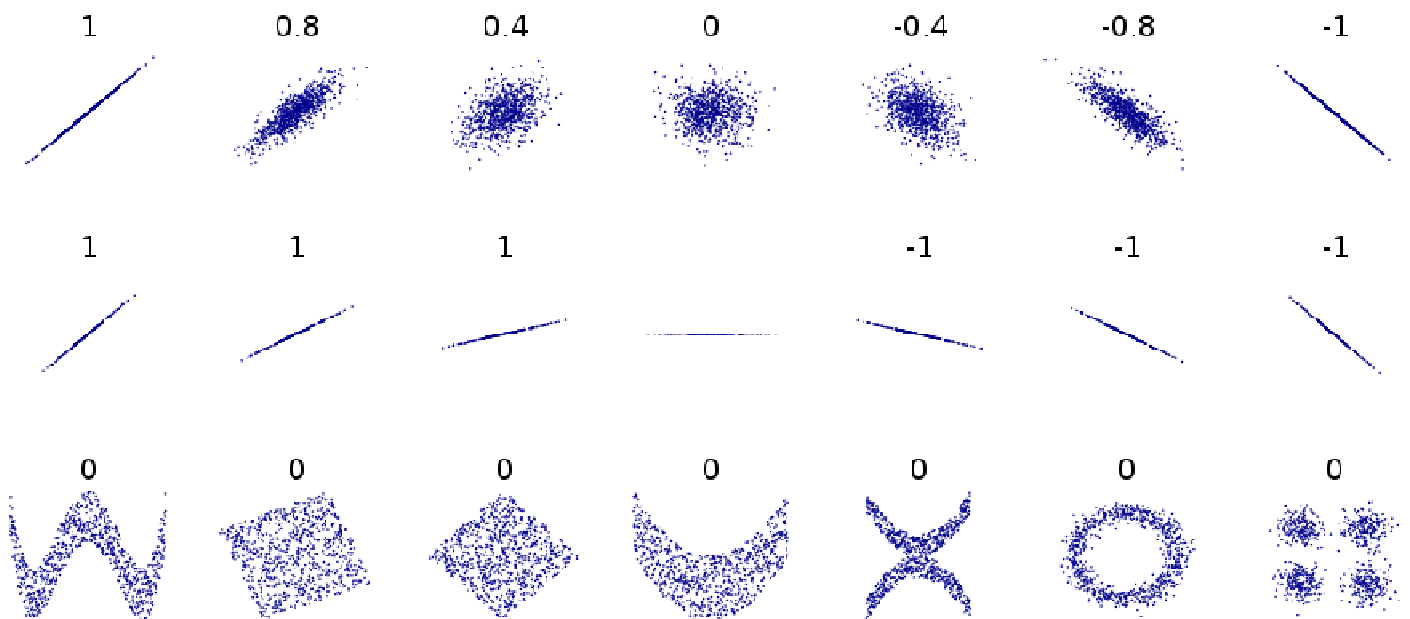
$$b_0 = \bar{Y} - b_1 \bar{X}$$

Cuando el tipo de relación estudiada es lineal, se puede ver el grado de ésta usando el **Coefficiente de Correlación de Pearson**:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$



¡Cuidado! Puede haber una marcada relación pero que no sea lineal.



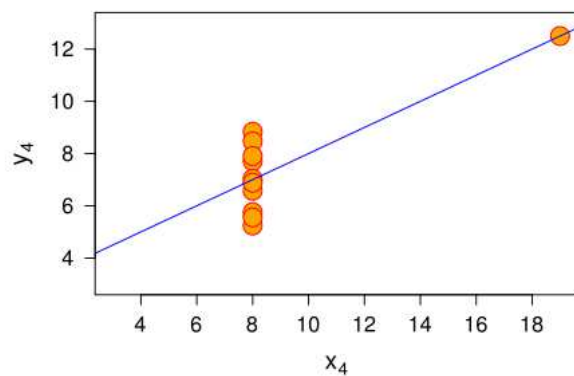
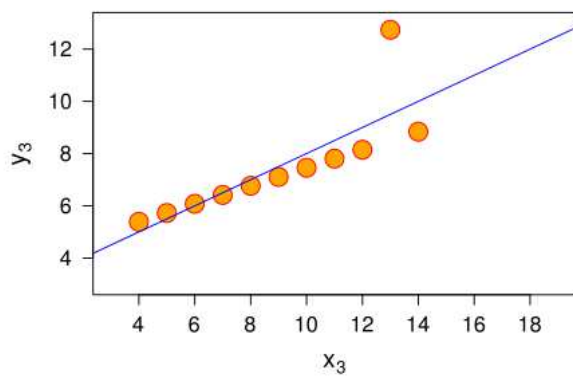
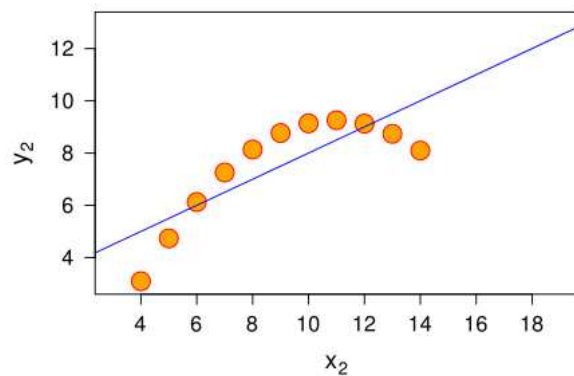
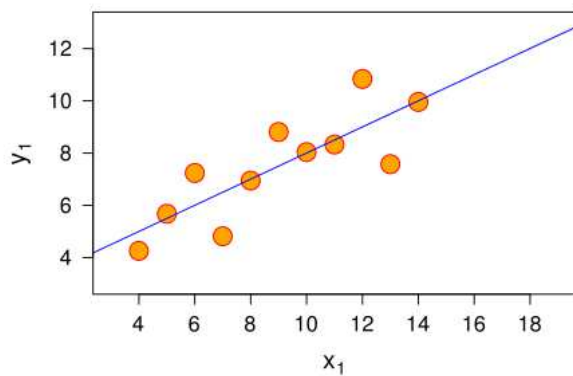
Correlación 0 → No hay relación lineal

Correlación 0 → Puede haber relación no lineal

Correlación alta (en valor absoluto) → Suele haber relación lineal, pero hay que corroborarlo viendo el diagrama de dispersión, tal y como muestra el cuarteto de Anscombe.

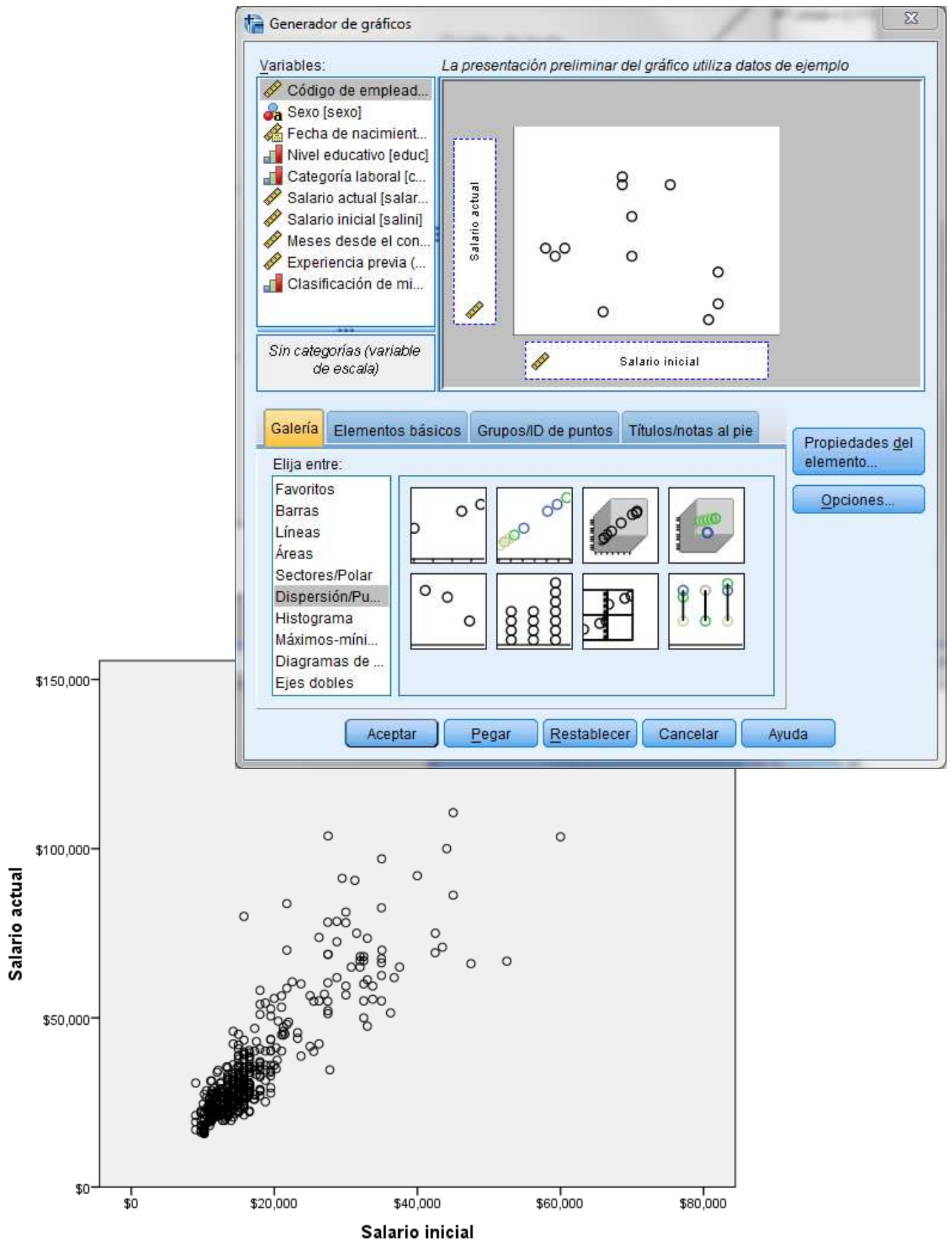
Propiedad	Valor
Media de cada una de las variables $x$	9.0
Varianza de cada una de las variables $x$	11.0
Media de cada una de las variables $y$	7.5
Varianza de cada una de las variables $y$	4.12
Correlación entre cada una de las variables $x$ e $y$	0.816
Recta de regresión	$y = 3 + 0.5x$

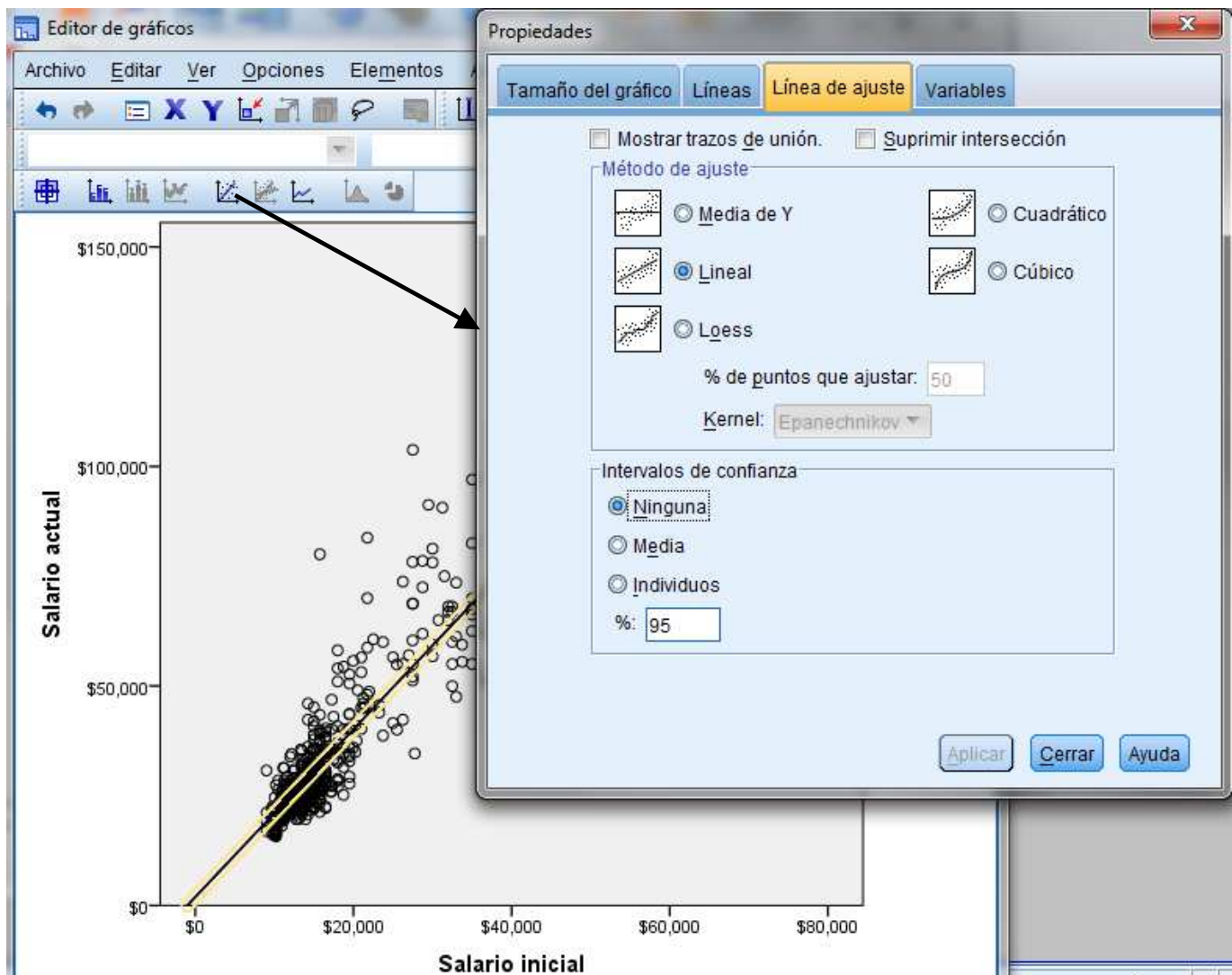
"Cuarteto de Anscombe"





⇒ Gráficos/Generador de Gráficos/Dispersión.





## 11 Análisis Estadísticos de dependencia.

### Numérica-Numérica. Regresión.

Con variables de escala, el estudio estadístico a realizar es el de regresión.

Cuando vimos el ANOVA, se comparaban las medias de cada uno de los grupos correspondientes a cada valor de la variable independiente. Cuando ésta es continua en vez de discreta, se procede a una descomposición similar de las sumas de cuadrados:

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$H_0. b_1 = 0$$

$$H_1. b_1 \text{ distinto de } 0$$

Tabla ANOVA				
Fuentes de Variación	Sumas de Cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	$SCR_{eg} = \sum (\hat{y}_i - \bar{y})^2$	1	$MCR_{eg}$	$\frac{MCR_{eg}}{MCE}$
Error	$SCE = \sum (y_i - \hat{y}_i)^2$	$n - 2$	$MCE = \frac{SCE}{n - 2}$	
Total	$SCT = \sum (y_i - \bar{y})^2$	$n - 1$	$\frac{SCT}{n - 1}$	

F se distribuye bajo la hipótesis nula según una F de Snedecor con n-2 grados de libertad.

Nota. El coef. de determinación  $R^2$  se define como la proporción de varianza total explicada por el modelo (en este caso de regresión)

$$R^2 = \frac{SCR_{eg}}{SCT} = 1 - \frac{SCE}{SCT} ; \quad 0 \leq R^2 \leq 1$$

En el caso de una variable independiente, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación de Pearson.

⇒ **Analizar/Regresión/lineal.** Hacedlo con las variables anteriores. En Estadísticos, seleccionad Estimaciones, Intervalos de Confianza y Ajuste del modelo. SPSS incluye un test de hipótesis (a través de estadístico F) de independencia.

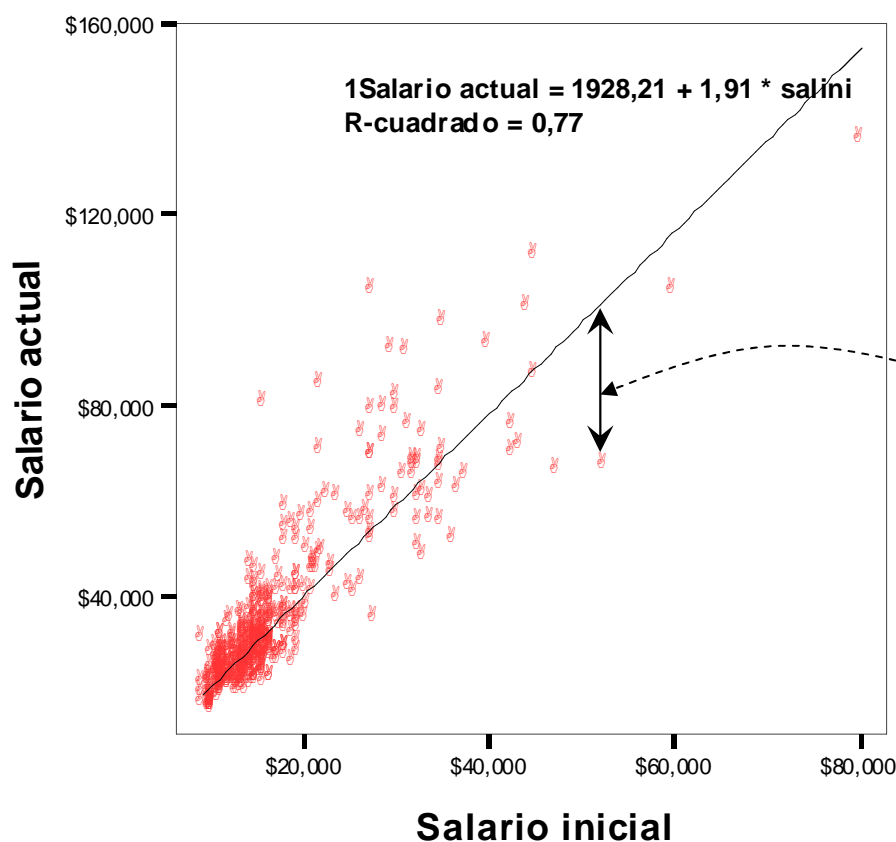
**ANOVA<sup>b</sup>**

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,068E+11	1	1,07E+11	1622,118	,000 <sup>a</sup>
	Residual	3,109E+10	472	65858997		
	Total	1,379E+11	473			

a. Variables predictoras: (Constante), Salario inicial

b. Variable dependiente: Salario actual

Se rechaza independencia, por lo que aceptamos que el Sal Inic influye en el Sal Act



Regresión lineal

El análisis de SPSS también muestra un ajuste de los parámetros:

Coeficientes <sup>a</sup>								
		Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	1928,206	888,680		2,170	,031	181,947	3674,464
	Salario inicial	1,909	,047	,880	40,276	,000	1,816	2,003

a. Variable dependiente: Salario actual

Valor poco significativo (por eso, el IC es muy grande) No debemos fiarnos mucho del valor de la cte (1928.206)

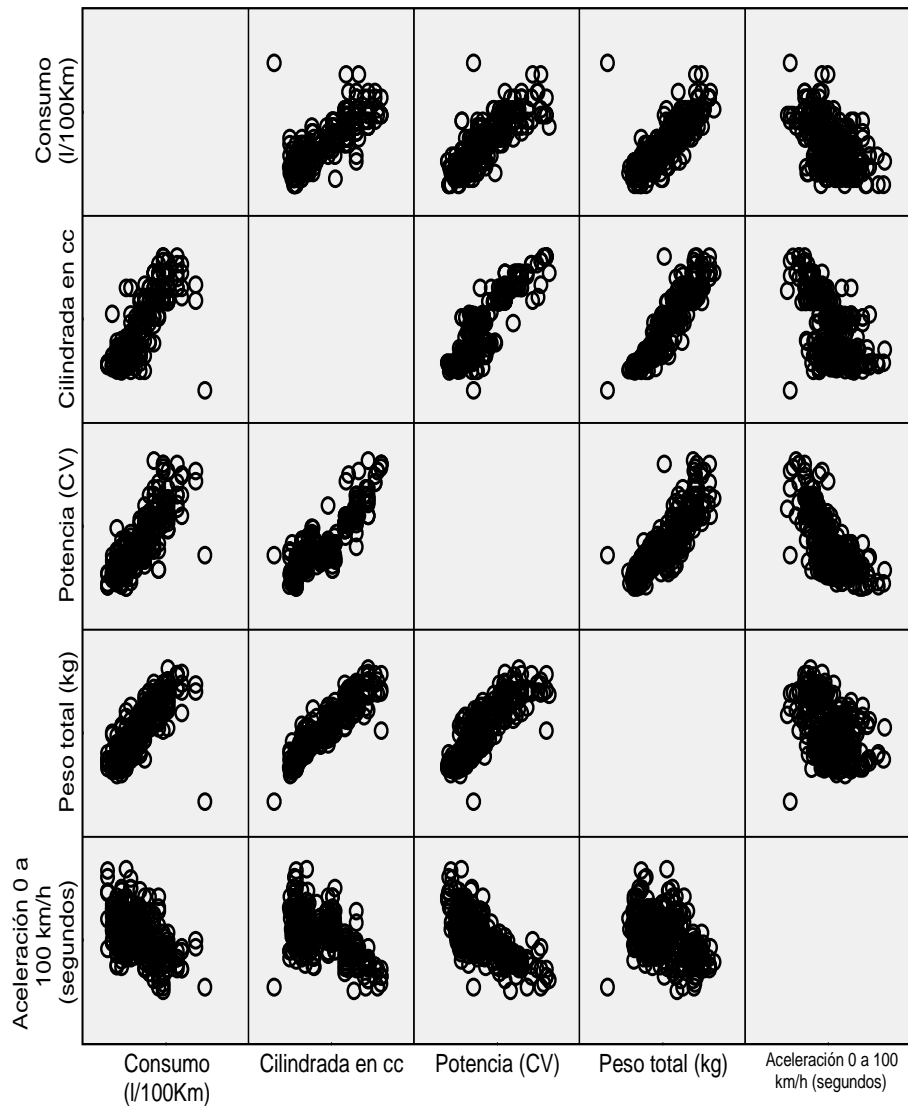
Valor muy significativo. La pendiente de la recta (1.909) es bastante fiable. El IC es pequeño.

**Ejercicio:** ¿Depende el salario inicial de la experiencia previa?

Es posible obtener un diagrama de dispersión de varias variables, cruzadas dos a dos.

Cargad la bd de coches

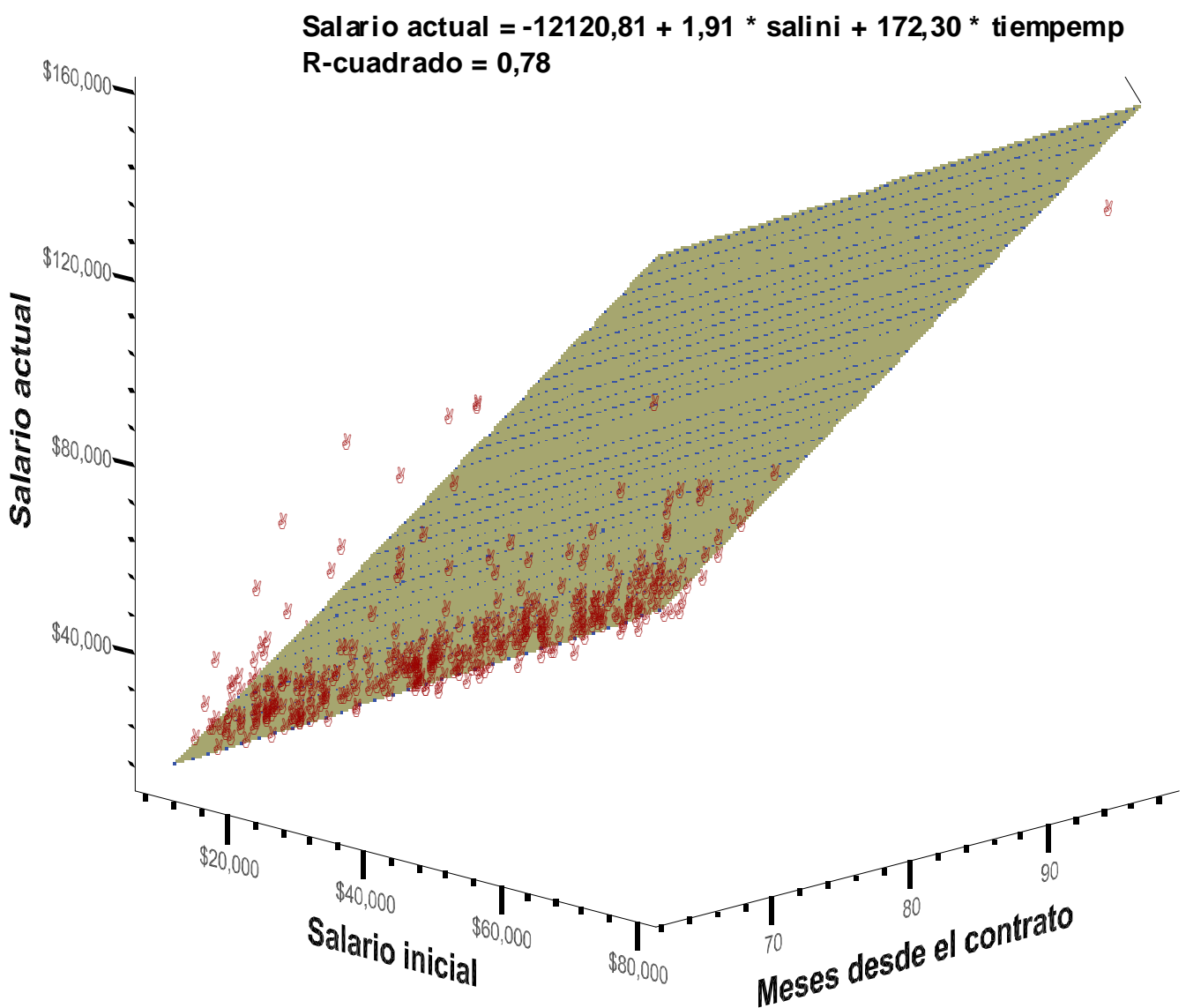
⇒ **Gráficos/Dispersión/Dispersión Matricial.** Seleccionad el consumo, cilindrada, potencia, peso, aceleración.



¿Qué pasa cuando tenemos varias variables independientes? En este caso, se busca una combinación lineal de las variables independientes para explicar la dependiente. Es importante destacar que el objetivo es encontrar una fórmula lineal. Es decir, que para cualquier conjunto de valores de las variables independientes, podemos aplicar siempre una misma fórmula lineal que pronostique el valor de la dependiente. En la práctica, no suele ocurrir que existan dependencias lineales.

⇒ ¿Depende el salario actual del salario inicial y de los meses del contrato?

Podemos plantear un AED, seleccionando Gráficos/Interactivos/Dispersión/Coordenadas 3D



Realizamos la regresión y obtenemos:

**Coeficientes<sup>a</sup>**

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-12120,8	3082,981		-3,932	,000
	Salario inicial	1,914	,046	,882	41,271	,000
	Meses desde el contrato	172,297	36,276	,102	4,750	,000

a. Variable dependiente: Salario actual

Todos los coeficientes son significativos

**ANOVA<sup>b</sup>**

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,083E+11	2	5,41E+10	859,383	,000 <sup>a</sup>
	Residual	2,966E+10	471	62982310		
	Total	1,379E+11	473			

a. Variables predictoras: (Constante), Meses desde el contrato, Salario inicial

b. Variable dependiente: Salario actual

**Ejercicio:** Sobre la base de datos de las encuestas, ¿depende el número de años de escolarización del encuestado del número de años de escolarización del padre y la madre?

Nota: En general, la regresión sólo se usa con variables de escala, pero podemos hacer un pequeño truco para ver como influye una variable nominal (con pocos valores, usualmente sólo 2) en la regresión. En el ejemplo anterior, podemos transformar la variable "Sexo" en una numérica con dos valores: 0 (hombre) y 1 (mujer). Si la incluimos como variable independiente, el coeficiente de la recta de regresión nos va a indicar la diferencia porcentual del salario en función del sexo.



## 12 AED: Informes y gráficos sobre varias variables

### Numérica-Nominal

Queremos ver la relación que existe entre variables cuando una de ellas es numérica y la otra nominal. En estos estudios la variable numérica es la variable *dependiente* mientras que la(s) otra(s) es(son) la(s) independiente(s) o *factores*. Un primer tipo de preguntas en las que podríamos estar interesados sería:

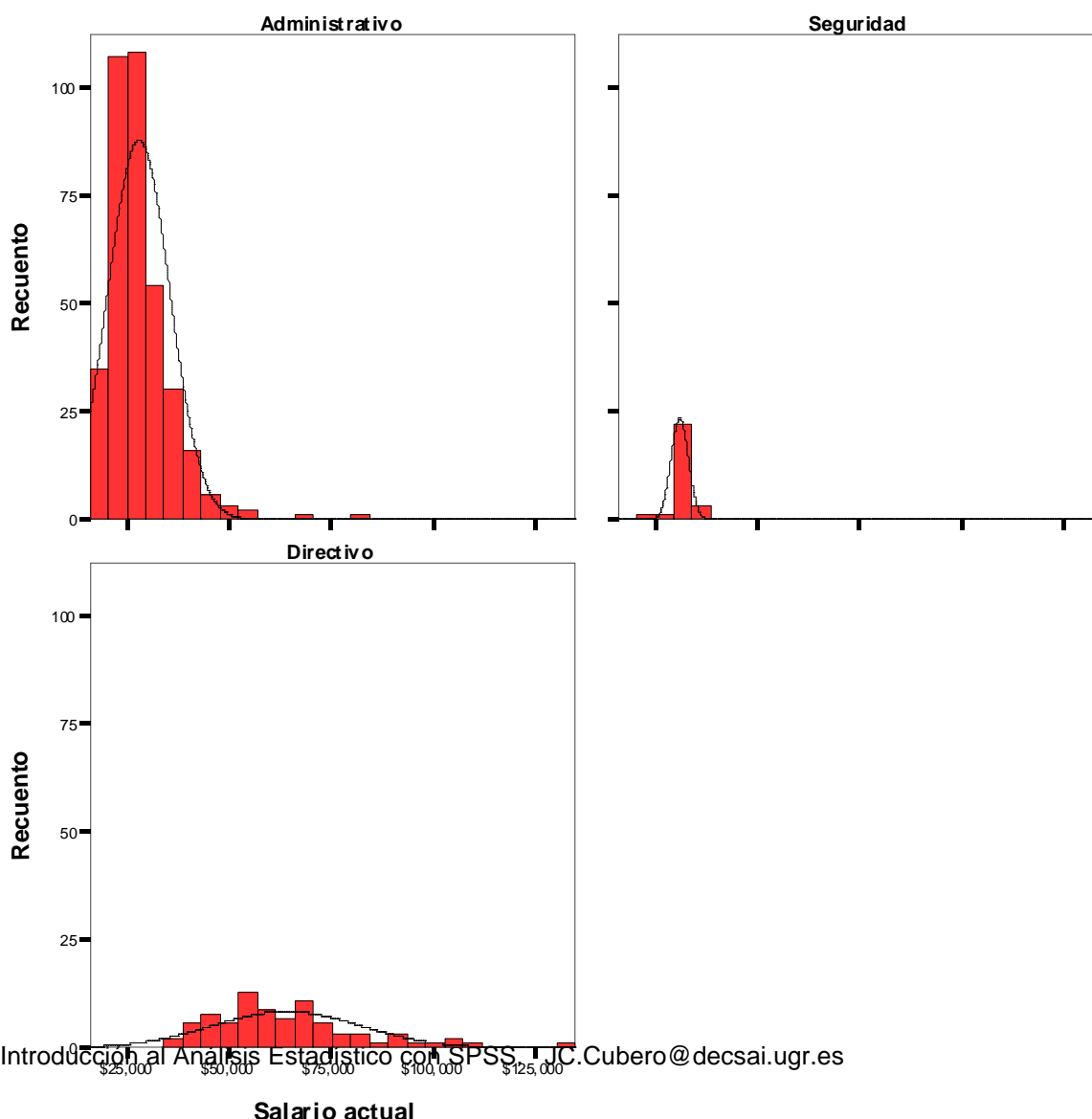
**¿Cómo se distribuye el salario de los empleados en función de su categoría laboral?**

En general:

**¿Cómo se distribuye una variable cuantitativa para cada valor posible de otra variable nominal (factor)?**

Como la variable dependiente (consecuente) es numérica, usamos un histograma.

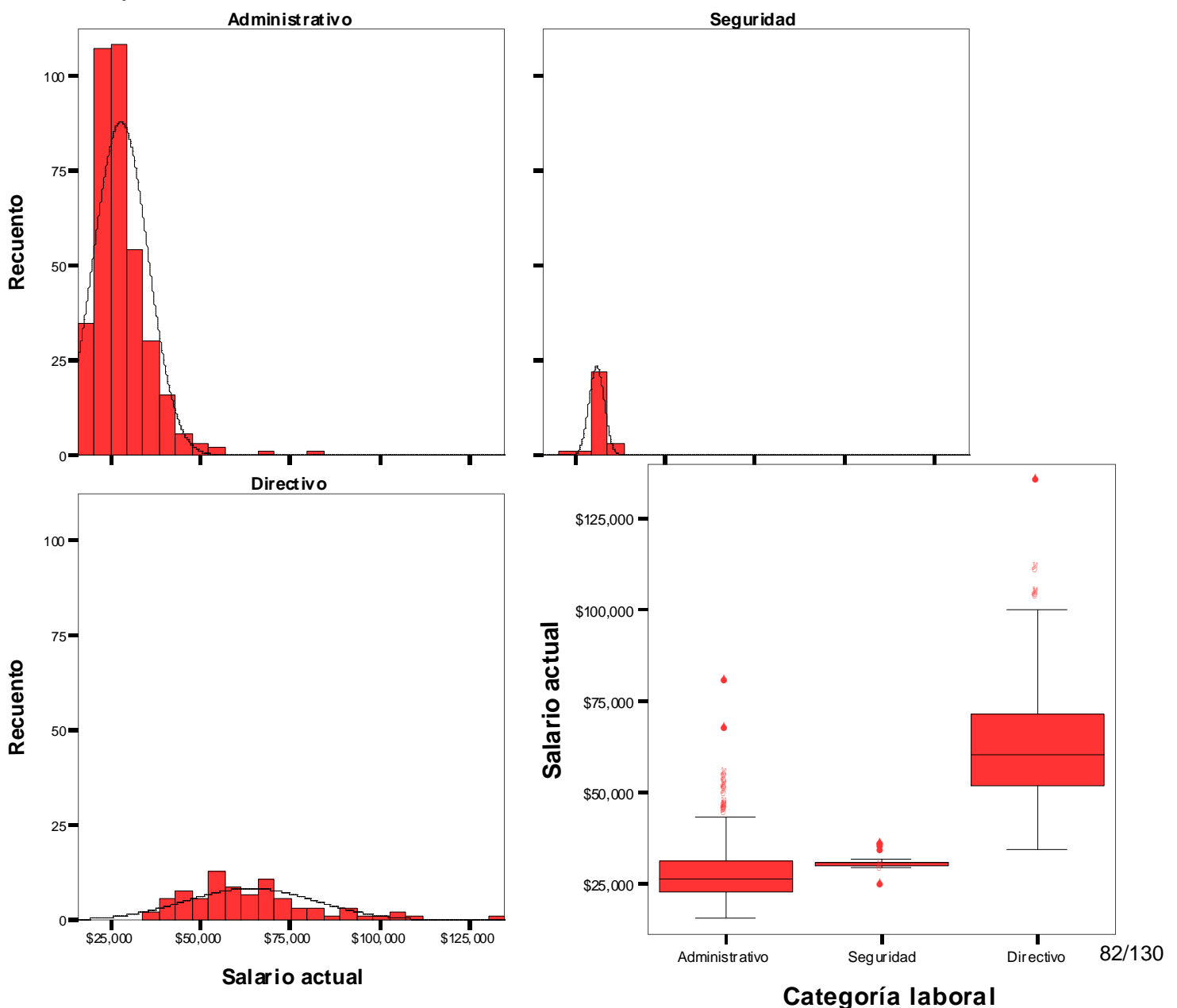
⇒ Gráficos/Cuadros.. antiguos/Histogramas/ Seleccionamos Sal Actual y agrupamos por categoría laboral (No disponible en PSPP)



Observamos que hay muchos menos directivos que administrativos (barras de conteo muy altas en los administrativos), pero que su salario es mayor (dichas barras están más a la derecha del eje de las abscisas que representa el salario). Si nos fijamos en el eje de abscisas que representa el salario, observamos que hay directivos con sueldos altos y otros bajos (aunque la distribución es muy parecida a una normal y por lo tanto los salarios están concentrados en una zona central), mientras que los administrativos y personal de seguridad tienen menos variabilidad en los salarios. Una forma de ver la dispersión de los salarios es con diagramas de cajas:

⇒ Gráficos/Cuadros.. antiguos/Diagramas de Cajas/ Seleccionamos “Salario Actual”

arriba (variable dependiente) “ y Categoría laboral” abajo (variable independiente) Podemos ver las diferencias existentes, en cuanto al salario, entre las distintas categorías laborales. Aquí ya no representamos número de datos (como en el histograma) sino que nos estamos fijando en la distribución relativa.



Ya que hemos echado un vistazo a la distribución del Salario en función de la categoría laboral, podríamos estar interesados en responder a preguntas del tipo:

**¿Cual es la media aritmética del salario de los empleados que son administrativos?**

**¿Cual es la media aritmética del salario de los empleados varones que son administrativos?**

⇒ **Analizar/Informes/Resúmenes de casos.** Quitad "Mostrar los casos". Seleccionad la variable "Salario Actual" y agrupad resultados por "Categoría laboral". Includ como estadísticos la media y la desviación típica.

#### Resúmenes de casos

Salario actual

Categoría laboral	N	Media	Desv. típ.
Administrativo	363	\$27,838.54	\$7,567.995
Seguridad	27	\$30,938.89	\$2,114.616
Directivo	84	\$63,977.80	\$18,244.776
Total	474	\$34,419.57	\$17,075.661

⇒ Haced lo mismo agrupando también por Sexo.

#### Resúmenes de casos

Salario actual

Categoría laboral	Sexo	N	Media	Desv. típ.
Administrativo	Hombre	157	\$31,558.15	\$7,997.978
	Mujer	206	\$25,003.69	\$5,812.838
	Total	363	\$27,838.54	\$7,567.995
Seguridad	Hombre	27	\$30,938.89	\$2,114.616
	Total	27	\$30,938.89	\$2,114.616
Directivo	Hombre	74	\$66,243.24	\$18,051.570
	Mujer	10	\$47,213.50	\$8,501.253
	Total	84	\$63,977.80	\$18,244.776
Total	Hombre	258	\$41,441.78	\$19,499.214
	Mujer	216	\$26,031.92	\$7,558.021
	Total	474	\$34,419.57	\$17,075.661

⇒ **Opcional: Analizar/Informes/Cubos OLAP.** Lo mismo que antes, pero se muestra un gráfico interactivo, dónde vamos seleccionando la categoría y el sexo. Podemos tener en cuenta relaciones parciales (por ejemplo Sexo: Hombre y Cat Lab: total o bien Sexo: Mujer y Cat Lab: Directivo)

Viendo los estadísticos y los gráficos, podemos comprobar, por ejemplo, que la media aritmética de los directivos mujeres es inferior a la media aritmética de los directivos hombres. Así que un tipo de pregunta en la que podríamos estar interesados sería:

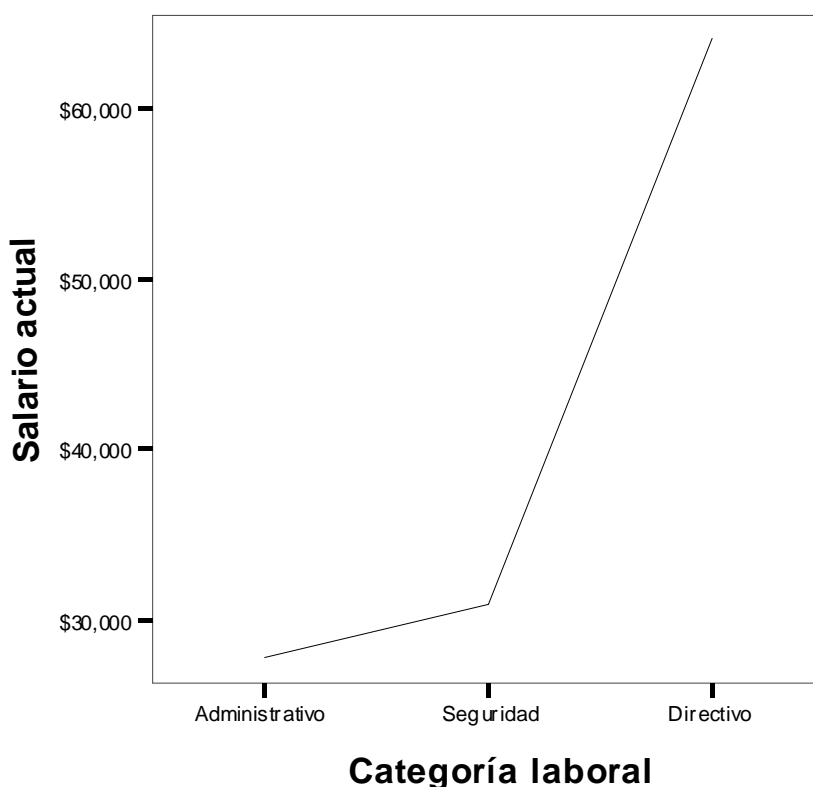
**¿Depende el salario de los empleados de su categoría laboral?**

es decir,

**¿Varía el salario de los empleados en función de la categoría laboral?**

Para ello, vamos a ver hasta qué punto las medias de los salarios son distintas en cada categoría laboral. Una forma gráfica cómoda de comparar las medias (en vez de recurrir a las tablas numéricas anteriores) es:

⇒ **Gráficos/Cuadros.. antiguos/Líneas/** Sustituid Recuento por Salario Actual y que las líneas representen medias. Seleccionad Categoría laboral en el antecedente.



Los puntos/líneas muestran Medias

Si queremos controlar más variables, basta incluirlas en variables de leyenda, color, etc.

⇒ **Opcional:** Includ la variable “Sexo” como variable de leyenda (para que aparezcan juntas las cajas) o como variable de panel (para que aparezcan separadas las cajas)

## 13 Análisis Estadísticos de dependencia.

### Numérica-Nominal

#### 13.1 Prueba para múltiples muestras: ANOVA

Supongamos que queremos ver si las medias de varias muestras, no necesariamente del mismo tamaño, son las mismas. Por ejemplo, queremos ver cómo influyen tres dietas distintas. Partimos de un conjunto de individuos, los asignamos de forma aleatoria a cada dieta y vemos si el incremento de peso es el mismo en cada uno de los tres grupos (dietas)

O supongamos que tenemos una variable nominal (factor) y agrupamos una variable de escala según los grupos determinados por el factor. Queremos ver si la media de cada grupo es la misma. Por ejemplo, queremos ver si el Salario medio es el mismo en cada uno de los grupos formados por la Categoría Laboral.

Podríamos estar tentados en plantear todos los test de hipótesis (T-tests) resultado de todas las posibles combinaciones:

$H_{0ij}$ . La media del grupo  $i$  = media del grupo  $j$

$H_{1ij}$ . Son distintas

Esto no es viable por lo siguiente. Supongamos que hay  $k$  grupos. Por lo tanto, tendríamos que construir todas las posibles combinaciones (uno con dos, dos con tres, uno con tres) lo que nos llevaría a plantear  $m = k(k-1)/2$  tests de hipótesis. Supongamos un nivel de significación  $\alpha$  (0.05 por ejemplo). La probabilidad de no cometer un error de tipo I en un test es 0.95. La probabilidad de no cometer un error de tipo I en tres test ( $m=3$ ) es  $0.95^3 = 0.86$ . Por lo tanto, la probabilidad de cometer al menos un error de tipo I es  $1-0.86 = 0.14$ . En general:

$$\text{Prob(Cometer al menos un error de tipo I)} = 1 - (1-\alpha)^m$$

Este error se conoce como **Family Wise Experiment Error (FWER)**

Para  $\alpha = 0.05$  y  $m=10$  por ejemplo, dicha probabilidad es 0.4, un valor extremadamente alto.

**El razonamiento anterior se aplica en cualquier situación en la que tenemos que contrastar de forma simultánea  $m$  tests de hipótesis.**

Por lo tanto, se recurre al siguiente test:

$H_0$ . **Todas** las medias (de la variable dependiente) de todos los grupos son iguales

$H_1$ . Hay una media (de la variable dependiente) de algún grupo que es distinta a la de otro grupo

Este tipo de test se conoce como "**Multisample Hypotheses**"

La idea en la construcción del test es la siguiente: El valor  $i$ -ésimo de la variable dependiente, en un grupo  $g$  lo denotamos por  $x_{gi}$ , por ejemplo  $x_{Directivo,34} = \$34.150$

Cada valor se puede descomponer de la siguiente forma:

$$(x_{gi} - \bar{x}) = (\bar{x}_g - \bar{x}) + (x_{gi} - \bar{x}_g)$$

donde  $\bar{x}_g$  representa la media del grupo correspondiente (por ejemplo, si  $g = \text{Directivos}$ , representaría la media en el salario de los Directivos)

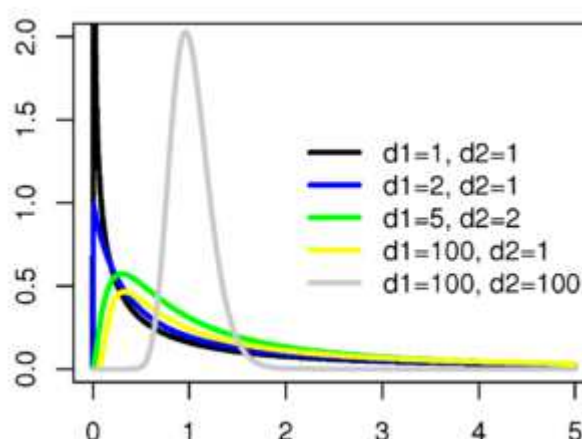
Es decir, estamos diciendo que la desviación de cada valor con respecto a la media global, se puede poner como la suma de la desviación con respecto a la media de su grupo, más la desviación de la media del grupo con respecto a la media global.

Para tener en cuenta todos los valores y para que no se compensen negativos con positivos, podemos realizar la suma al cuadrado de todas esas cantidades y obtendríamos:

$$\sum_g \sum_i (x_{gi} - \bar{x})^2 = \sum_g n_g (\bar{x}_g - \bar{x})^2 + \sum_g \sum_i (x_{gi} - \bar{x}_g)^2$$
$$SS_T = SS_B + SS_W$$

Dónde  $B$  representa "Between groups" (inter-grupos), aunque estaría mejor dicho "Among groups", es decir, la suma de las desviaciones las medias de cada grupo con respecto a la media global, y  $W$  representa "Within groups" (intra-grupos), es decir, la suma de las desviaciones de cada valor a la media de su grupo. Ponderando por los correspondientes grados de libertad ( $df_B = \text{número de grupos} - 1$  y  $df_W = \text{número de valores} - \text{número de grupos}$ ) se construye el estadístico  $F$ :

$$F = \frac{SS_B / df_B}{SS_W / df_W} = \frac{GMS}{EMS}$$



GMS = Groups Mean Square

EMS = Error Mean Square

Dentro de cada grupo, los individuos presentan una variabilidad dada por el denominador de la anterior fórmula. Si las medias fuesen iguales en todos los grupos, la variabilidad medida por el numerador no aportaría nada nuevo, por lo que el cociente  $F$  sería próximo a 1. En caso contrario (medias distintas), el estadístico  $F$  sería mayor que 1.

Se puede demostrar que cuando la hipótesis nula de igualdad de medias es cierta, y si la distribución de cada grupo o factor es una normal, entonces el estadístico  $F$  sigue una distribución denominada  $F$  de Snedecor. Como siempre, bastará con comprobar los valores resultantes del estadístico con los de la tabla de la distribución, a un  $p$ -nivel fijado a priori.

Los parámetros de la distribución  $F$  son  $df_B$  y  $df_W$

El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos deben ser simétricos, unimodales y con la misma varianza. Si no se satisface alguna de estas restricciones o si el tamaño de la muestra es pequeño, se recurren a los **tests no paramétricos** (ver el último apartado)

⇒ **Analizar/Comparar Medias/Anova de un Factor**. Podemos incluir en Opciones, el gráfico de líneas de las medias que habíamos generado anteriormente.

Efectivamente, como suponíamos, podemos afirmar que hay alguna categoría laboral que tiene un salario actual medio distinto a la de otra categoría laboral

### ANOVA

Salario actual

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	8,944E+10	2	4,47E+10	434,481	,000
Intra-grupos	4,848E+10	471	1,03E+08		
Total	1,379E+11	473			

**Ejercicio:** ¿Hay diferencia de salarios atendiendo al sexo?



## 13.2 Post-hoc

Supongamos que rechazamos la hipótesis nula de un ANOVA. ¿Qué podemos deducir? Que hay alguna(s) media(s) distinta(s). ¿Pero cuál? Podrían ser iguales las medias de dos grupos pero distintas a la de un tercer grupo, o podrían ser distintas las tres.

Después de aplicar el ANOVA, si se rechaza la hipótesis nula, se procede a un análisis "post-hoc" o de comparaciones múltiples para ver qué medias son distintas entre sí.

¿Cómo hacemos el post-hoc?

Se procede a plantear todos los test de hipótesis correspondientes a los posibles cruces. Por ejemplo, si tenemos 4 grupos, 1 con 2, 1 con 3, 1 con 4, 2 con 3, 2 con 4, 3 con 4. Ya vimos en la introducción del ANOVA que si planteamos todos los test de hipótesis resultantes en los  $k$  grupos, vamos a tener  $m = k(k-1)/2$  comparaciones distintas.

$H_{0\ ij}$ . La media del grupo  $i$  = media del grupo  $j$

$H_{1\ ij}$ . Son distintas

Y el error FWER se dispara. Recordad que

$$\text{FWER} = \text{Prob}(\text{Cometer al menos un error de tipo I}) = 1 - (1-\alpha)^m$$

Para resolver este problema, se procede a ejecutar dichos tests pero ajustando el nivel de significación. Para ello, se recurre a dos tipos de soluciones:

- Single test procedures (STPs).
  - O bien Simplemente ajustando (Bonferroni, Sidak, etc) el nivel de significación y haciéndolo más pequeño para controlar el error FWER
  - Desarrollando tests específicos como el test de Tukey.
- Multiple-Stage tests (MSTs). Al igual que los STPs, se establecen tantos tests de hipótesis como comparaciones haya, pero se ajusta el nivel de significación para cada una de las comparaciones. Si una comparación no es significativa se deja de comparar el resto.

Una vez aplicado el post-hoc se nos pueden dar varias situaciones:

$$\begin{array}{cccc} \overline{X}_1 & \overline{X}_2 & \overline{X}_3 & \overline{X}_4 \\ \hline & & & \end{array}$$

1 y 2 similares. 3 y 4 similares. 1 y 2 diferentes de 3 y 4.

$$\begin{array}{cccc} \overline{X}_1 & \overline{X}_2 & \overline{X}_3 & \overline{X}_4 \\ \hline & & & \end{array}$$

1 y 2 similares. 2, 3 y 4 similares. 1 diferente de 3 y 4.

No es posible determinar a qué población pertenece el segundo grupo, por lo que se ha cometido al menos un error de tipo II en el post-hoc. Esto es posible ya que los tests post-hoc tienen menos potencia que el ANOVA original.

Incluso podría llegarse al caso:

$$\begin{array}{cccc} \overline{X}_1 & \overline{X}_2 & \overline{X}_3 & \overline{X}_4 \\ \hline & & & \end{array}$$

cuando el ANOVA habría dado significativo en rechazar igualdad.

## STP: Test de Tukey.

Se aplica en el caso concreto de estar comparando las medias de varios grupos que son independientes, normalmente distribuidos y homocedásticos.

$H_{0\ ij}$ . La media del grupo  $i$  es la misma que la del grupo  $j$

$H_{1\ ij}$ . Son distintas

La idea consiste en calcular la diferencia entre las medias de cada par de grupos

$$q_{ij} = \frac{\bar{X}_i - \bar{X}_j}{EMS\sqrt{2/n}}$$

y comparar todos los valores obtenidos con una misma distribución, a saber, la distribución del estadístico construido a partir de la mayor diferencia observada entre dos grupos:

$$q = \frac{\bar{X}_{\max} - \bar{X}_{\min}}{EMS\sqrt{2/n}}$$

dónde  $EMS^2$  (mean square error) es el denominador del estadístico  $F$ .

Tukey demostró que el estadístico  $q$  sigue una distribución Q "**rango estudentizado**" (studentized range) que dependen de dos parámetros:  $df$  (grados de libertad) y  $k$ . Si la hipótesis nula es cierta,  $q$  sigue una Q con parámetros  $N-k$  y  $k$  ( $N$  = Número total de datos,  $k$ =número de grupos)

Como siempre, si el valor observado  $q_{ij}$  (ahora tenemos tantos valores como pares de medias estemos comparando) es muy poco frecuente en dicha distribución, se rechaza la hipótesis nula. Recordemos que el nivel de significación  $\alpha$  representa en este caso la probabilidad de cometer al menos un error tipo I, es decir, la probabilidad de rechazar erróneamente al menos una de las hipótesis nula. No es el error cometido en una comparación aislada (tasa de error por comparación, Comparison wise error rate), sino el Family Wise Experiment Error FWER (aunque Tukey lo denominó inicialmente Experiment wise error rate)

En definitiva, se usa la **misma** distribución y el **mismo** nivel de significación para todos los estadísticos resultantes de todas las comparaciones.

## STP: Ajuste en un sólo paso el nivel de significación.

Se parte de un conjunto de tests de hipótesis que se realizan simultáneamente sobre un mismo experimento. Cada test saldrá con un p-value diferente. Se ajusta el nivel de significación y se comparan todos los tests con el valor ajustado.

El ajuste más conocido es el método de **Bonferroni** simplemente divide el nivel de significación  $\alpha$  por el **número de comparaciones ( $m$ )**.

$$\hat{\alpha} = \frac{\alpha}{m}$$

Cada test será significativo si el p-value correspondiente es menor que  $\hat{\alpha}$

### Example: Bonferroni

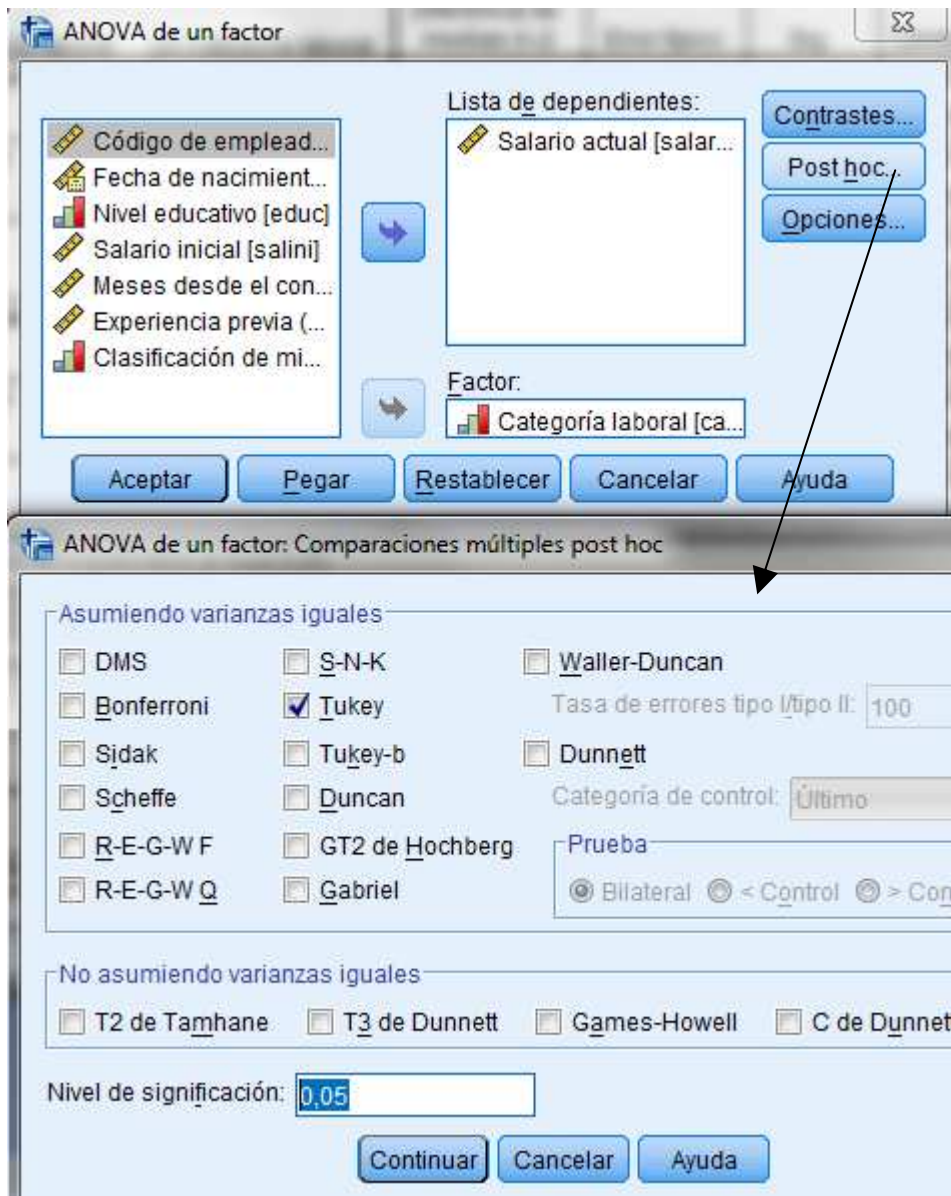
- P-values (sorted):  
 $H_{0(1)}$ : 0.005,  $H_{0(2)}$ : 0.011,  $H_{0(3)}$ : 0.02,  $H_{0(4)}$ : 0.04,  $H_{0(5)}$ : 0.13
- $M = 5$  tests; Significance level: 0.05
- Corrected p-value:  $0.005 \cdot 5 = 0.025 < 0.05$ : Reject  $H_{0(1)}$
- Corrected p-value:  $0.011 \cdot 5 = 0.055$ : Don't reject  $H_{0(2)}$
- Corrected p-value:  $0.02 \cdot 5 = 0.1$ : Don't reject  $H_{0(3)}$
- Corrected p-value:  $0.04 \cdot 5 = 0.2$ : Don't reject  $H_{0(4)}$
- Corrected p-value:  $0.13 \cdot 5 = 0.65$ : Don't reject  $H_{0(5)}$
  
- Conclusion:  
Reject  $H_{0(1)}$ , don't reject  $H_{0(2)}$ ,  $H_{0(3)}$ ,  $H_{0(4)}$ ,  $H_{0(5)}$

Si el número de comparaciones es alto, no es adecuado al tener poca potencia.

En el caso particular de ANOVAs en los que se comparan  $k$  medias, hay un total de  $m = k(k-1)/2$  comparaciones (valor que puede dispararse)

⇒ Analizar/Comparar Medias/Anova de un Factor.

**Post Hoc** seleccionamos el método de Tukey.



## Comparaciones múltiples

Variable dependiente: Salario actual

HSD de Tukey

(I) Categoría laboral	(J) Categoría laboral	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Administrativo	Seguridad	-\$3,100.35	\$2,023.760	,277	-\$7,858.50	\$1,657.80
	Directivo	-\$36,139.26*	\$1,228.352	,000	-\$39,027.29	-\$33,251.22
Seguridad	Administrativo	\$3,100.35	\$2,023.760	,277	-\$1,657.80	\$7,858.50
	Directivo	-\$33,038.91*	\$2,244.409	,000	-\$38,315.84	-\$27,761.98
Directivo	Administrativo	\$36,139.26*	\$1,228.352	,000	\$33,251.22	\$39,027.29
	Seguridad	\$33,038.91*	\$2,244.409	,000	\$27,761.98	\$38,315.84

\*. La diferencia entre las medias es significativa al nivel .05.

## Salario actual

HSD de Tukey<sup>a,b</sup>

Categoría laboral	N	Subconjunto para alfa = .05	
		1	2
Administrativo	363	\$27,838.54	
Seguridad	27	\$30,938.89	
Directivo	84		\$63,977.80
Sig.		,227	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

- Usa el tamaño muestral de la media armónica = 58,031.
- Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

Nivel de significación en el test de disparidad dentro de cada subconjunto.

Subconjunto 1 (Grupos Adm y Seg): Sig = .227, por lo que no pueden considerarse distintos Adm/Seg.

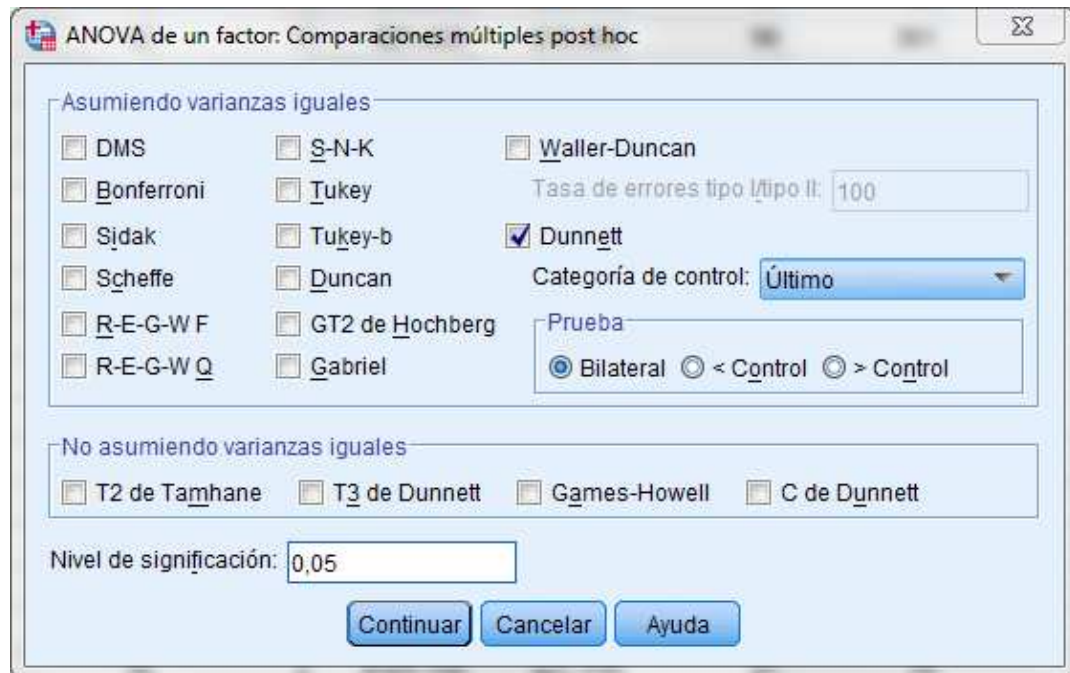
Subconjunto 2 (Grupo Dir): Al sólo haber 1 valor (Directivo), no hay disparidad DENTRO, por lo que Sig=1

Directivo      Seguridad      Administrativo

---

**IMPORTANTE:** Cuanto menor sea el número de comparaciones a realizar, mayor será la potencia del test. Por lo tanto, si estamos interesados en comparar el valor de una variable de **CONTROL** con el resto de variables (k-1) y no todos los posibles cruces de todos con todos, solamente necesitaríamos  $m=(k-1)$  comparaciones, por lo que aumentaría la potencia del test.

En SPSS no podemos seleccionar el control en todos los métodos. Únicamente con el de **Dunnett** (que es del tipo STP):



**Nota:** Si seleccionamos Comparar Medias / Medias e incluimos varias variables independientes, se hace un estudio de un ANOVA de la v. dependiente cruzada de forma independiente con cada una de las independientes. Si queremos fijar valores concretos de la v. independiente podemos seleccionar Comparar Medias / Muestras Independientes.

## Multiple-Stage tests (MSTs) -step wise-

Se parte de un conjunto de tests de hipótesis que se realizan simultáneamente sobre un mismo experimento. Cada test saldrá con un p-value diferente. Se ajusta el nivel de significación **de forma distinta para cada test** y se compara cada p-value con el correspondiente nivel de significación. Las comparaciones se ordenan (de mayor a menor valor o al revés) Si una comparación no es significativa se deja de comparar el resto (de ahí el nombre de step wise)

En definitiva, los MST funcionan secuencialmente, proporcionando un nivel de significación diferente para cada test. Los MST tienen más potencia que los STP.

Uno de los métodos más conocidos es el de **Holm** (Holm-Bonferroni): Se ordenan los p-values de los distintos test de hipótesis, de menor a mayor (de más significativo a menos significativo) obteniendo  $\{p_1, \dots, p_m\}$  Se escoge el primero de ellos ( $p_1$ , el más pequeño) y se compara con  $\alpha/m$ . Si es más pequeño, se rechaza la hipótesis nula y se procede a comparar  $p_2$  con  $\alpha/(m-1)$ . Si es más grande, el procedimiento para  $y$  ya no puede rechazarse ningún test posterior. Así se va aplicando sucesivamente al resto de valores. En general, se compara  $p_i$  con

$$\hat{\alpha}_i = \frac{\alpha}{(m - i + 1)}$$

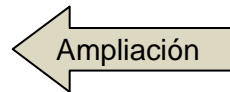
### Example: Holm-Bonferroni

- P-values:  
 $H_{0(1)}$ : 0.005,  $H_{0(2)}$ : 0.011,  $H_{0(3)}$ : 0.02,  $H_{0(4)}$ : 0.04,  $H_{0(5)}$ : 0.13
- $M = 5$  tests; Significance level: 0.05
- Corrected p-value:  $0.005 \cdot 5 = 0.025 < 0.05$ : Reject  $H_{0(1)}$
- Corrected p-value:  $0.011 \cdot 4 = 0.044$  : Reject  $H_{0(2)}$
- Corrected p-value:  $0.02 \cdot 3 = 0.06$ : Don't reject  $H_{0(3)}$  and stop
  
- Conclusion:  
Reject  $H_{0(1)}$  and  $H_{0(2)}$ , don't reject  $H_{0(3)}$ ,  $H_{0(4)}$ ,  $H_{0(5)}$

En el caso de **Hochberg**, el procedimiento es similar pero se empieza con el menos significativo. Otros métodos: **Holland**, **Hommel**, **Rom**, **Li**. etc.



### 13.3 ANOVA de medidas repetidas



Si se toman varias medidas sobre un mismo sujeto hay una evidente dependencia, por lo que se viola esta restricción del ANOVA. Es la misma situación que el T-test de muestras pareadas, sólo que ahora tenemos más de dos medidas.

Por ejemplo, se aplican tres tratamientos a cada uno de los sujetos y se ven los valores de un indicador después de aplicar los tres tratamientos (es un caso de tres *medidas repetidas*)

O por ejemplo, se aplica un tratamiento y se ve un indicador anotándose durante 5 semanas el valor de dicho indicador (es un caso de cinco *medidas repetidas*)

Para ver cómo se construye el estadístico correspondiente consultad:

<https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php>

En SPSS habría que seleccionar Analizar/Modelo Lineal General/Medidas Repetidas

## 13.4 Anova de varias vías (Multi-way ANOVA)

Ampliación

¿Y qué pasa cuando tenemos **varias variables independientes**? En el caso de que queramos agrupar considerando varias variables independientes, estudiando las posibles interacciones entre ellas, debemos seleccionar Analizar/Modelo Lineal General/Univariante/

Nota. No confundir con MANOVA (Multivariate ANOVA) que se aplica cuando se tienen en cuenta simultáneamente varias variables dependientes.

En el directorio de instalación de SPSS/Tutorial/Sample\_files, se encuentra la base de datos grocery\_coupons. Contiene información sobre ventas realizadas en una tienda. Hay muchos datos repetidos ya que cada tupla corresponde a los datos de compra de un cliente en una semana determinada. En la base de datos **grocery\_1month** se han fundido estas tupla (roll up), a través de la variable week. Se han eliminado aquellas variables que no se podían resumir, y se ha calculado la suma de las ventas en el campo amtspent. Así pues, este campo representa la suma de las compras realizadas por un mismo cliente durante un mes.

⇒ Cargad en SPSS el conjunto de datos grocery\_1month

⇒ Analizar/Modelo Lineal General/Univariante -> Factores Fijos

Seleccionad Cantidad Gastada (gasto) –en inglés: Amount Spent (amtspent)- como variable dependiente y Tipo de compra (stilo) –en inglés: Shopping Style (style)- como factor fijo. En Gráficos, seleccionad la variable independiente Tipo de compra (stilo) en el eje horizontal y seleccionad “Añadir”. Finalmente, Post-Hoc para style -> Tukey

### Comparaciones múltiples

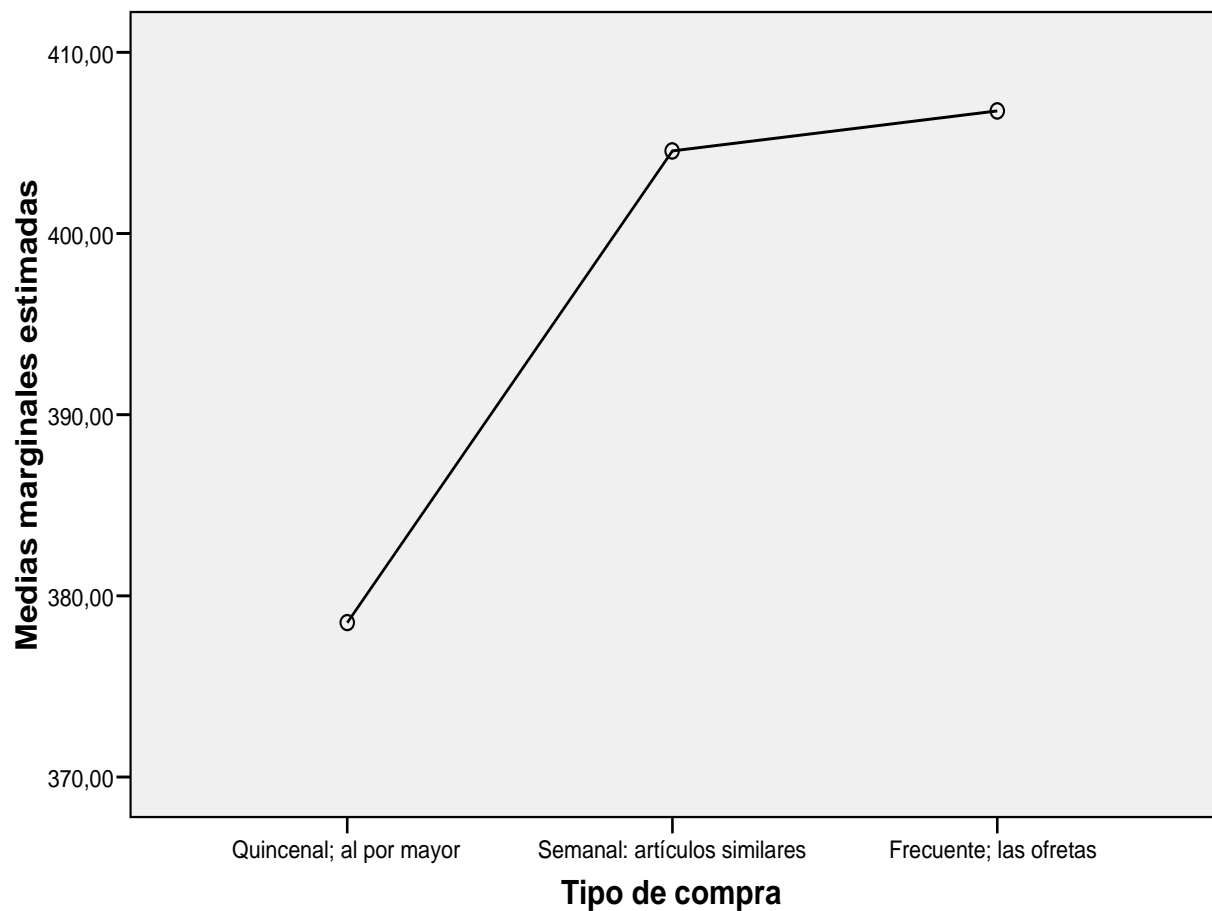
Variable dependiente: Cantidad gastada

DHS de Tukey

		Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%.	
(I) Tipo de compra	(J) Tipo de compra				Límite inferior	Límite superior
Quincenal; al por mayor	Semanal: artículos similares	-26,0342	13,44903	,130	-57,6903	5,6220
	Frecuente; las ofretas	-28,2471	17,33984	,235	-69,0614	12,5671
Semanal: artículos similares	Quincenal; al por mayor	26,0342	13,44903	,130	-5,6220	57,6903
	Frecuente; las ofretas	-2,2130	14,37063	,987	-36,0383	31,6124
Frecuente; las ofretas	Quincenal; al por mayor	28,2471	17,33984	,235	-12,5671	69,0614
	Semanal: artículos similares	2,2130	14,37063	,987	-31,6124	36,0383

Basado en las medias observadas.

### Medias marginales estimadas de Cantidad gastada

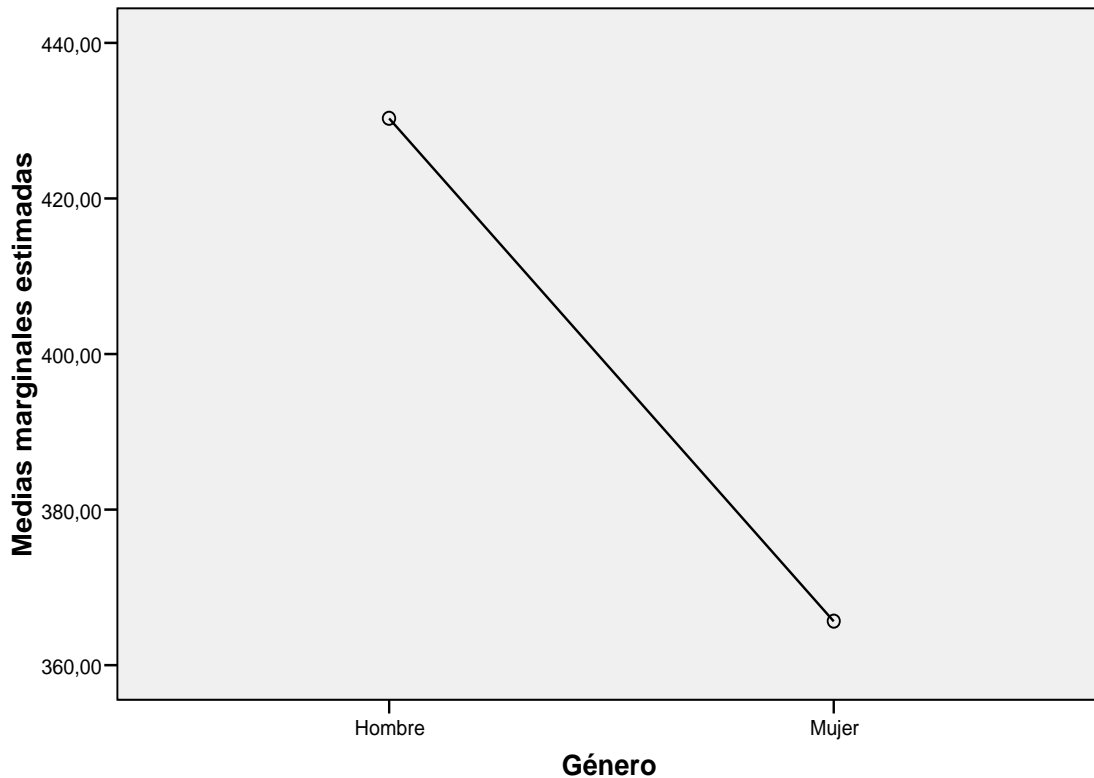


En el ANOVA resultante, no es significativa la diferencia entre las medias. Estos mismos resultados pueden obtenerse desde Analizar/Comparar Medias/Anova de un factor.

Concluimos que no merece la pena hacer campañas de marketing para fomentar a los consumidores que compren menos espaciadamente en el tiempo.

Haced lo mismo con el sexo (Género).

### Medias marginales estimadas de Cantidad gastada



### Pruebas de los efectos inter-sujetos

Variable dependiente: Cantidad gastada

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	365542,589 <sup>a</sup>	1	365542,589	42,188	,000
Intersección	55432908,0	1	55432908	6397,686	,000
genero	365542,589	1	365542,589	42,188	,000
Error	3023919,231	349	8664,525		
Total	59475118,4	351			
Total corregida	3389461,820	350			

a. R cuadrado = ,108 (R cuadrado corregida = ,105)

Ahora, las diferencias sí son significativas (aunque pueda parecer que la pendiente es similar a la del anterior gráfico, hay que tener en cuenta la **escala** de las ordenadas, que ahora está entre 360 y 430 y antes era entre 380 y 405).

Sin embargo, queremos ver si hay alguna interacción entre Tipo de compra (stilo) y Género. Seleccionamos la misma dependiente y ambas como factores fijos. En Gráficos, stilo en el Eje Horizontal y género en "líneas distintas". Añadimos el gráfico stilo\*género. Opciones. Mostrar medias para todas.

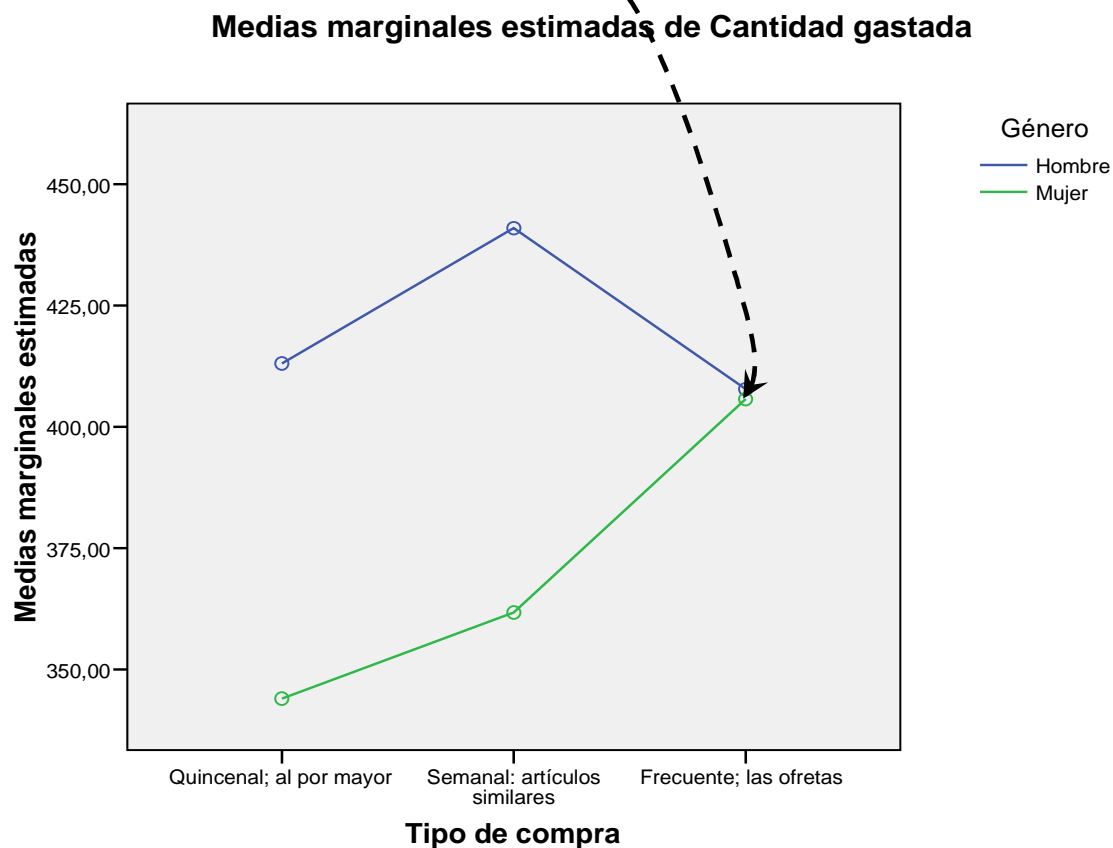
**Pruebas de los efectos inter-sujetos**

Variable dependiente: Cantidad gastada

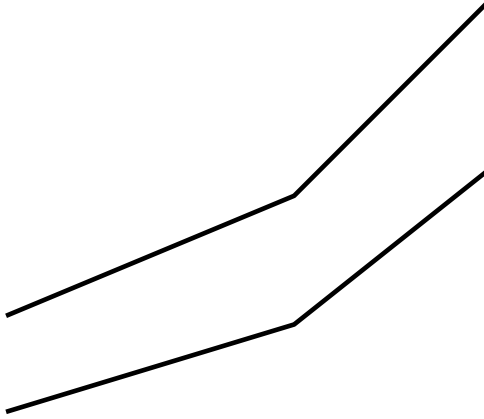
Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	469402,996 <sup>a</sup>	5	93880,599	11,092	,000
Intersección	39359636,4	1	39359636	4650,274	,000
genero	158037,442	1	158037,442	18,672	,000
stilo	33506,210	2	16753,105	1,979	,140
genero * stilo	69858,325	2	34929,163	4,127	,017
Error	2920058,824	345	8463,939		
Total	59475118,4	351			
Total corregida	3389461,820	350			

a. R cuadrado = ,138 (R cuadrado corregida = ,126)

Sí es significativa la interacción



Si no hubiese interacción, los gráficos habrían salido similares:



**Ejercicio:** ¿Hay alguna interacción entre el sexo y la categoría laboral, en relación al Salario Actual?

**Ejercicio:** Sobre la BD Encuesta General USA 1991, ¿depende el número de años de escolarización de la raza del encuestado? ¿y de la región?

## 13.5 Requisitos ANOVA

Para lanzar un ANOVA es necesario que se verifiquen ciertas hipótesis. En el siguiente enlace se enumeran con detalle:

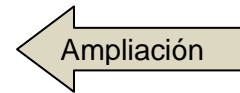
[http://www.basic.northwestern.edu/statguidefiles/oneway\\_anova\\_ass\\_viol.html](http://www.basic.northwestern.edu/statguidefiles/oneway_anova_ass_viol.html)

Las principales hipótesis son:

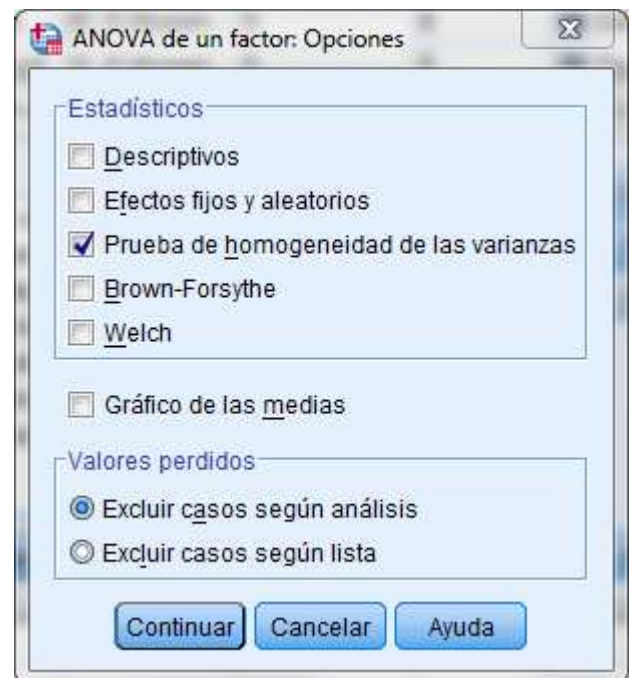
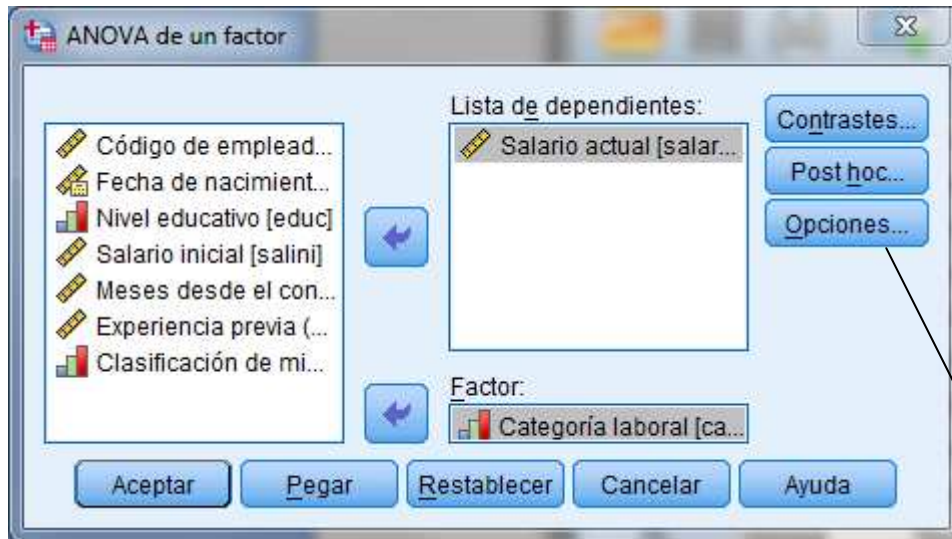
- Independencia. Las observaciones son independientes unas de otras. Esta hipótesis se viola, por ejemplo, en el caso de medidas repetidas (ANOVAr)
- Normalidad. La variable dependiente debe seguir una distribución normal, en cada uno de los grupos formados por la variable independiente. Para comprobarlo se pueden usar varios tests, como por ejemplo el test de Kolmogorov Smirnov (se ve posteriormente) Si se viola, hay que usar tests no paramétricos (se ve posteriormente)
- Homocedasticidad. La varianza de la variable dependiente debe ser similar en cada uno de los grupos formados por la variable independiente. Para comprobarlo, se puede usar el test de Levene (se ve en este apartado) Si se viola, hay que usar tests no paramétricos (se ve posteriormente)

Cuando los tamaños de cada grupo son semejantes y hay bastantes observaciones en cada uno (50 o más), el test ANOVA es bastante robusto a las violaciones de normalidad y homocedasticidad. En caso contrario, debemos recurrir a tests no paramétricos (se ve posteriormente)

## Comprobación de la Homocedasticidad: Test de Levene



Aunque el ANOVA es bastante robusto frente a la violación de la homocedasticidad, si ésta es muy acusada o hay pocos datos, es conveniente recurrir a otros tests no paramétricos. Para comprobarlo, se recurre, por ejemplo, al test de **Levene**:





## 14 Tests no paramétricos

Los tests ANOVA vistos en apartados anteriores se denominan paramétricos ya que presuponen que los datos se ajustan a una distribución concreta, usualmente la normal. Si no se asume ninguna hipótesis sobre la distribución subyacente, debemos recurrir a los tests no paramétricos. La mayor parte de ellos, se basan en el teorema central del límite (visto en el apartado de Estimación puntual)

### 14.1 Ajuste de la distribución

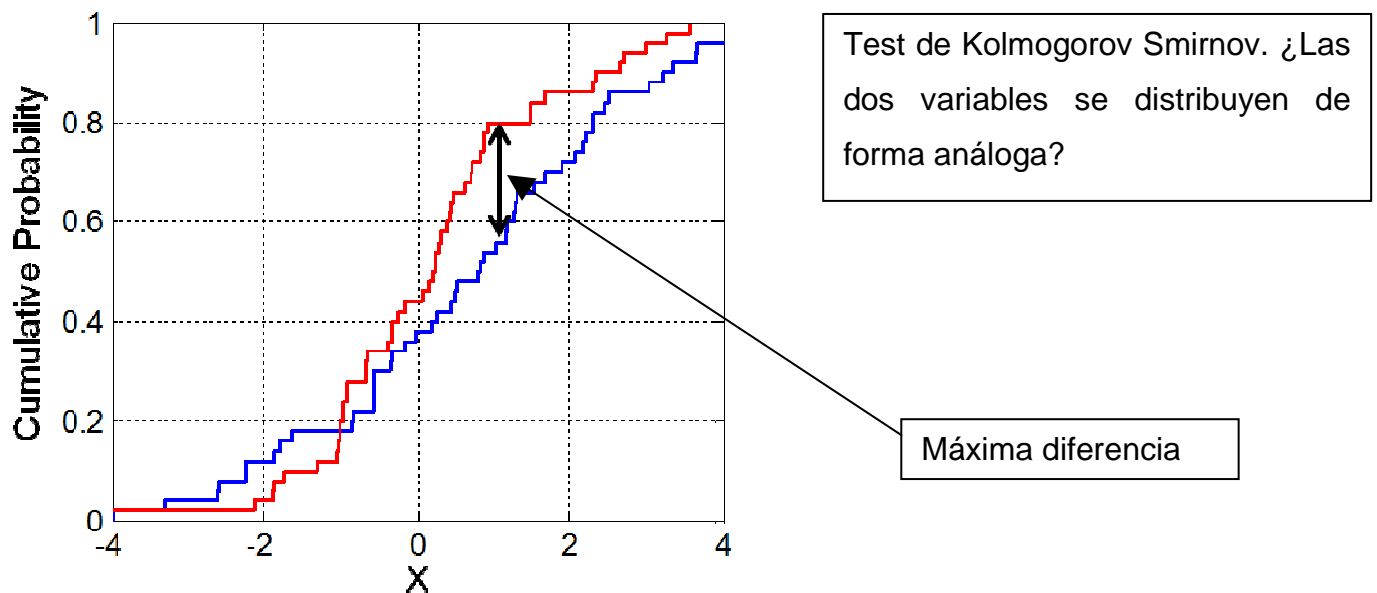


Recordemos que los gráficos Q-Q permiten ver de forma intuitiva si dos distribuciones son similares. Los gráficos P-P son similares pero usan la distribución de probabilidad acumulada. Usando una idea similar, el test de **Kolmogorov Smirnov** compara si dos distribuciones son iguales, para lo que usa la mayor diferencia entre sus funciones de distribución acumulativas.

El test se puede aplicar para comparar una muestra con una distribución concreta o para comparar dos muestras. Veámoslo:

$H_0$ . Dos variables continuas tienen la misma distribución.

$H_1$ . Las variables no tienen la misma distribución.



En este test se usa el siguiente estadístico:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$

$F$  es la función de distribución empírica

$$\frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \leq t\}$$

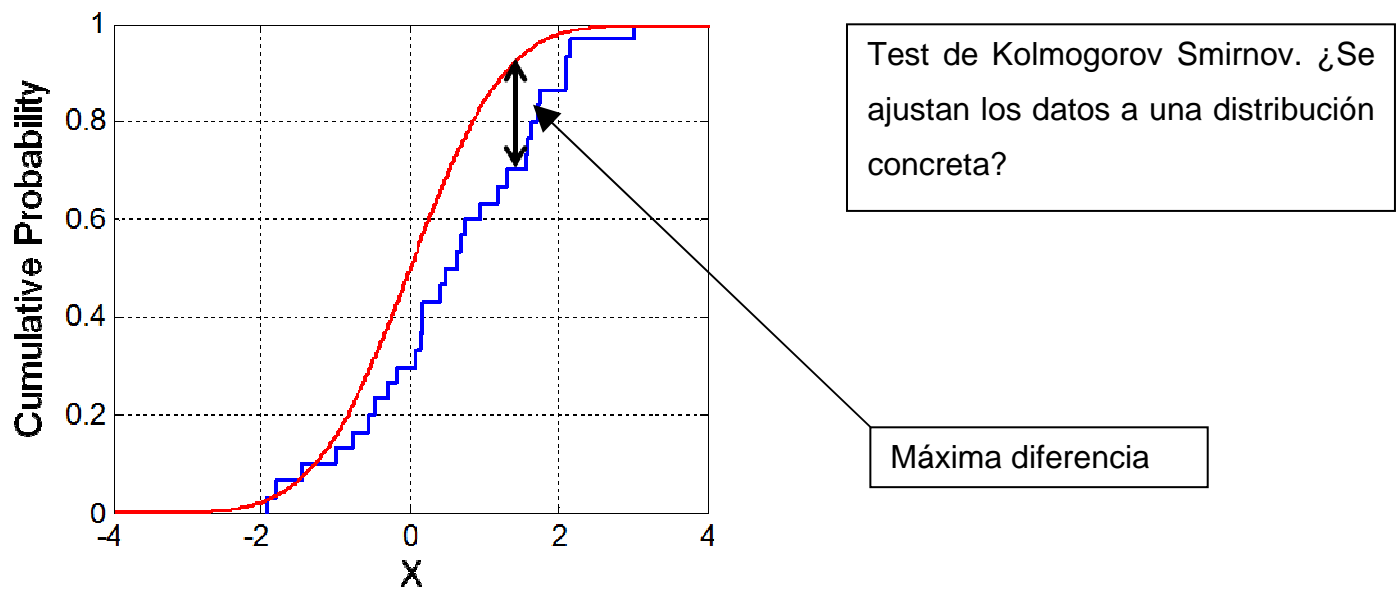
Se rechaza la hipótesis nula si se verifica la siguiente desigualdad:

$$D_{n,n'} > c(\alpha) \sqrt{\frac{n+n'}{nn'}}$$

$\alpha$	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

$H_0$ . La distribución de una variable continua coincide con una distribución concreta (por ejemplo la normal)

$H_1$ . La distribución de la variable no coincide con la distribución fijada.



Como caso particular, podemos contrastar con la distribución normal.

⇒ ¿El Salario actual sigue una distribución normal?

⇒ **Analizar/Pruebas no paramétricas/Una muestra.**

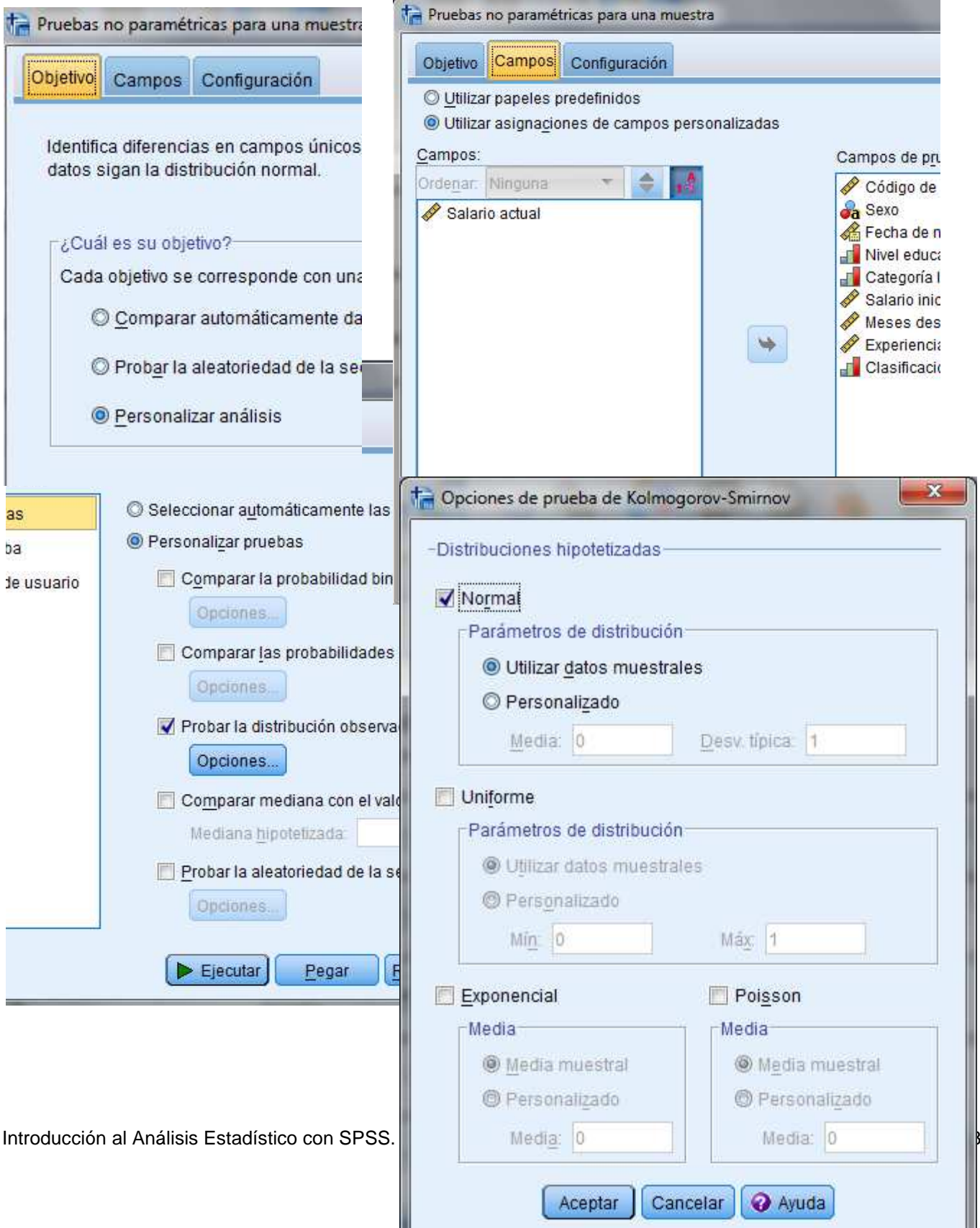
Objetivo: Personalizar Análisis

Campos: Salario Actual

Configuración: Personalizar pruebas → Probar la distribución observada con el valor hip.

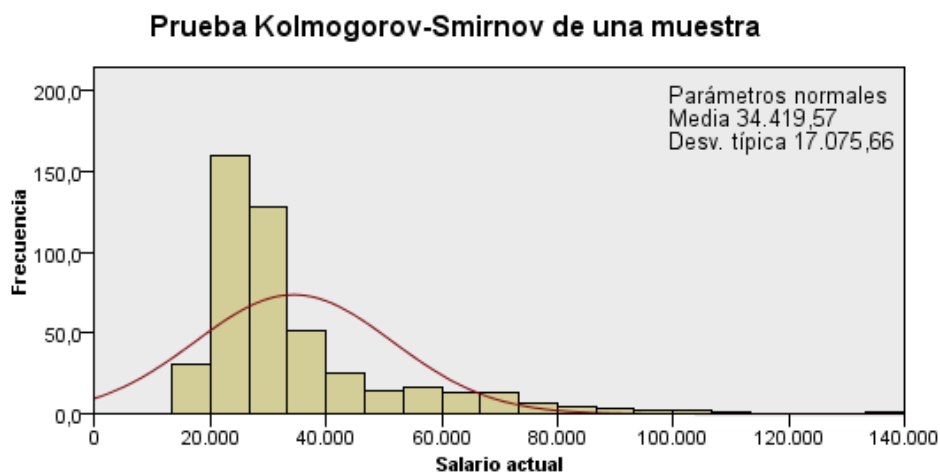
Opciones → Normal

SPSS sólo ofrece cuatro distribuciones.



Resumen de prueba de hipótesis				
	Hipótesis nula	Test	Sig.	Decisión
1	La distribución de Salario actual es normal con la media 34.419,57 y la desviación típica 17.075,66.	Prueba Kolmogorov-Smirnov de una muestra	,000	Rechazar la hipótesis nula.
Se muestran las significancias asintóticas. El nivel de significancia es ,05.				

Sig. muy baja. Rechazamos la hipótesis de que el Salario Actual siga una distribución Normal.



<b>N total</b>		474
<b>Absolutos</b>		,208
<b>Diferencias más extremas</b>	<b>Positivos</b>	,208
	<b>Negativos</b>	-,143
<b>Probar estadística</b>		4,525
<b>Sig. asintótica (prueba de dos caras)</b>		,000

**Ejercicio:** Haced lo mismo con la aceleración en la BD de coches

En este caso no se rechaza, por lo que la muestra no contradice la hipótesis nula.

**Ejercicio:** Plantead distintos test de hipótesis para las medias de algunas variables numéricas de los datos de Mundo 95

## 14.2 Comparación de Medianas. Variables Independientes

A veces, únicamente queremos comprobar si **dos** distribuciones son similares, en el sentido de que su valor medio o sus medianas sean similares. Por tanto, en vez de usar el test de KS (que es más restrictivo), se recurren a otros tests, que pueden verse como la versión no paramétrica de los T-tests y ANOVA.

### Dos variables independientes. Test U de Mann-Whitney

Se aplica cuando tenemos dos muestras de variables independientes no necesariamente del mismo tamaño. Se contrasta la hipótesis de que las medianas son iguales frente a que son distintas (también caben otras alternativas)

Por ejemplo ¿La altura de hombres y mujeres es la misma? ¿Es la mediana del Salario actual la misma en los dos grupos formados por la variable "Clasificación de Minorías"?

Únicamente se tienen en cuenta las situaciones relativas (ranks) de los valores y no sus valores absolutos. Se *mezclan* los datos, se ordenan y se anota el número de orden. Si hay empate (dos valores iguales) se asigna la media aritmética de los ranks correspondientes

Altura de Hombres	Altura de Mujeres	Ranks de hombres	Ranks de mujeres
193	178	1	6/7 => (6+7)/2=6.5
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
178		6/7 => (6+7)/2=6.5	
170		9	
n1 = 7	n2 = 5	R1 = 30.5	R2 = 47.5

Estadístico usado:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

También se podría haber usado R2.

U1 Sigue una distribución U de Mann-Whitney que depende de dos parámetros n1 y n2 (número de valores en el primer y segundo grupo)

Si hay diferencia de altura entre hombres y mujeres, R1 será un valor pequeño y R2 un valor grande

⇒ ¿El salario actual depende de si es minoría étnica o no?

⇒ Analizar/Pruebas no paramétricas/Dos o más muestras independientes.

Pruebas no paramétricas: Dos o más muestras independientes

Objetivo Campos Configuración

☐ Utilizar papeles predefinidos  
☒ Utilizar asignaciones de campos personalizadas

Campos:

Ordenar: Ninguna

- Código de empleado
- Sexo
- Fecha de nacimiento
- Nivel educativo
- Categoría laboral
- Salario inicial
- Meses desde el contrato
- Experiencia previa (meses)

Campos de prueba:

Salario actual

Grupos:

Clasificación de minorías

Pruebas no paramétricas: Dos o más muestras independientes

Objetivo Campos Configuración

Seleccione un elemento:

- Seleccionar pruebas
- Opciones de prueba
- Valores perdidos de usuario

☐ Seleccionar automáticamente las pruebas en f  
☒ Personalizar pruebas

Comparar distribuciones entre grupos

- ☒ U de Mann-Whitney (2 muestras)
- ☐ Kolmogorov-Smirnov (2 muestras)
- ☐ Secuencia de prueba de aleatoriedad (Wald-Wolfowitz para 2 muestras)

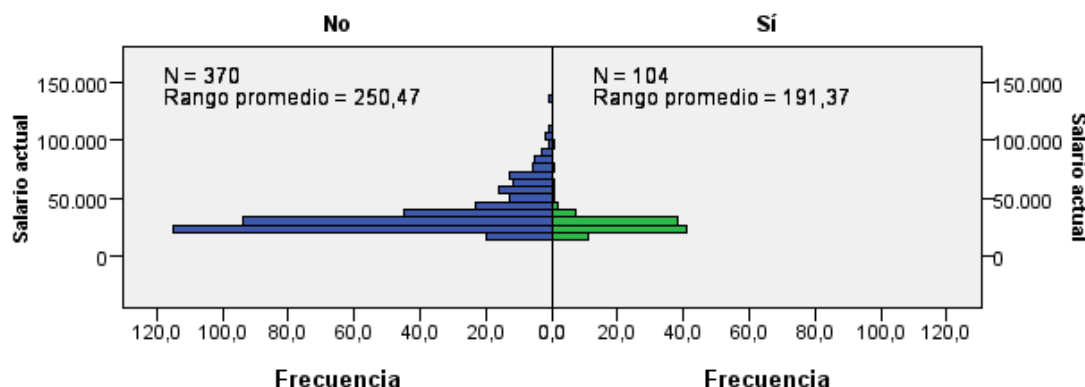
### Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La distribución de Salario actual es la misma entre las categorías de Clasificación de minorías.	Prueba U de Mann-Whitney de muestras independientes	,000	Rechazar la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

### Prueba U de Mann-Whitney de muestras independientes

#### Clasificación de minorías



## ANOVA no paramétrico. Rank test de Kruskal-Wallis

Se aplica en las mismas situaciones que el ANOVA de un factor, pero cuando se violan los requisitos del ANOVA como por ejemplo la normalidad de los datos o la homocedasticidad. Extiende el test U de Mann-Whitney a más de dos variables o grupos. Así pues, se aplica cuando tenemos varias (dos o más) variables independientes o una variable agrupada según factores de otra y queremos comparar las correspondientes medianas.

Se contrasta la hipótesis de que las medianas son iguales frente a que son distintas (también caben otras alternativas)

Por ejemplo, ¿Es diferente el tiempo medio en que se fundirán las bombillas de 100 vatios de tres marcas distintas? (la variabilidad puede ser muy distinta en cada marca)

¿La mediana del Salario actual es la misma en cada categoría laboral? (las distribuciones no son normales)

Únicamente se tienen en cuenta las situaciones relativas (ranks) de los valores y no sus valores absolutos.

Al igual que se hacía en el test de Mann-Whitney, se *mezclan* todos los datos independientemente del grupo al que pertenecen, se ordenan y se anota el número de orden. Si hay empate (dos valores iguales) se asigna la media aritmética de los ranks correspondientes

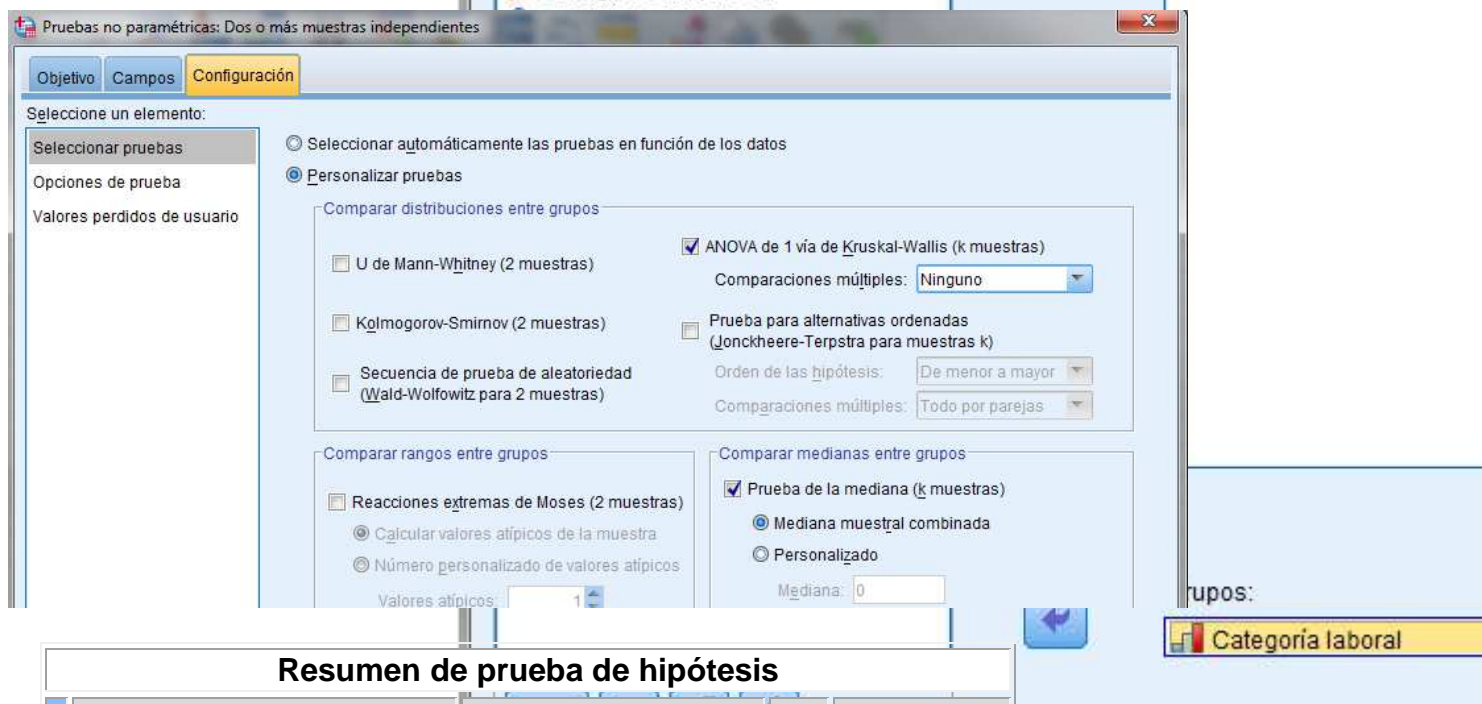
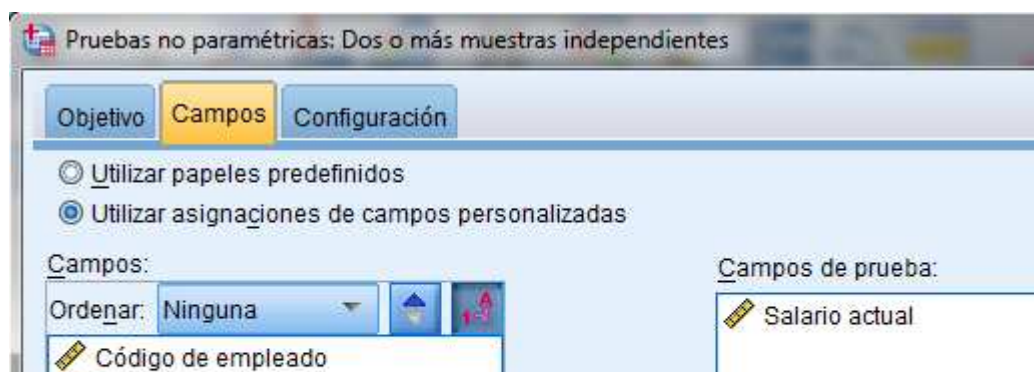
$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \left( \frac{\left( \sum_{j=1}^k R_j \right)^2}{n_j} \right) - 3(N+1)$$

H se distribuye aproximadamente según una Chi Cuadrado de g-1 grados de libertad (g=número de grupos) -al menos 5 datos en cada grupo- En el caso de que haya muchos empates, se recurre a una corrección (ver Sheskin 985)



⇒ ¿Hay diferencias de salario en las distintas categorías laborales?

⇒ Analizar/Pruebas no paramétricas/Dos o más muestras independientes.

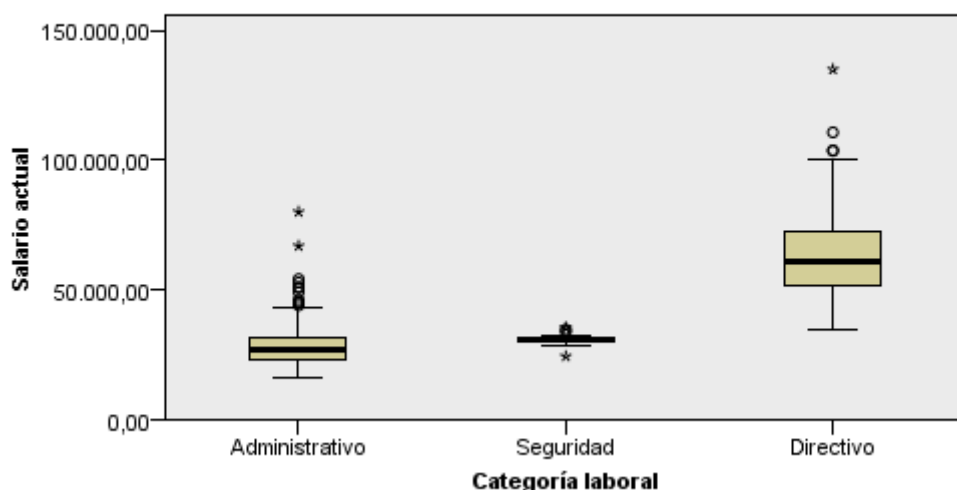


### Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La distribución de Salario actual es la misma entre las categorías de Categoría laboral.	Prueba Kruskal-Wallis de muestras independientes	,000	Rechazar la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

### Prueba Kruskal-Wallis de muestras independientes



## 14.3 Comparación de Medianas. Variables Pareadas

### Dos variables pareadas. Wilcoxon Matched-pairs signed-rank test

Se aplica cuando tenemos dos muestras de variables dependientes del mismo tamaño. Por ejemplo, ¿mide lo mismo la longitud de la pata trasera y delantera de un tipo de ciervo (Zar)? ¿Es el salario inicial igual que el salario final?

Al igual que el test de Mann-Whitney, se contrasta la hipótesis de que las medianas son iguales frente a que son distintas (también caben otras alternativas) y únicamente se tienen en cuenta las situaciones relativas (ranks) de los valores (no se tiene en cuenta las magnitudes de las diferencias). Este test también es válido para datos ordinales.

El proceso es el siguiente:

- Se calcula las diferencias de los pares y se excluyen aquellos con diferencia nula (quedando un total de  $N_r$  valores).
- Se ordenan según el valor absoluto de la diferencia (no se tiene en cuenta el signo).
- Se asigna el rank correspondiente (se sigue el mismo proceso que Mann-Whitney para los empates).
- Se calcula  $K^+$  como la suma de los ranks de aquellas diferencias que eran positivas y  $K^-$  como la suma de los ranks de aquellas diferencias que eran negativas
- Se calcula el estadístico  $W$  como:  **$W = \min\{K^+, K^-\}$**

$W$  sigue una distribución  $W$

de Wilcoxon de parámetro  $n=N-1$ .

Nota: Hay otro test llamado Wilcoxon signed-rank test (**sin matched**) que se utiliza cuando queremos ver si la mediana de una muestra es igual a un valor concreto (es la versión no paramétrica del T-test para una muestra)

n	Two-Tailed Test		One-Tailed Test	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Ejemplo: Medimos el número de palabras reconocidas correctamente por un mismo sujeto por su oído izquierdo y por el derecho. Queremos ver si hay diferencias significativas.

**Number of words reported:**

Participant	Left ear	Right ear
1	25	32
2	29	30
3	10	7
4	31	36
5	27	20
6	24	32
7	27	26
8	29	33
9	30	32
10	32	32
11	20	30
12	5	32
<b>median:</b>	<b>24.08</b>	<b>32.00</b>

Si hay diferencia entre left ear y right ear, se obtendrá un valor de K+ pequeño y un valor de K- grande (o al revés)

Participant	Left ear	Right ear	Difference (d)	ranked difference
1	25	32	-7	7.5
2	29	30	-1	1.5
3	10	7	3	4
4	31	36	-5	6
5	27	20	7	7.5
6	24	32	-8	9
7	27	26	1	1.5
8	29	33	-4	5
9	30	32	-2	3
10	32	32	0	ignore
11	20	30	-10	10
12	5	32	-27	11

Suma de Ranks+ :  $K+ = 4 + 7.5 + 1.5 = 13$

Suma de Ranks-:  $K- = 7.5 + 1.5 + 6 + 9 + 5 + 3 + 10 + 11 = 53$ .

Cuidado: Cuanto más pequeño sea K+, mayor será la fuerza de rechazo

$W = \min\{K+, K-\} = 13$ . ~~N=12~~, por lo que  $n = N - 1 = 11$ , Mirando la tabla con  $\alpha = 0.05$  obtenemos el valor crítico de  $10 < 13$ , por lo que no podemos rechazar la hipótesis de que sean distintos.

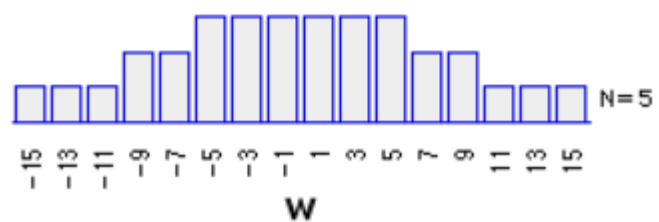
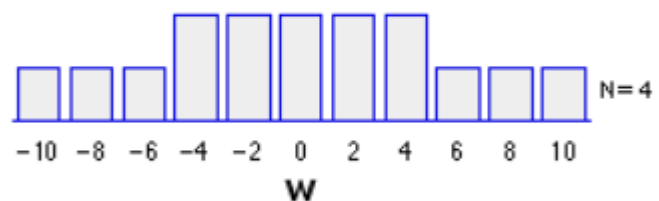
Para un valor de n grande (n mayor de 30, aunque algunos autores lo reducen a 12) se puede recurrir a la aproximación normal (z es una N(0,1))

$$z = \frac{\left| T - \frac{N(N+1)}{4} \right| - 0.5}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

0.5 es un factor de corrección por continuidad. En el caso de que haya muchos empates, se recurre a otra corrección (ver Sheskin, pag 798)

Justificación intuitiva:

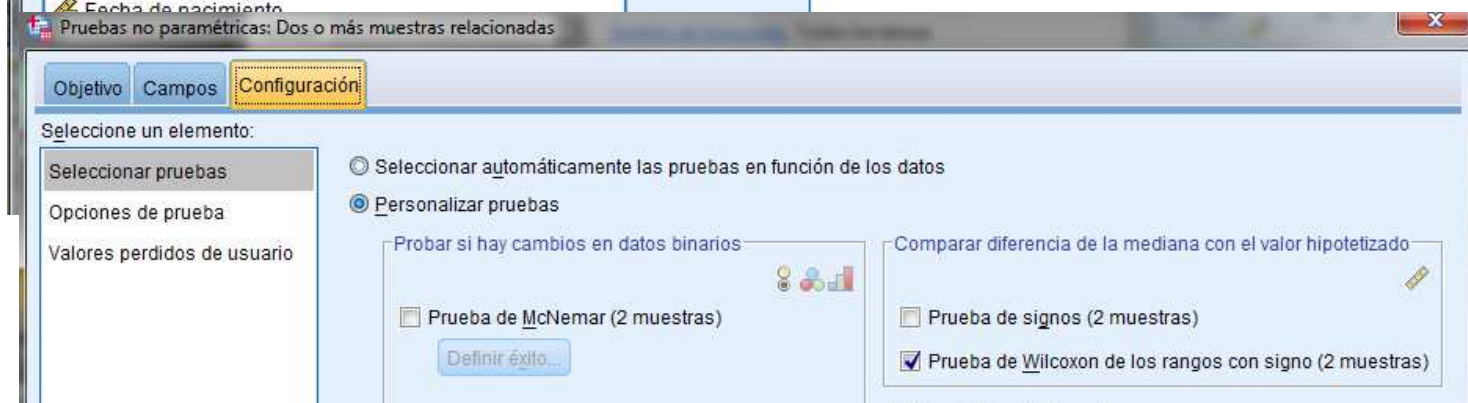
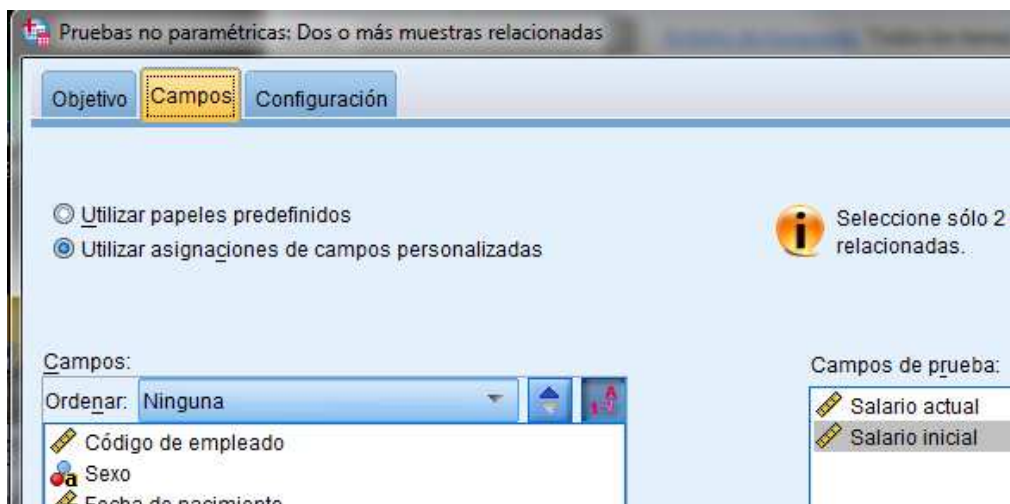
<b>Ranks</b>			
<b>1</b>	<b>2</b>	<b>3</b>	<b>W</b>
+	+	+	+6
-	+	+	+4
+	-	+	+2
+	+	-	0
-	-	+	0
-	+	-	-2
+	-	-	-4
-	-	-	-6



¿Son similares el Salario Actual y el Salario Final?

⇒ Analizar/Pruebas no paramétricas/Muestras relacionadas

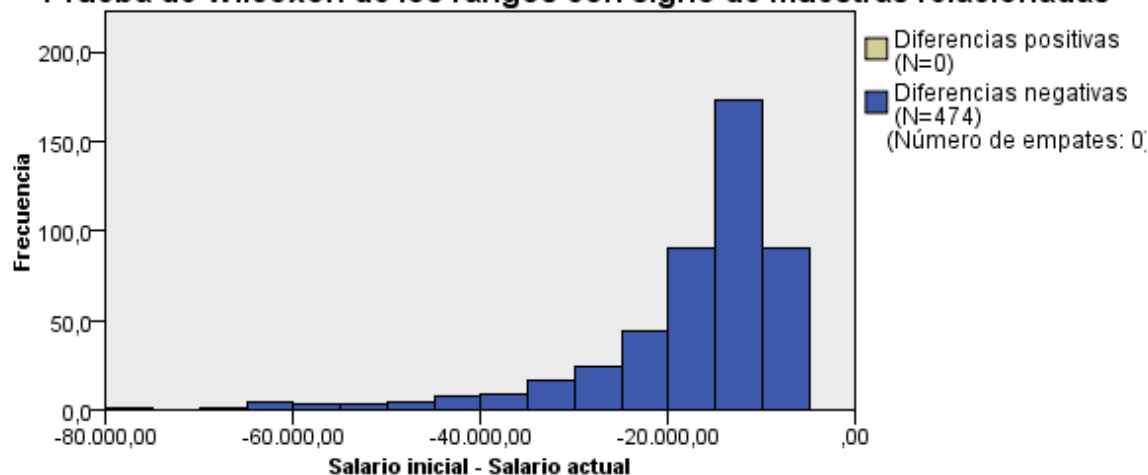
Personalizar Análisis



Resumen de prueba de hipótesis				
	Hipótesis nula	Test	Sig.	Decisión
1	La mediana de las diferencias entre Salario actual y Salario inicial es igual a 0.	Prueba de Wilcoxon de los rangos con signo de muestras relacionadas	,000	Rechazar la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

**Prueba de Wilcoxon de los rangos con signo de muestras relacionadas**



## ANOVA de medidas repetidas no paramétrico: Test de Friedman.

Es la extensión del test de Wilcoxon cuando hay más de dos variables que son dependientes entre sí. O visto de otra forma, es la extensión no paramétrica del ANOVA de medidas repetidas.

¿Asocia la gente diferentes niveles de prestigio a doctores, abogados, policías y profesores? Se pide a varias personas que cada una de ellas valoren cada una de las cuatro profesiones (k=4 medidas repetidas)

La asignación de ranks se hace ahora por cada fila.

Es decir, se ordenan los valores que cada sujeto tiene en cada columna y se asigna un rank de 1 a k, aplicando el mismo criterio de asignación de rank cuando hay empate que ya se vio con el test de Mann-Whitney.

Group 1	Group 2	Group 3	Group 4
70	61	82	74
77	75	88	76
76	67	90	80
80	63	96	76
84	66	92	84
78	68	98	86

Group 1	Group 2	Group 3	Group 4
2	1	4	3
3	1	4	2
2	1	4	3
3	1	4	2
2.5	1	4	2.5
2	1	4	3

Rank medio de cada columna (grupo)

Suma de ranks (en cada grupo)

$$\bar{R}_j = \frac{1}{N} \sum_{i=1}^N r_i^j$$

$$R_j = \sum_{i=1}^N r_i^j$$

$$\chi_r^2 = \frac{12N}{k(k+1)} \sum_{j=1}^k \left( \bar{R}_j^2 - \frac{k+1}{4} \right)^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^k R_j^2 - 3N(k+1)$$

El estadístico tiene una distribución exacta de Friedman (ver tablas en Internet). Se aproxima muy bien con una Chi cuadrado con k-1 grados de libertad.

Una mejora de Friedman viene dada por **Iman, Davemport**:

$$F_F = \frac{(N-1)\chi_r^2}{N(k-1) - \chi_r^2}$$

que sigue una  $F$  con  $(k-1)$  y  $(k-1)(N-1)$  gl

Cargad los datos disponibles en:

[http://www.spss-tutorials.com/downloads/commercial\\_ratings.sav](http://www.spss-tutorials.com/downloads/commercial_ratings.sav)

y convertid las variables en medida de Escala.

Contienen una evaluación (de 1 a 11) sobre los gustos que una persona le da a cuatro versiones de un producto comercial.

	id	com_1	com_2	com_3	com_4
1	1	2	6	7	6
2	2	1	7	7	6
3	5	4	1	5	5
4	8	5	5	6	5
5	9	1	6	7	4
6	10	2	6	6	8
7	11	5	7	3	4
8	12	4	4	4	6

⇒ ¿Hay diferencias de aceptación de las personas entre cuatro productos comerciales?

⇒ Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/k muestras relacionadas, o bien Analizar/Pruebas no paramétricas/Muestras relacionadas



**Estadísticos de contraste<sup>a</sup>**

N	40
Chi-cuadrado	32,808
gl	3
Sig. asintót.	,000

a. Prueba de Friedman



## ⇒ Analizar/Pruebas no paramétricas/Muestras relacionadas

Pruebas no paramétricas: Dos o más muestras relacionadas

Objetivo Campos Configuración

☐ Utilizar papeles predefinidos  
☒ Utilizar asignaciones de campos personalizadas

Seleccione sólo 2 campos de prueba para ejecutar 2 p... relacionadas.

Campos:

Ordenar: Ninguna

Unique respondent identifier

Campos de prueba:

- How appealing did you find the first commercial?
- How appealing did you find the second commercial?
- How appealing did you find the third commercial?
- How appealing did you find the fourth commercial?

Pruebas no paramétricas: Dos o más muestras relacionadas

Objetivo Campos Configuración

Seleccione un elemento:

- Seleccionar pruebas
- Opciones de prueba
- Valores perdidos de usuario

☐ Seleccionar automáticamente las pruebas en función de los datos  
☒ Personalizar pruebas

Probar si hay cambios en datos binarios

- ☐ Prueba de McNemar (2 muestras)  
Definir éxito...
- ☐ Q de Cochran (k muestras)  
Definir éxito...  
Comparaciones múltiples: Todo por parejas

Probar si hay cambios en datos multinomiales

- ☐ Prueba de homogeneidad marginal (2 muestras)

Comparar diferencia de la mediana con el valor hipotetizado

- ☐ Prueba de signos (2 muestras)
- ☐ Prueba de Wilcoxon de los rangos con signo (2 muestras)

Estimar intervalo de confianza

- ☐ Hodges-Lehman (2 muestras)

Cuantificar asociaciones

- ☐ Coeficiente de concordancia de Kendall (k muestras)  
Comparaciones múltiples: Todo por parejas

Comparar distribuciones

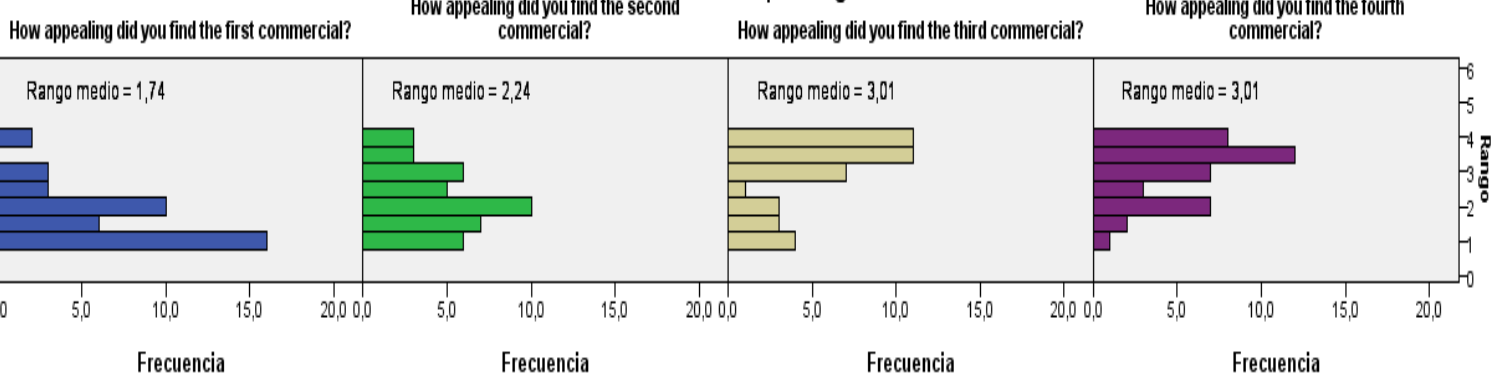
- ☒ ANOVA de 2 vías de Friedman por rangos (k muestras)  
Comparaciones múltiples: Todo por parejas



## Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	Las distribuciones de How appealing did you find the first commercial?, How appealing did you find the second commercial?, How appealing did you find the third commercial? and How appealing did you find the fourth commercial? son las mismas.	Análisis de dos vías de Friedman de varianza por rangos de muestras relacionadas	,000	Rechazar la hipótesis nula.
Se muestran las significancias asintóticas. El nivel de significancia es ,05.				

## Análisis de dos vías de Friedman de varianza por rangos de muestras relacionadas



N total	40
Probar estadística	32,808
Grados de libertad	3
Sig. asintótica (prueba de dos caras)	,000

## Post-hoc en los tests no paramétricos.

Al igual que en el caso paramétrico, se supone que hemos aplicado un test (en este caso no paramétrico) y ha salido significativo. Nos preguntamos qué medianas son distintas. Procedemos a hacer un post-hoc. Tenemos dos alternativas:

- a) Single test procedures (STPs) de una distribución: Test de **Nemenyi** (los tamaños de los grupos han de ser iguales) Similar al de Tukey en el caso paramétrico.
- b) Ajustes del nivel de significación. (STPs de un único ajuste como Bonferroni y Multiple-Stage tests (MSTs) de ajuste por pasos)

Se aplica un test para comparar las medianas de la variable  $i$  y de la variable  $j$  y se usa como nivel de significación cualquiera de las correcciones que ya se vieron en ANOVA (corrección única de Bonferroni, corrección por pasos de Holm, etc)

Para los tests de comparación de cada par de medianas se utilizan los correspondientes tests para el caso de 2 variables, es decir:

- Usaremos el test de Mann-Whitney como post-hoc del test de Kruskal Wallis
- Usaremos el test de Wilcoxon de muestras pareadas como post-hoc de Friedman.

En el caso de Friedman, también se recurre como post-hoc al siguiente estadístico (aunque la muestra no debe ser pequeña)

$$z = (R_i - R_j) / \sqrt{\frac{k(k+1)}{6N}}$$

Que sigue aproximadamente una  $N(0,1)$ . El valor obtenido se compara, como ya hemos comentado, con el valor de  $\alpha$  previamente ajustado según se utilice Bonferroni (el mismo ajuste para todos), Holm (ajuste de  $\alpha$  por pasos), etc

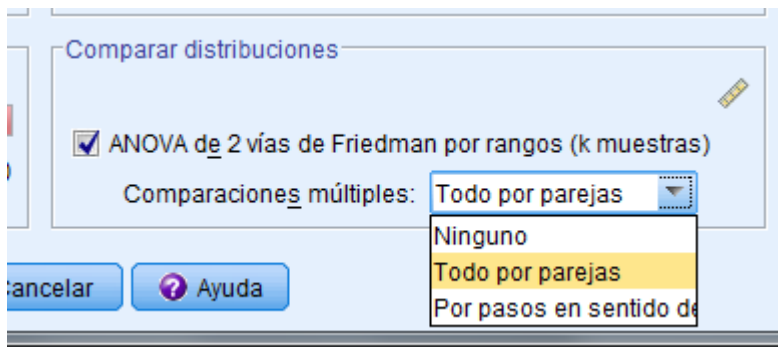
Como ya dijimos cuando vimos las correcciones FWER, si no hacemos todas las comparaciones de todas las variables con todas sino de todas con una variable fija de control, entonces aumenta mucho la potencia de los post hoc.

SPSS sólo permite dos tipos de comparaciones múltiples en los tests no paramétricos:

- Todo por parejas. Ajuste de Bonferroni
- Por pasos. Aplica un método secuencial del tipo step-down pero no va ajustando el nivel de significación en cada test. Aplica el método de **Campbell and Skillings** para formar grupos homogéneos. Se ordenan los niveles de significación de mayor a menor y se compara el primero con el segundo y se van añadiendo los siguientes hasta que sale significativo (al llegar a la  $j$ -ésima). Todos ellos forman un grupo homogéneo. Se repite el proceso con la  $j+1$ -ésima.

Ya se use un método u otro, SPSS no permite fijar un control en los tests no paramétricos.

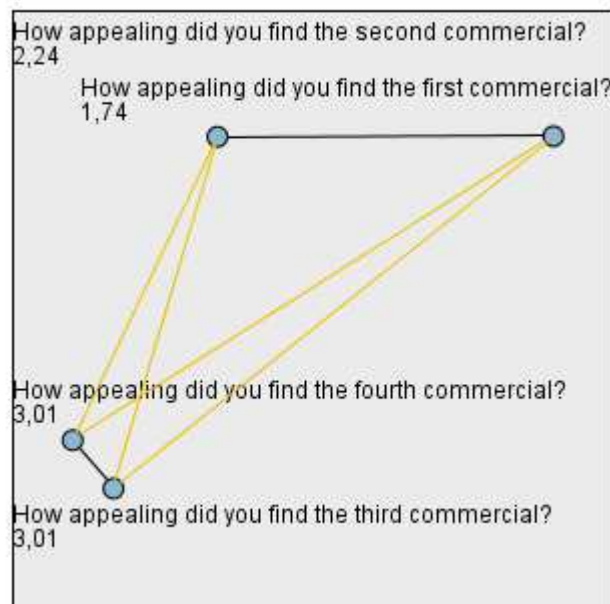
⇒ Analizar/Pruebas no paramétricas/Muestras relacionadas



### Comparaciones por parejas

En negro: No significativo.  
En naranja: Significativo

Todo por parejas



Por pasos

Subconjuntos homogéneos		
		Subconjunto
		1 2
Muestra <sup>1</sup>	How appealing did you find the first commercial?	1,738
	How appealing did you find the second commercial?	2,238
	How appealing did you find the third commercial?	3,013
	How appealing did you find the fourth commercial?	3,013
Probar estadística		3,600 ,025
Sig. (prueba de 2 caras)		,058 ,874
Sig. ajustada (prueba de 2 caras)		,112 ,984
Los subconjuntos homogéneos se basan en significancias asintóticas. El nivel de significancia es ,05.		
<sup>1</sup> Cada casilla muestra el rango de media de muestras.		

	Paramétrico	No Paramétrico
<b>Dos</b> variables independientes. Muestras del mismo o distinto tamaño	<b>T-test.</b> SPSS: Analizar/Comparar Medias/Prueba T para muestras independientes.	<b>Mann-Whitney U Test</b> SPSS: Analizar/Pruebas no paramétricas/Dos o más muestras independientes
<b>Dos</b> variables <b>pareadas</b> (dos medidas repetidas)	<b>Paired samples T-test</b> SPSS: Analizar/Comparar Medias/Prueba T para muestras relacionadas.	<b>Wilcoxon Signed-rank test</b> SPSS: Analizar/Pruebas no paramétricas/Dos o más muestras relacionadas
A ) Comparación de <b>varias</b> variables independientes B) Una variable dependiente y una independiente (factor) que forma dos o más grupos	<b>ANOVA</b> Post Hoc: Tukey, (Bonferroni,Holm, etc.)  SPSS: Analizar/Comparar Medias/Anova de un factor	<b>Kruskall-Wallis</b> Post Hoc: Dunn, (Bonferroni, Holm, etc. usando Mann-Whitney U para cada uno de los tests) SPSS: Analizar/Pruebas no paramétricas/Muestras independientes
<b>Varias</b> variables <b>pareadas</b> (más de dos medidas repetidas)	<b>ANOVA</b>  SPSS: Modelo Lineal General/Medidas repetidas	<b>Friedman</b> Post Hoc: Nemenyi, Z, (Bonferroni, Holm, etc. usando Wilcoxon para cada uno de los tests)  SPSS: Analizar/Pruebas no paramétricas/muestras relacionadas

Los tests no paramétricos se pueden aplicar en cualquier situación, pero si se dan las condiciones para aplicar un test paramétrico, se aplicará este último ya que un test no paramétrico siempre tienen menos potencia que el correspondiente paramétrico.

## 15 Comparación de Clasificadores

Queremos responder la siguiente pregunta:

¿Es un clasificador mejor que otro(s)?

Metodología:

- Se seleccionan los clasificadores que se van a comparar (columnas).
- Se seleccionan un conjunto de datasets como conjuntos de prueba (filas).
- Se mide la tasa de acierto de cada algoritmo en cada dataset.

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	<b>81.2</b>	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	<b>85.7</b>
bal	45.0	76.2	87.2	<b>90.4</b>	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	<b>68.2</b>
bps	51.6	83.2	82.8	78.6	80.1	79.0	<b>83.3</b>
bre	65.5	96.0	<b>96.7</b>	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	<b>72.6</b>
h-c	54.5	77.4	83.2	<b>83.6</b>	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	<b>95.3</b>	<b>95.3</b>	94.7	<b>95.3</b>	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	<b>95.2</b>	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	<b>74.9</b>	75.1	74.8
lym	55.0	83.3	83.6	<b>85.6</b>	77.0	71.5	79.0
mmg	56.0	63.0	<b>65.3</b>	64.7	64.8	61.9	63.4
mus	51.8	<b>100.0</b>	<b>100.0</b>	96.4	<b>100.0</b>	<b>100.0</b>	99.8
mux	49.9	78.6	99.8	61.9	99.9	<b>100.0</b>	<b>100.0</b>
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	<b>50.8</b>	41.6	39.8	43.7
seg	14.3	<b>97.4</b>	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	<b>98.4</b>	97.0	96.7
soyb	13.5	89.5	90.3	<b>92.8</b>	91.4	90.3	76.2
tao	49.8	<b>96.1</b>	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	<b>92.1</b>	<b>92.1</b>	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	<b>96.5</b>	95.4
vow	9.1	99.1	<b>96.6</b>	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	<b>97.8</b>	94.6	92.9	96.3
zoo	41.7	94.6	92.5	<b>95.4</b>	91.6	92.5	92.6
Avg	<b>44.8</b>	<b>80.0</b>	<b>82.4</b>	<b>78.0</b>	<b>82.1</b>	<b>81.8</b>	<b>81.7</b>

¿Qué tipo de test debemos aplicar?

- Al mismo individuo (dataset) se le aplican varios tratamientos (algoritmos). Por tanto, es un caso de **medidas repetidas**.
- El número de ejemplos (datasets) no suele ser elevado (no más de treinta). De hecho (desafortunadamente) hay infinidad de artículos de clasificación que únicamente utilizan de 5 a diez conjuntos de datos.

Por lo tanto, sería preferible un test **no paramétrico**.

- El número de aciertos de cada algoritmo es bastante variable respecto a los datasets, por lo que no suele seguir una distribución normal ni presentar varianzas similares. Por ello, no debemos usar la media aritmética de las tasas de acierto (avg en la tabla anterior) sino las medianas. Por tanto debemos usar un test no paramétrico.

Así pues:

- Si queremos comparar dos clasificadores, utilizaremos el test de **Wilcoxon**  
En el caso de que el test sea significativo, rechazaremos que los algoritmos sean iguales y podremos decir que el mejor es el que tenga mayor mediana (suponemos que estamos midiendo tasa de acierto)
- Si queremos comparar un conjunto de clasificadores, usaremos el test de **Friedman**.  
En el caso de que el test sea significativo y se rechace que los algoritmos tienen medianas similares, tenemos que realizar un post-hoc que nos diga qué algoritmos pueden considerarse similares entre sí, controlando el error FWER. Se nos presentan varias situaciones:
  - Si queremos comparar todos con todos, tenemos dos alternativas:
    - a) Aplicar un ajuste STP como por ejemplo el ajuste de Bonferroni, que divide el nivel de significación (0.05 usualmente) por  $k(k-1)/2$  donde  $k$  es el número de algoritmos. Este es el ajuste que viene por defecto en SPSS.
    - b) Como Bonferroni es muy conservador, mejor deberíamos usar cualquier método secuencial MST que va cambiando el nivel de significación en cada uno de los contrastes, como por ejemplo **Holm** o **Hochberg**. No está disponible en SPSS. Se verá con R.
  - Si queremos comparar uno fijado de antemano (**control**) con el resto (por ejemplo, un algoritmo nuevo diseñado por el investigador) el ajuste que hay que hacer es mucho menor ya que, por ejemplo, si se aplica una corrección STP como Bonferroni, hay que dividir el nivel de significación por  $(k-1)$  ya que sólo hay que hacer  $k-1$  comparaciones. Al igual que antes, sería preferible recurrir a un ajuste con algún método MST, preferiblemente Holm. No está disponible en SPSS. Se verá con R.

Más detalles en:

Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets.  
Journal of Machine Learning Research 7 (2006) 1–30.



## Extensión:



Si estamos interesados en darle más importancia a los casos en los que un algoritmo se porta bastante mejor que los otros algoritmos (la tasa de acierto es notablemente mayor para ciertos datasets) se pueden usar variaciones del test de Friedman:

- Aligned ranks Friedman.

A cada celda  $i, j$  (accuracy del algoritmo  $j$  en el dataset  $i$ ) se le resta la media de la fila  $i$  (accuracy medio de los algoritmos en el dataset  $i$ )

Con todos ( $kN$ ) los valores así obtenidos se forma un único conjunto de ranks y se le aplica el mismo procedimiento que Kruskal-Wallis

- Quade test.

Se calculan los  $k$ -ranks asociados a cada una de las  $N$  filas tal y como se hace en Friedman (total de  $kN$  ranks)

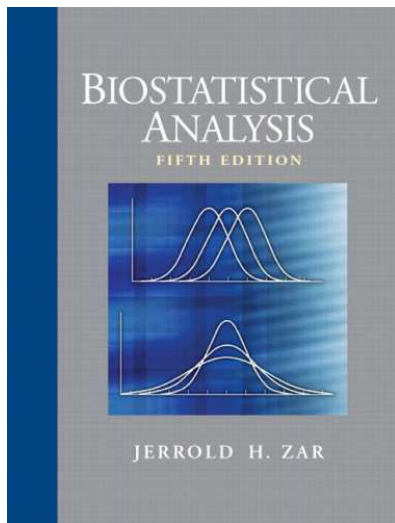
Por cada fila se calcula el máximo menos el mínimo de dicha fila y con los valores así obtenidos se obtiene un conjunto de  $N$  ranks de filas. Estos valores se utilizan para ponderar los ranks obtenidos en el paso anterior.

Más detalles en:

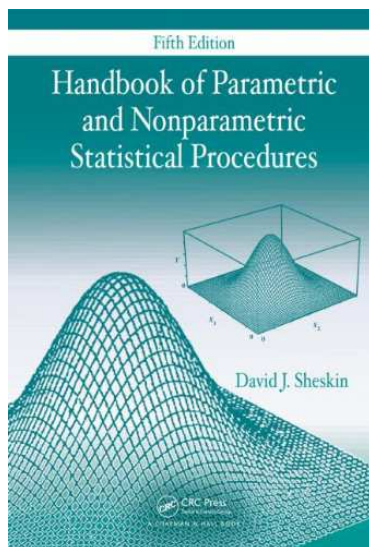
S. García, F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research* 9 (2008) 2677-2694

García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental Analysis of Power. *Information Sciences* 180 (2010) 2044–2064.

## 16 Bibliografía



Biostatistical Analysis. Zar, 2009. 5th edition.  
Pearson



Handbook of Parametric and Nonparametric  
Statistical Procedures, 2011. Fifth Edition.  
Chapman and Hall