Máster en Sistemas Inteligentes y Soft Computing:

M1. Fundamentos de Minería de datos

Sesión 3. Inferencia Estadística. Modelos Paramétricos.

Juan Carlos Cubero

Universidad de Granada

http://decsai.ugr.es/~carlos/main.htm

Grupo IDBIS: Intelligent Databases and

Information Systems

http://idbis.ugr.es

Noviembre 2007

1 Motivación

La Estadística es fundamental en la minería de datos:

1. En sí misma, proporciona técnicas de minería de datos, como por ejemplo, el análisis de componentes principales, regresión, análisis factorial, etc.
2. 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?
3. 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:

1. Suelen requerir que el experto diga exactamente lo que quiere comprobar.
2. Cuando se aplican técnicas estadísticas "clásicas", hay que tener cuidado de que se cumplan ciertos "requerimientos" o "hipótesis de partida". En caso contrario, hay que aplicar técnicas estadísticas "no paramétricas"

Usaremos SPSS como programa de ordenador, debido a su uso generalizado en estudios estadísticos.

2 Introducción a SPSS: Datos y Variables

2.1 ¿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 13.0.

2.2 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"

Hay una vista de datos y vista de variables (tipo hoja de cálculo) - para la versión 10 en adelante-.

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)

2.3 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.

Lo que digamos sobre nominales también sirve para ordinales.

Queremos responder a la pregunta:

¿Cómo se distribuye una variable nominal?

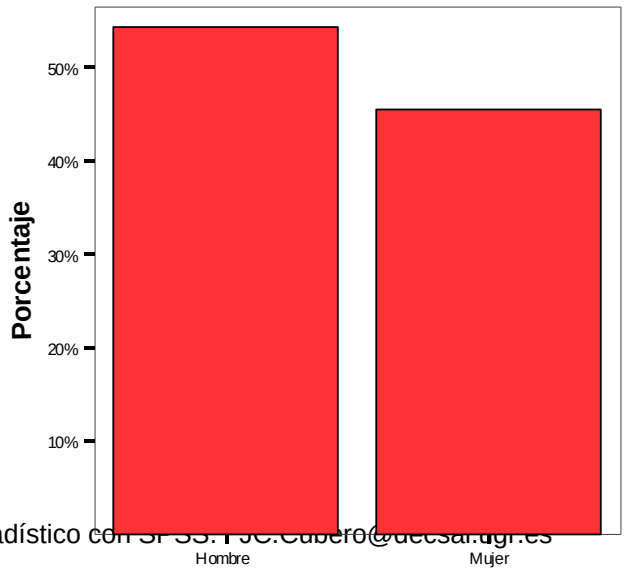
SPSS no ofrece muchas facilidades sobre las variables nominales (por ejemplo, que liste automáticamente los valores distintos). Para ello, tendremos que construir un gráfico y verlo en él.

Usaremos un **gráfico de puntos o de barras**.

⇨ Gráficos/Barras/Simples/Resúmenes para distintas variables/Seleccionad Sexo

y que las barras representen %casos. Aparece el visor de resultados (se guarda en un fichero con extensión spo) Si quisiésemos suprimir cualquier elemento del visor de resultados, basta con borrarlo del panel izquierdo.

Observad que aparecen las etiquetas de los valores



⇨ Gráficos/Interactivos/Barras/Seleccionad Sexo

Podemos ver que la diferencia con un gráfico interactivo es que se incluye un gráfico normal, pero no se abrirá ningún marco interactivo al hacer doble click, sino simplemente un formulario para cambiar el formato de la imagen. Si hacemos doble click sobre el gráfico interactivo se abre un marco dónde 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.

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

4 Análisis Exploratorio de Datos (AED). Gráficos y Estadística Descriptiva sobre una variable de escala.

4.1 Histograma

Queremos responder a la pregunta:

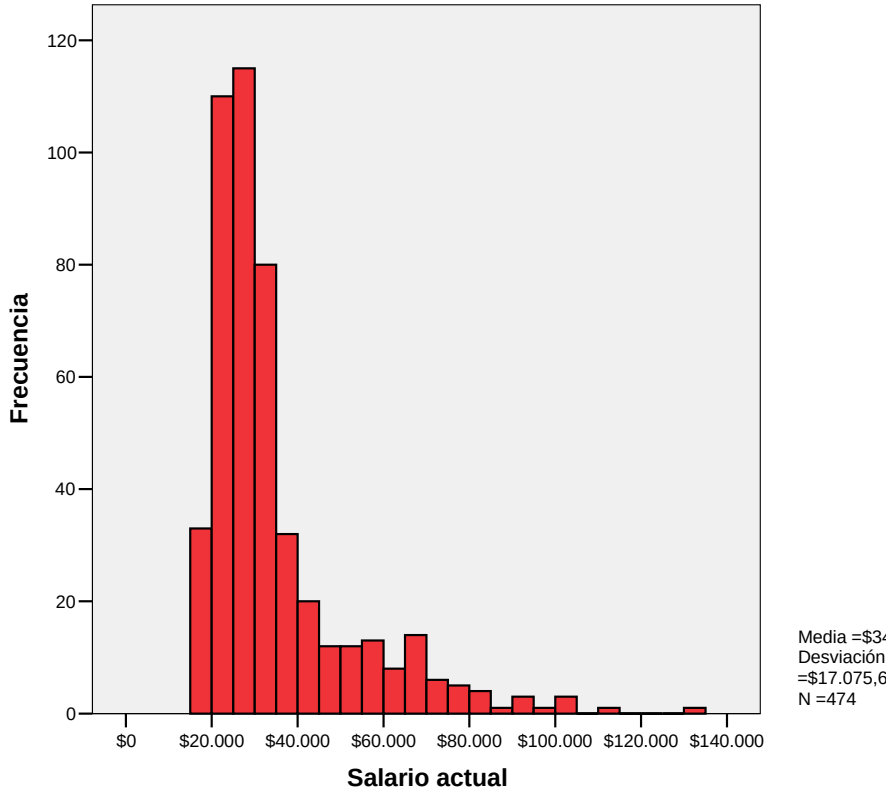
¿Cómo se distribuye una variable numérica?

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.

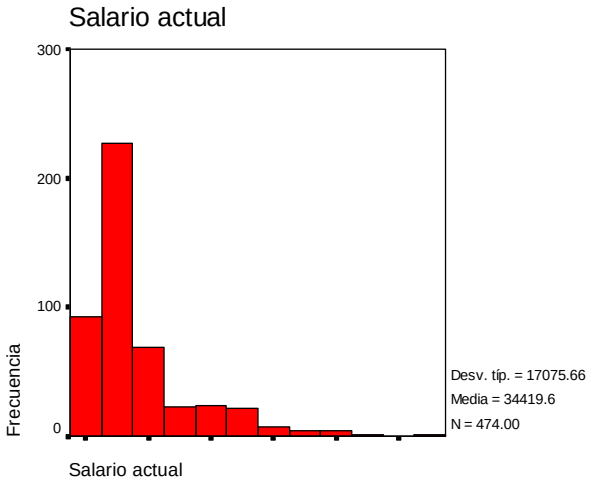
¿Cómo se distribuye el salario entre los empleados?

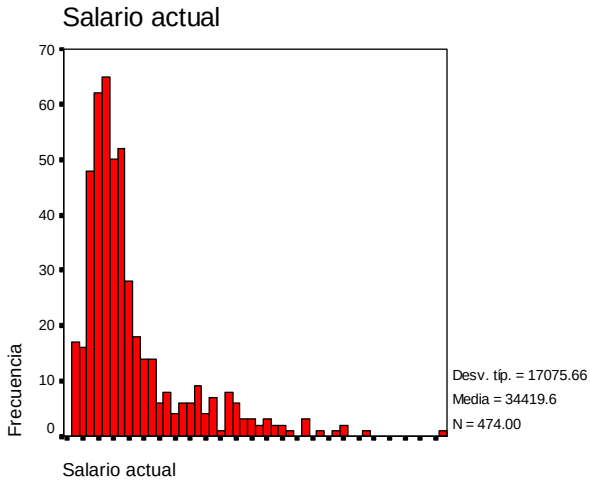
⇨ Gráficos/Interactivos/Histograma/Seleccionad Salario Actual



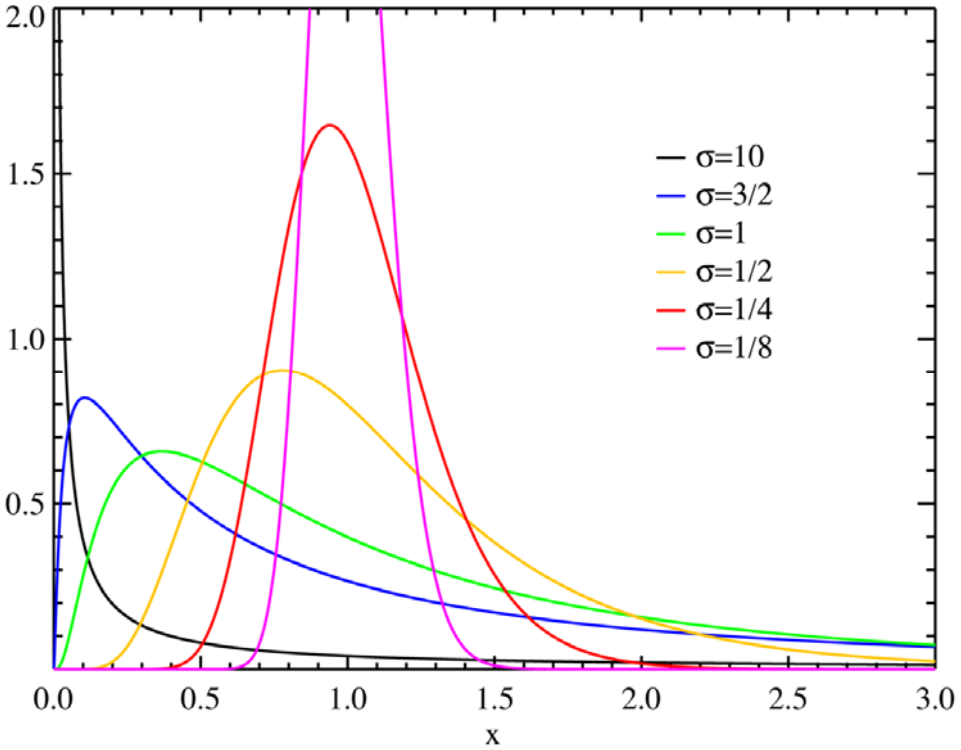
El ancho de los rectángulos (que determina el número de ellos) afecta a la información que muestra el histograma y, específicamente, puede afectar a su forma. Cambiando esta característica (desde la pestaña Histograma) podemos obtener información más precisa y detallada. Por ejemplo, podemos partir de pocos rectángulos e ir detallando progresivamente si

es necesario. Sólo se puede cambiar desde los gráficos interactivos (y no desde Gráficos/Histograma)





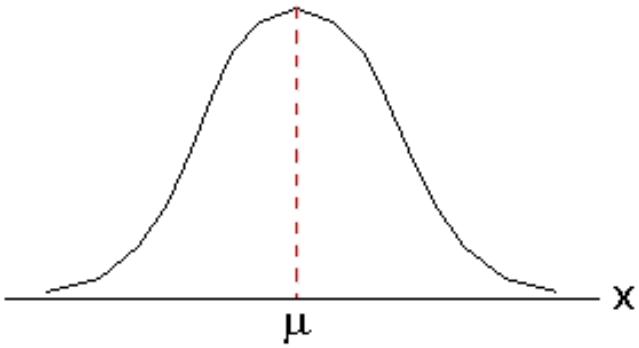
Si extrapolamos el histograma, obtendríamos una función matemática que determinaría la probabilidad con la que se da cada valor. En Estadística se han estudiado muchas funciones. Por ejemplo, la anterior se asemeja a una distibución log-normal:



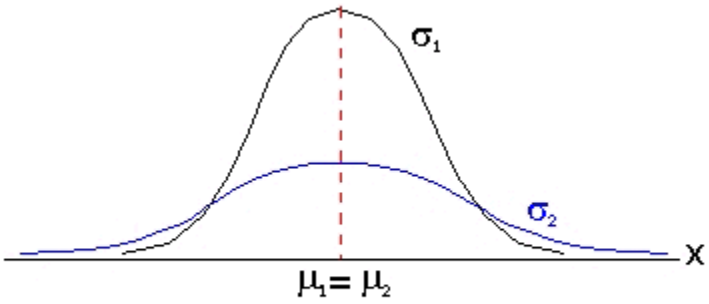
$$f(x;\mu,\sigma) = \frac{e^{-(\ln x-\mu)^2/(2\sigma^2)}}{x\sigma\sqrt{2\pi}}$$

Una distribución muy conocida es la normal:

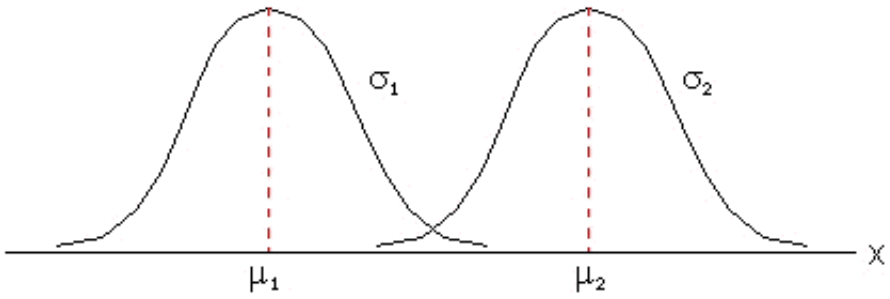
$$N(\mu,\sigma) \quad f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$



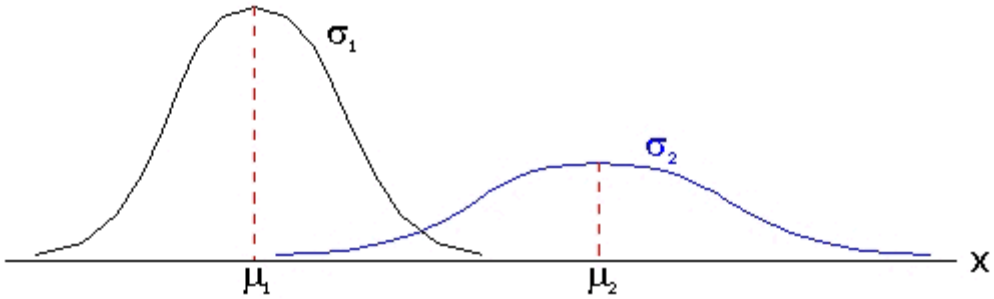
The normal curve



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



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



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

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:

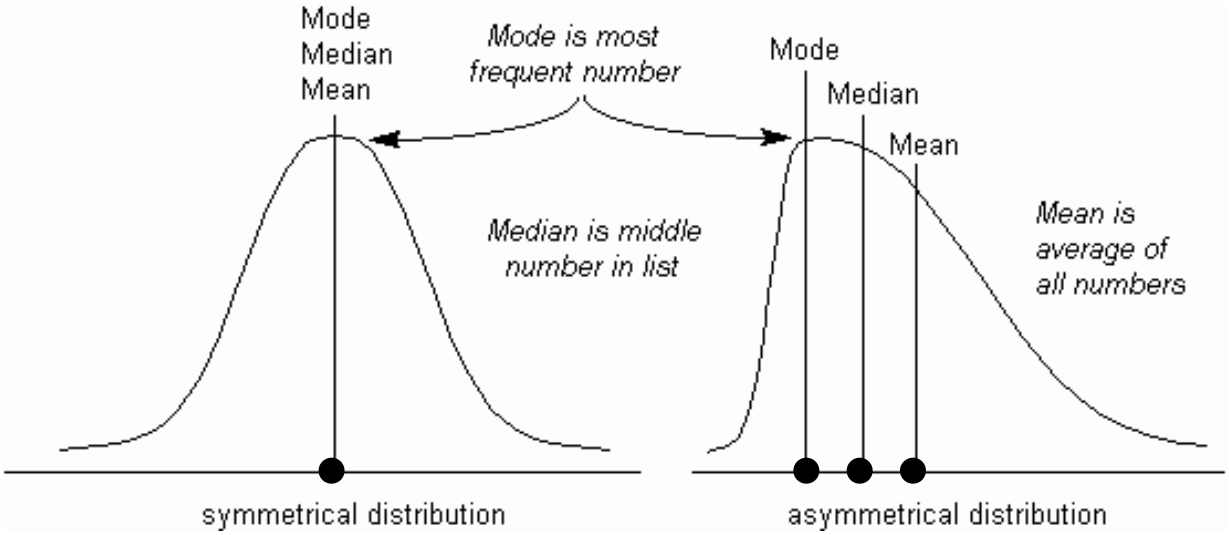
Tamaño de la muestra: N, es el número de casos en la muestra. Por ejemplo, el valor de N para administrativo, seguridad y directivo es 363, 27 y 84 respectivamente.

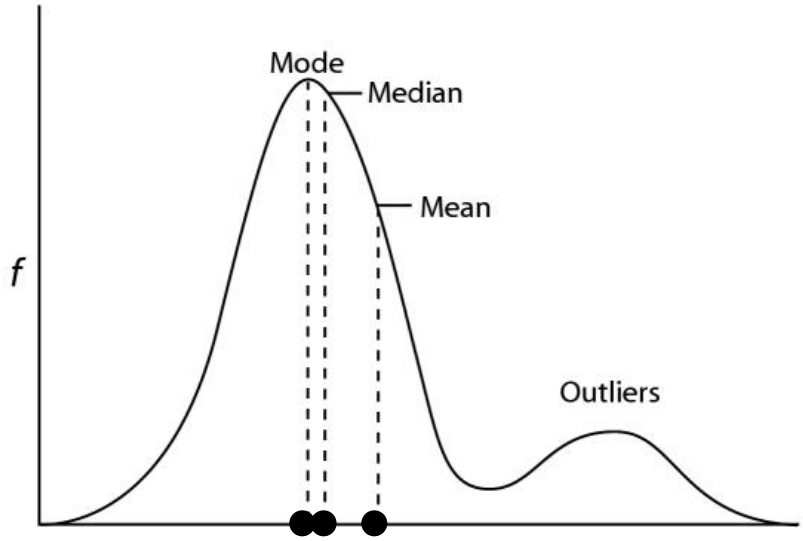
Aparte del tamaño muestral, hay dos tipos importantes de estadísticos:

- ✓ **De localización**. Dan una idea de cuales son los valores habituales de la distribución. Tratan de decirnos dónde la distribución es más densa.
  - **Media** muestral. Problema: sensible a casos aislados (outliers).

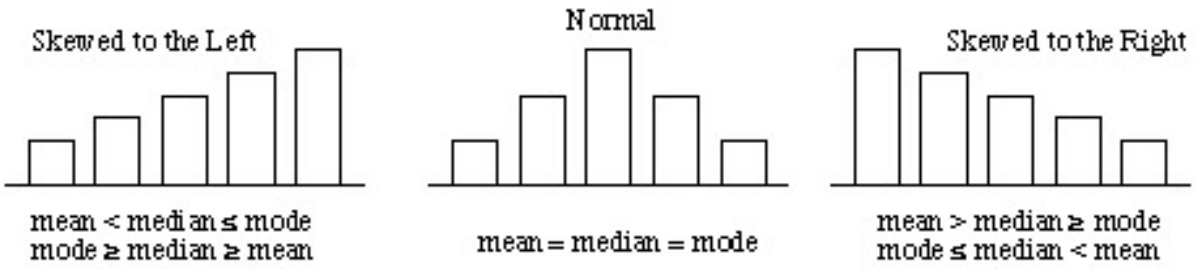
    $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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





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



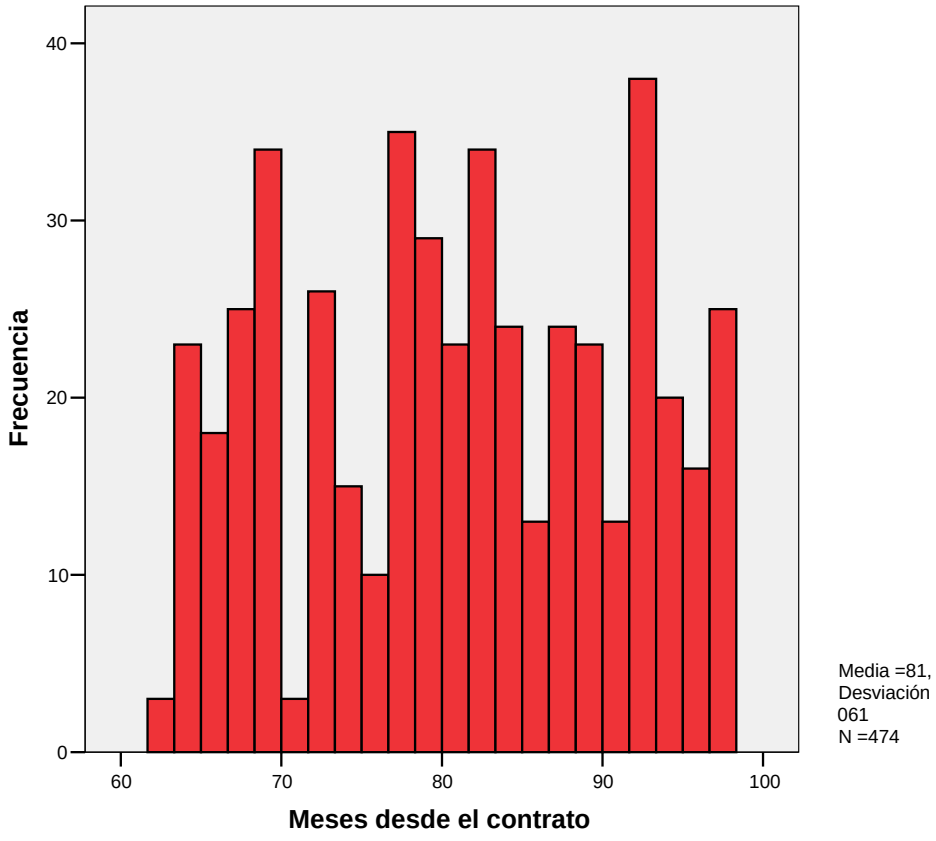
- ✓ **De dispersión.** Dan una idea de cual es la variabilidad en los datos.
  - **Desviación típica** muestral. Representa cómo de dispersos están los datos con respecto a la media aritmética. Es una medida global.

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

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

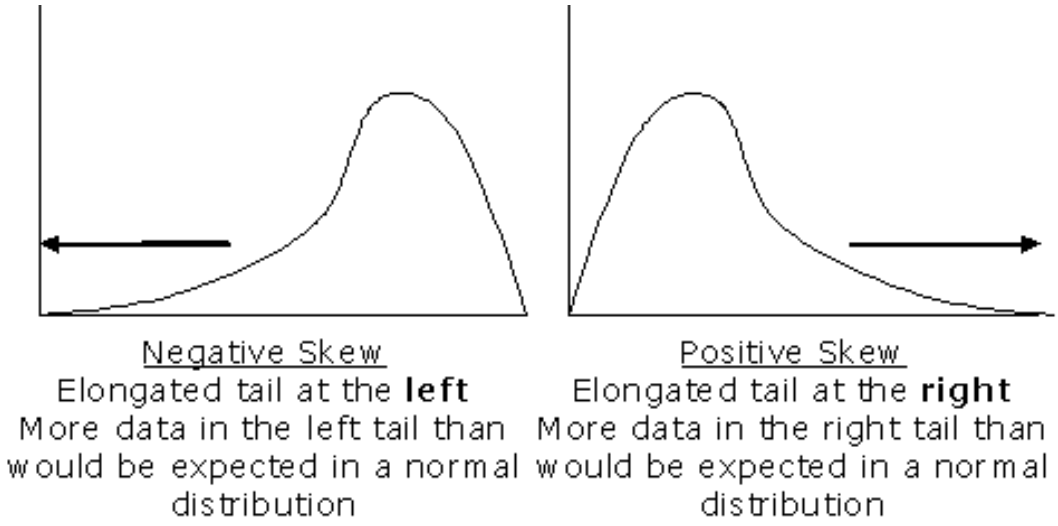
La varianza es el cuadrado de la desviación típica

Mucha varianza



- ✓ **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.



Su definición exacta es el tercer momento tipificado de la distribución, aunque Pearson dió una forma aproximada de calcularlo como (media-moda)/desv.tip
La distribución normal es simétrica por lo que tiene un valor de asimetría 0. Una distribución que tenga una asimetría positiva significativa tiene una cola derecha larga. Una distribución que tenga una asimetría negativa significativa tiene una cola izquierda larga. Un valor de asimetría mayor que 1, en valor absoluto, indica generalmente una distribución que difiere de manera significativa de la distribución normal.

  - **Kurtosis (curtosis).** Medida del grado en que las observaciones están agrupadas en torno al punto central. Para una distribución normal, el valor del estadístico de curtosis es 0. Una curtosis positiva indica que las observaciones se concentran más y presentan colas más largas que las de una distribución normal. Una curtosis negativa indica que las observaciones se agrupan menos y presentan colas más cortas.

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

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

⇨ Analizar/Estadísticos Descriptivos/Frecuencias/Seleccionad 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.

Con apenas cuatro valores de resumen, nos hacemos una idea muy aproximada de cual es la distribución de los datos. La media está en torno a los $34.400. La mitad de los trabajadores ganan menos de $28.875 y la otra mitad gana más. 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 $[34.419-2*17.075, 34.419+2*17.075] = [0.419, 68.4]$

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

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

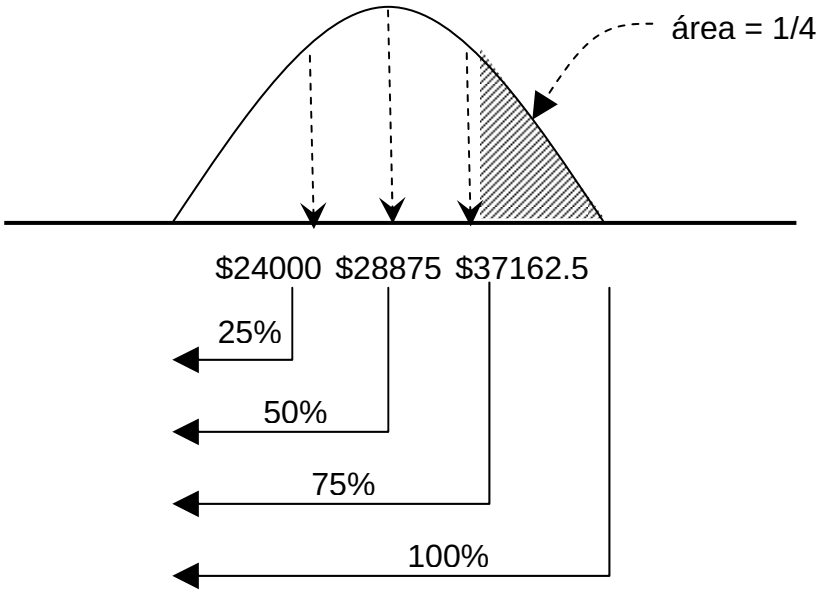
	Salario actual	Experiencia previa (meses)
N	474	474
Media	$34,419.57	95,86
Desv. típ.	$17,075.661	104,586

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

Podemos incluir varias variables en la misma tabla de resumen:

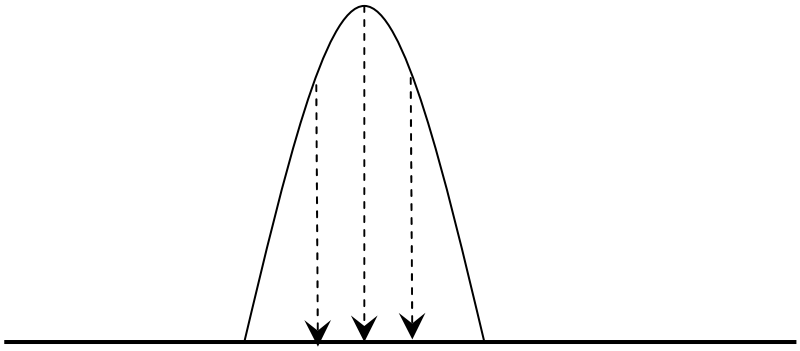
⇨ Analizar/ Informes/Resúmenes de Casos/ Seleccionad Salario Actual y Experiencia Previa. No mostrad los casos y como estadísticos, seleccionad la media y desviación típica

Los **percentiles** son unos estadísticos de tendencia central, pero que también ofrecen información sobre la dispersión de los datos. El percentil 25 es un valor tal que el 25% de los valores de la muestra son menores que él. Y así con el resto (obviamente, el percentil 50 es la mediana)



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

Si dos cuartiles 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.

Estadísticos

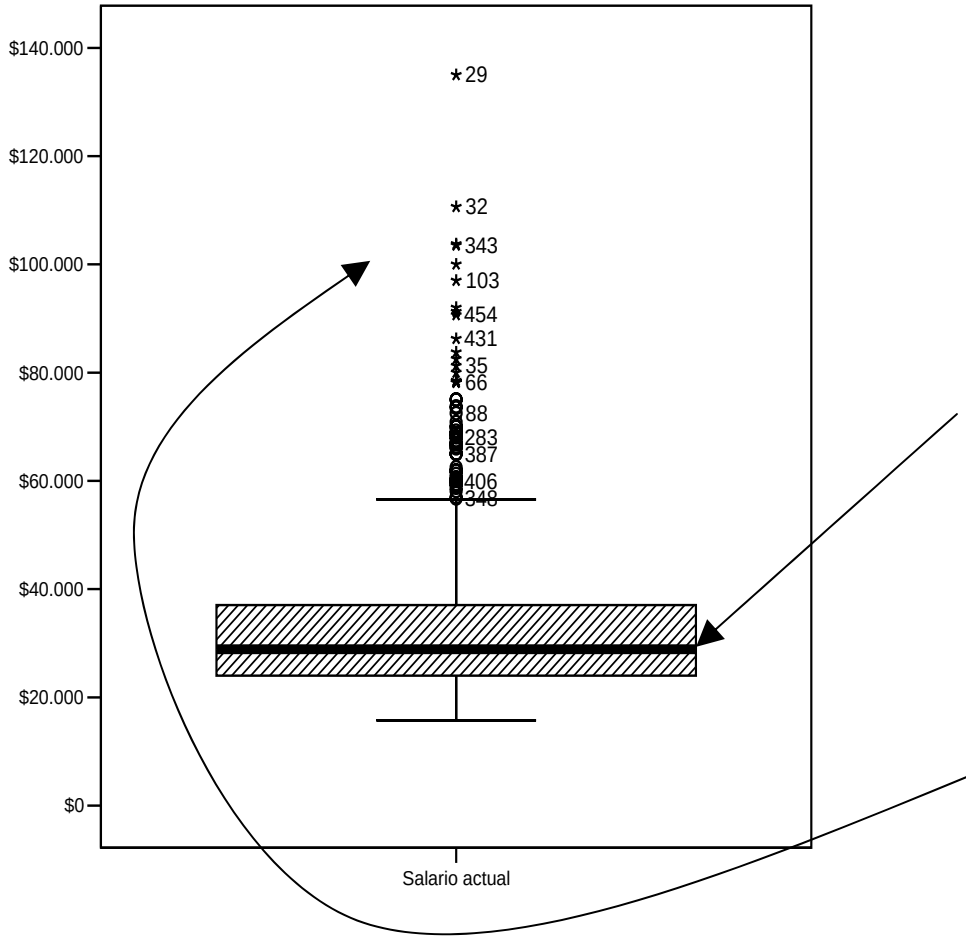
Salario actual

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

Una forma de representar gráficamente esta idea, es usando los *diagramas de cajas.* En las ordenadas se representan los valores de la variable y se muestran cuatro cajas correspondientes a las divisiones de los cuartiles. Para una sola variable no podemos seleccionar interactivo (inexplicablemente) sino que debemos irnos a:

⇨ Gráficos/Diagramas de caja/Simple/Resúmenes distintas variables/Definir

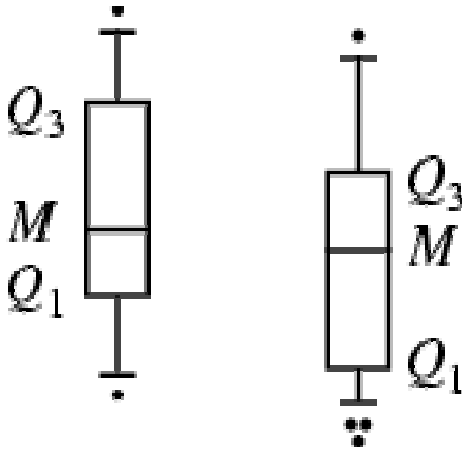
Seleccionad Salario Actual. El diagrama de caja contiene:



- 4 bloques correspondientes a los percentiles 25,50,75,100, es decir, los cuartiles). Se representan como una línea vertical, bloque, bloque y otra línea vertical.
- Mediana (la línea horizontal)
- Dispersión de los datos (cuanto más grande sean los bloques o las líneas verticales, mayor es la dispersión)
- Además también se ofrecen los casos aislados.

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).

Seleccionad "Etiquetar los casos mediante ... " y escoged la variable Sexo. Se aprecia que los valores atípicos corresponden a los hombres (y siempre en sueldos altos)



IQ = Q3 - Q1

P es un Outlier si *P > Q3 + 1.5 IQ*

P es un Outlier si *P < Q1 - 1.5 IQ*

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)

Si quisiéramos ver juntas todas las cajas correspondientes a varias variables, seleccionaríamos Diagrama de Cajas/Agrupado – Resúmenes para distintas variables.

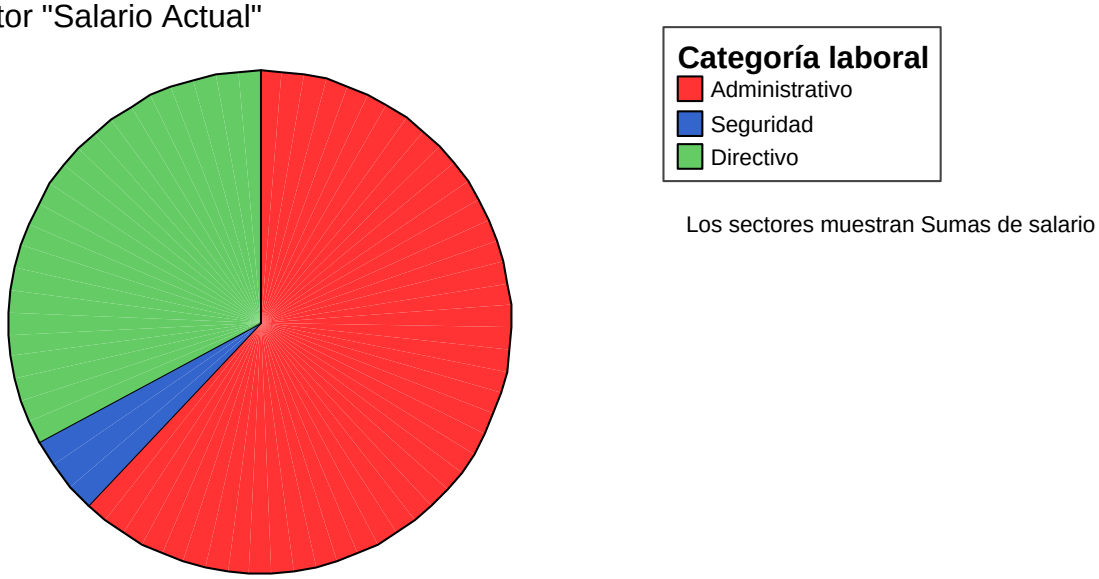
Si quisiéramos utilizar una variable de agrupación, usaríamos Resúmenes para grupos de casos. Por ejemplo, queremos ver la distribución del Salario para los Administrativos, de forma separada de los Directivos.

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:

⇨ Gráficos/Interactivos/Sectores/Simples. Sectores por "Categoría laboral" y Resumen del sector "Salario Actual"



5 Inferencia estadística

5.1 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**.

⇨ Abrid el fichero de Coches y calcular la media de la aceleración: 34419.56

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.**

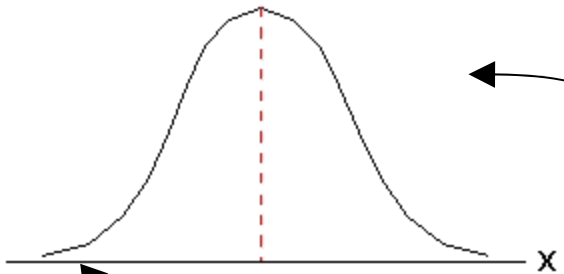
La inferencia estadística es el proceso de obtener conclusiones sobre una población a partir del análisis de una muestra. En la medida en que la muestra sea representativa de la población, los resultados podrán generalizarse.

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)

Por ejemplo, la famosa distribución normal para una variable numérica (de medida de escala)

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



The normal curve

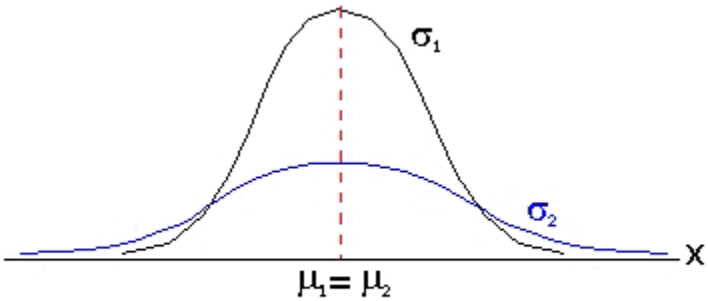
En el eje de las abscisas se representan los valores que puede tomar la variable (por ejemplo la edad)

En el eje de las ordenadas se representan los valores de probabilidad (entre 0 y 1). 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 doctorado.

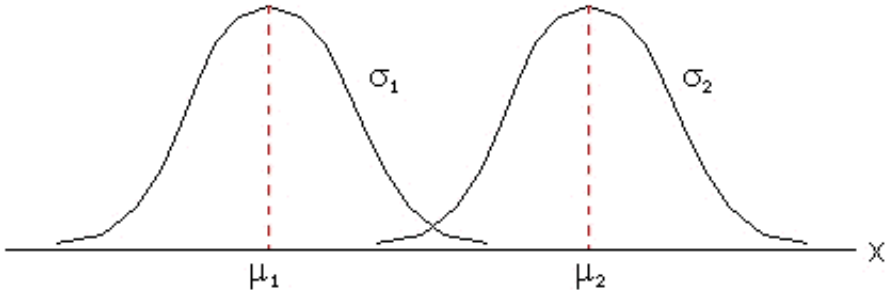
Dependiendo de lo queramos inferir tenemos distintos tipos de procedimientos estadísticos:

5.2 Estimación Puntual

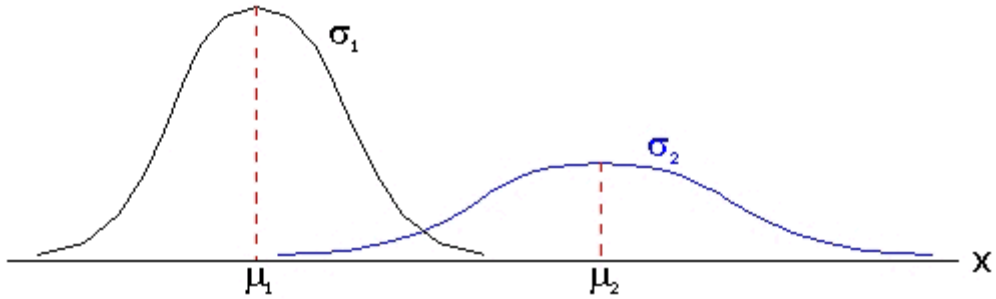
Podemos apreciar que la función de densidad de la normal depende únicamente de los valores que toman dos **parámetros**: μ y σ. Por ejemplo:



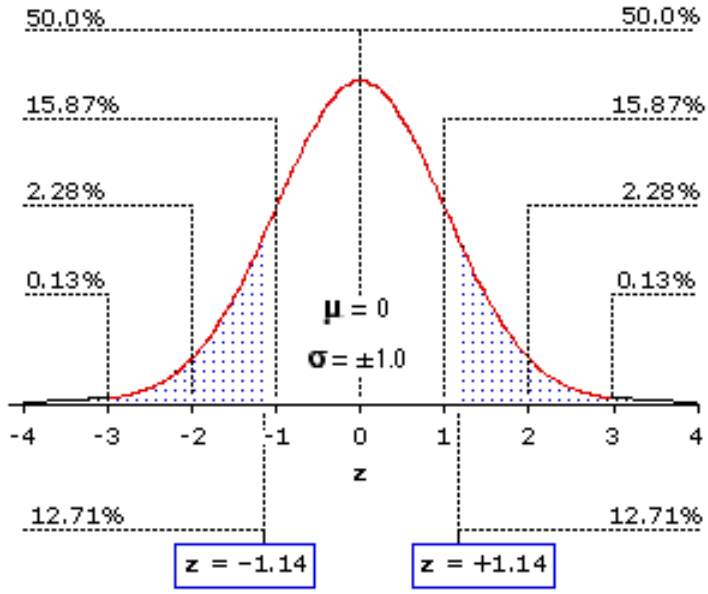
The normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$.



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

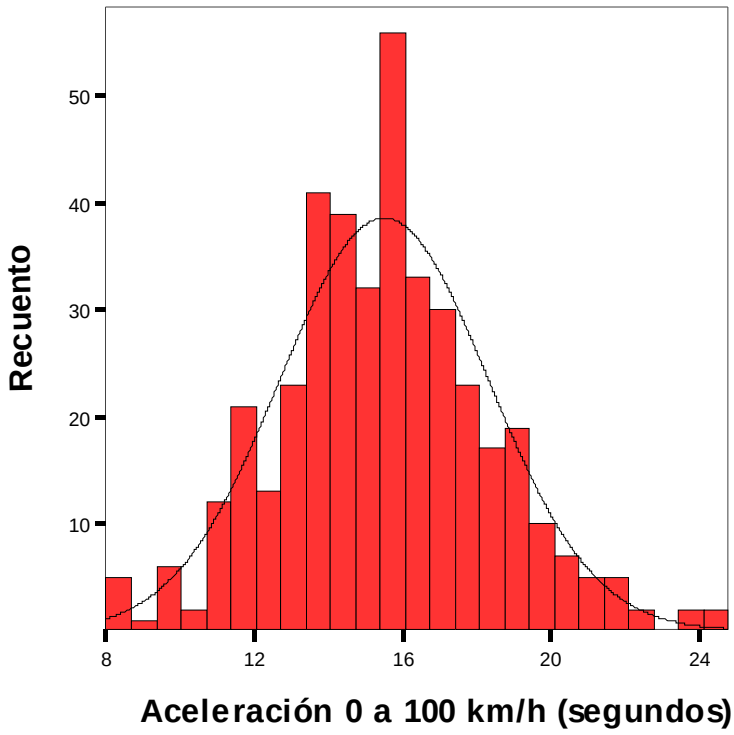


The normal curves with $\mu_1 \neq \mu_2$ and $\sigma_1 < \sigma_2$.



Conociendo dichos parámetros conoceríamos la función y tendríamos descrita la probabilidad con la que dicha variable puede tomar valores en el eje de las abscisas. Realmente, no llegaremos a *conocer* dichos parámetros, sino que los estimaremos con un estadístico. Por ejemplo, en el caso de la distribución normal, resulta que un buen estimador de μ es la media aritmética de la muestra y que un buen estimador de σ es la desviación típica de la muestra.

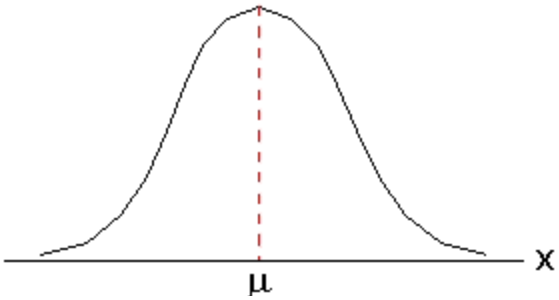
⇨ Abrid el fichero de Coches y ver el histograma de la aceleración.



Si los datos de la aceleración proviniesen, efectivamente de una distribución normal, entonces habrá muchos valores en torno a un valor medio (en este caso 16) y menos cuanto más lejanos estemos de dicho valor. Por lo tanto, el histograma se asemejaría a una función de densidad normal. Este es nuestro caso, con el ejemplo de la Aceleración. Así pues, si la muestra es representativa, podríamos decir que la población se distribuye según una normal. Por lo tanto podemos proceder a estimar la media y la desviación típica. Nos sale 15.66 y 2.8. Podemos estimar por tanto que la variable Aceleración se distribuye según una N(15.66, 2.8)

¿Cómo se formaliza matemáticamente este importante resultado? Veámoslo con la media. Es decir, vamos a ver por qué un buen estimador de μ es la media aritmética de la muestra

1. Suponemos que tenemos una población sobre la cual sabemos su distribución (su *forma*) pero no de forma exacta. Por ejemplo, sabemos que la forma es una Normal N(μ , σ) pero desconocemos los parámetros que la definen. Por ejemplo, si la población es la de los coches censados en Granada en los últimos 3 años, y estamos interesados en ver la aceleración de dichos coches, podríamos saber que la distribución subyacente es una normal, en el sentido de que la mayoría de las aceleraciones de los coches son cercanas a un valor central y que hay más o menos el mismo número (pequeño) de coches con grandes aceleraciones que con pequeñas aceleraciones.

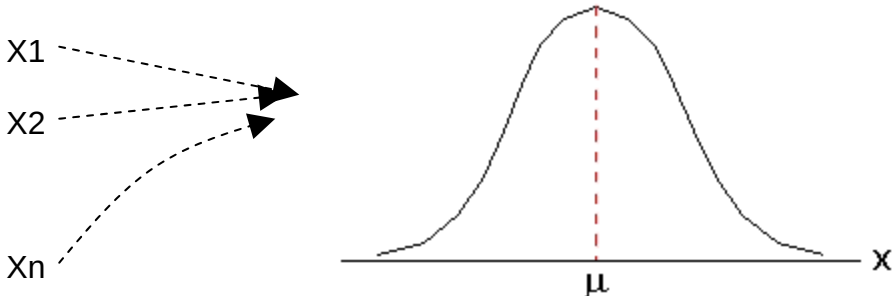


Aceleración de un coche censado en Granada en los últimos 3 años

2. Procedemos a seleccionar ***n*** individuos de una forma aleatoria (podrían ser todos los censados). En el ejemplo anterior, *n* coches del total de censados, y los numeramos del 1 al *n.*

   Xi = Aceleración del coche i (elegido al azar entre la población de coches en estudio)

   Antes de tomar los datos, no sabremos el valor concreto de Xi, pero sí sabremos que la probabilidad de que tenga, por ejemplo, una aceleración grande es baja. En general, podemos afirmar que Xi es una *variable aleatoria* que toma valores (de aceleración) según la distribución anterior:



   Podemos pensar que si hay n tuplas en la BD, y X es la columna de la Aceleración, entonces X1 .... Xn son n variables aleatorias idénticamente distribuidas (todas ellas tienen la misma distribución subyacente). Además, son independientes ya que la forma

de seleccionar los coches ha sido aleatoria y un valor de una no condiciona el valor de otra.

3. Se puede demostrar matemáticamente que si construimos el siguiente estadístico (que es una variable aleatoria)
   (X1+ .... + Xn)/n
   entonces, dicha variable aleatoria tiene, exactamente, la función de densidad de una
   $N(\mu , \sigma/\sqrt{n})$
   (esto es porque en el punto 1 habíamos supuesto que los Xi eran $N(\mu , \sigma)$)
   Esto significa que la variable aleatoria (X1+ .... + Xn)/n tomará valores en torno a μ, y que es poco probable que se aleje demasiado de μ. Si el valor de *n* fuese muy grande (una muestra muy grande), la desviación se hace casi cero, y por lo tanto más difícil será que el valor del estadístico esté lejos de su valor central μ,

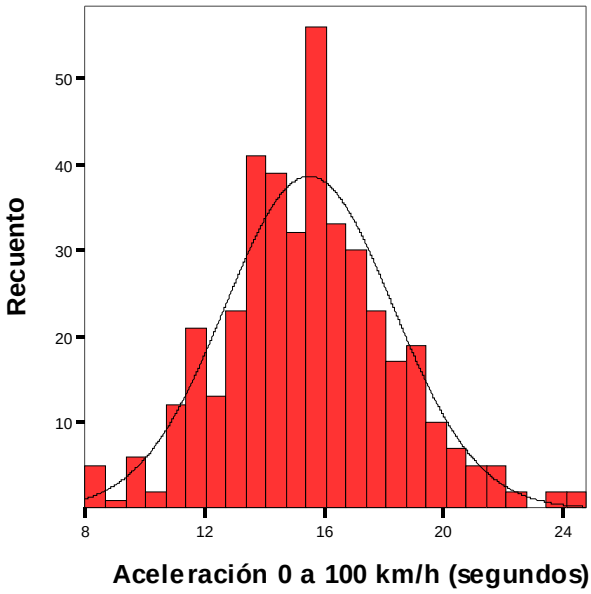
4. Si nos fijamos ahora en los valores concretos observados en la base de datos, el valor que la variable Xi toma en la tupla i, lo llamaremos **Xi**

Consumo	Aceleración	Potencia	Año	.....
	X1 = 16 **X2 = 14** **X3 = 14**			

   Según lo visto en el punto 3, estadísticamente hablando, lo usual será que si calculamos el número (**X1**+ .... + **Xn**)/n, este valor estará cerca del desconocido μ. En la tabla anterior, sería 14.66. Por eso es por lo que estimamos μ con el valor 14.66

Así pues, una forma de estimar un parámetro consiste en encontrar un estadístico (construido con los valores que toma la muestra,) tal que su función de densidad tenga como esperanza (valor medio) el parámetro que queremos estimar. (Esta es sólo una forma de hacerlo y dichos estadísticos se dice que son *insesgados*).

Observemos que en el punto 1 dijimos que el experto supone que dicha densidad común es una específica, por ejemplo, una normal N(μ , σ) de parámetros desconocidos. Esta suposición puede hacerla simplemente viendo el histograma:



¿Pero y si la población subyacente no sigue una distribución normal?

Caso 1

- El experto nos debería decir la distribución en concreto que él cree que sigue la población.
- Cada función de densidad vendrá determinada por ciertos parámetros. Veríamos los parámetros que determinan la función de densidad, y veríamos qué estadísticos son buenos estimadores de dichos parámetros, y veríamos cual es la función de densidad de dicho estadístico (punto 3). Hay cientos de artículos que tratan cientos de distribuciones posibles. Sin embargo, SPSS sólo ofrece los estadísticos descriptivos típicos que ya vimos anteriormente.

Caso 2

- Se recurren a técnicas estadísticas “no paramétricas”.

⇨ Abrid de nuevo los datos de empleados. Comprobad que el histograma del salario actual no se acerca mucho a una normal. Menos aún los "Meses desde el contrato". Por lo tanto, no podemos hacer la misma estimación que hicimos con la Aceleración.

¡Nuestro gozo en un pozo!

Sin embargo, no todo está perdido. Aunque para distribuciones no normales, no podemos hacer una estimación tan poderosa (de los parámetros μ y σ que determinan la función de densidad), sí que podemos estimar otras cosas interesantes. Por ejemplo, si X es el nombre de la variable en cuestión y f(x) es la función de densidad de X, tenemos:

- El número calculado en la forma $\int x\ f(x)\ dx$ es conocido como la Esperanza de X, y se denota por E(X). Es la conocida medida de tendencia central que mide "el valor medio" de la población. En la normal, la esperanza es μ
- El número calculado como la raíz cuadrada positiva de $\int (x-E(X))^2\ f(x)\ dx$ es conocido como la Desviación Típica de X. Es la conocida medida de dispersión que mide cómo de dispersos están los valores de la población. En la normal, la desviación típica es σ

Así pues sería deseable estimar estos valores, porque nos dan bastante información de la población (aunque no **toda** la información como en el caso de la normal).

Hay un resultado importantísimo y es que la media y la desviación típica muestral, son unos buenos estimadores de la esperanza y la desviación típica de la población, sea cual sea la distribución subyacente de la población. Así pues, aunque no estemos estimando cual es la densidad de la población, al menos estamos estimando características importantes suyas.

¿Cómo se formaliza matemáticamente este importante resultado? Es el famoso teorema central del límite. Lo vemos para la media:

> Dadas n v.a.i.i.d (sea cual sea su función de densidad), la v.a. (X1+ .... + Xn)/n
> se distribuye aproximadamente según una Normal(μ , $\sigma/\sqrt{n}$)

Es más, se puede demostrar también que:

> Dadas n v.a.i.i.d (sea cual sea su función de densidad), la v.a. (X1+ .... + Xn)/n
> se distribuye aproximadamente según una Normal(μ , $S/\sqrt{n}$)

Claro está la primera aproximación es *más fina* pero requiere el conocimiento de σ, lo que en la realidad es inviable.

⇨ Abrid de nuevo los datos de empleados. Calculad la media aritmética del Salario Actual (recordad que el Salario no se distribuía según una normal). Debe salir 34419.56. Podemos afirmar, según el TCL, que la esperanza de la distribución subyacente del Salario Actual estará cercana a 34419.56

Finalmente, destaquemos que lo mismo que hemos hecho con la media se podría hacer con otros parámetros de la población, como por ejemplo la desviación típica.

En resumen, una forma de estimar un parámetro consiste, como ya dijimos, en encontrar un estadístico (construido con los valores que toma la muestra,) tal que su función de densidad tenga como esperanza (valor medio) el parámetro que queremos estimar. (Esta es sólo una forma de hacerlo y dichos estadísticos se dice que son *insesgados*). Dicha función de densidad podrá ser exacta si se conoce la distribución de los Xi o aproximada (TCL) en caso contrario.

5.3 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. Intuitivamente, cuanto mayor sea la dispersión de los datos, 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.

5.4 Contraste o Test de Hipótesis

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 (o *aceptar*) 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. μ =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

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

H_0. μ =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 pequeño

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

valor muy alto > 0.05

¿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. μ =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(X1, ... , Xn)
  En nuestro ejemplo, utilizamos el estadístico media muestral:
  T(X1, ... , Xn) = (X1+ .... + Xn)/n = $\overline{X}$
- Se calcula la distribución de dicho estadístico suponiendo que H_0 es cierta. Esto ha de hacerse de forma que en dicha distribución no aparezcan parámetros desconocidos (este es el pilar matemático). Llamamos a dicha función f_0.
  Por ejemplo, si la muestra corresponde a una N(μ , σ), recordemos que habíamos dicho que la media muestral se distribuía según una N(μ , $\sigma/\sqrt{n}$) , o lo que es lo mismo, $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ se distribuye según una N(0,1). Pero cuando H_0 es cierta, en nuestro caso, μ

=34000, resulta que el estadístico se nos queda $\frac{\sqrt{n}(\overline{X}-34000)}{\sigma}$ y σ es desconocida, por lo que no podemos calcular dicho valor.

Pero se puede demostrar que $\frac{\sqrt{n}(\overline{X}-\mu)}{S}$ 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, en nuestro caso, μ =34000, resulta que el estadístico se nos queda $\frac{\sqrt{n}(\overline{X}-34000)}{S}$ 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.

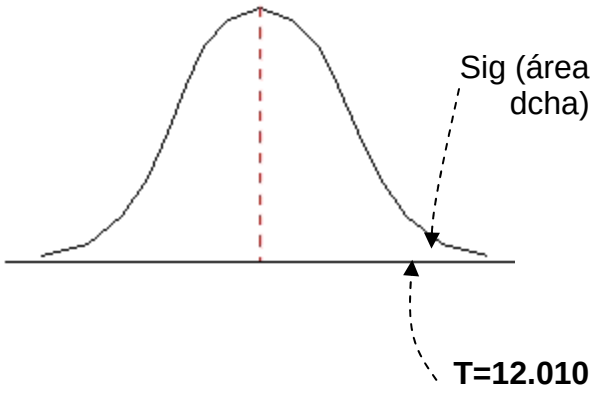
- Se calcula el valor numérico concreto que T toma en la muestra **X**1, ... , **X**n . Lo llamamos **T.** (es el valor llamado t en SPSS)

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

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

- Ahora, basta comprobar si dicho valor es poco probable que se de en la distribución definida por f_0 (en la práctica se consideran valores por debajo de 0.05) Si es así, 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. La Sig. es el área que deja a la derecha **T** (o a la izda). 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)

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. En muchas ocasiones se recurren a aproximaciones derivadas del TCL.
Nota. Además, también existen contrastes para otro tipo de hipótesis. Por ejemplo, podemos establecer un contraste para comprobar estadísticamente si una muestra se ajusta a una distribución determinada. En SPSS sólo puede hacerse para unas poquitas (en gráficos P-P pueden encontrarse otras distribuciones).

H_0. La muestra proviene de una distribución Normal
H_1. La muestra no proviene de una distribución Normal

⇨ Analizar/Pruebas no Paramétricas/Test de K-S de normalidad / Normal

Prueba de Kolmogorov-Smirnov para una muestra

		Salario actual
N		474
Parámetros normales [a,b]	Media	$34,419.57
	Desviación típica	$17,075.662
Diferencias más extremas	Absoluta	,208
	Positiva	,208
	Negativa	-,143
Z de Kolmogorov-Smirnov		4,525
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.

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

Ejercicio: Plantead 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

6 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?

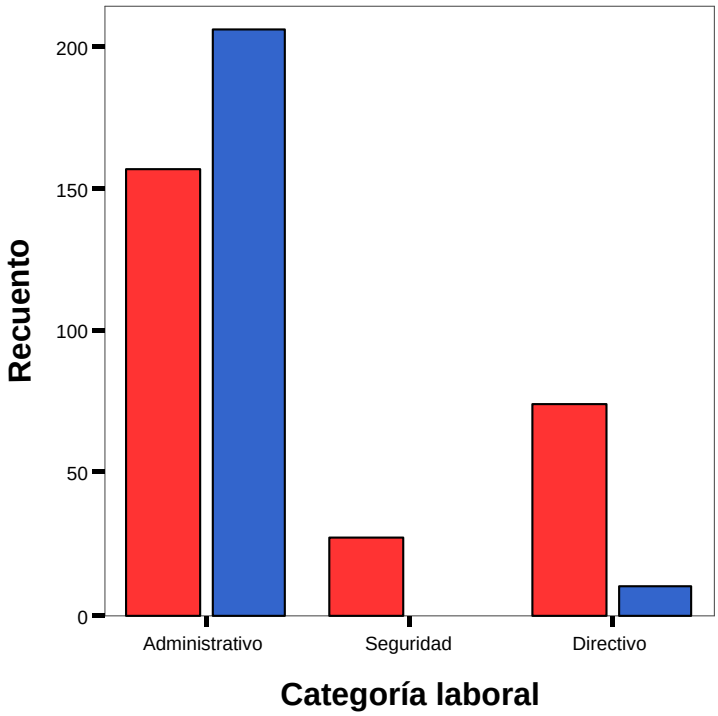
O más específicamente:

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

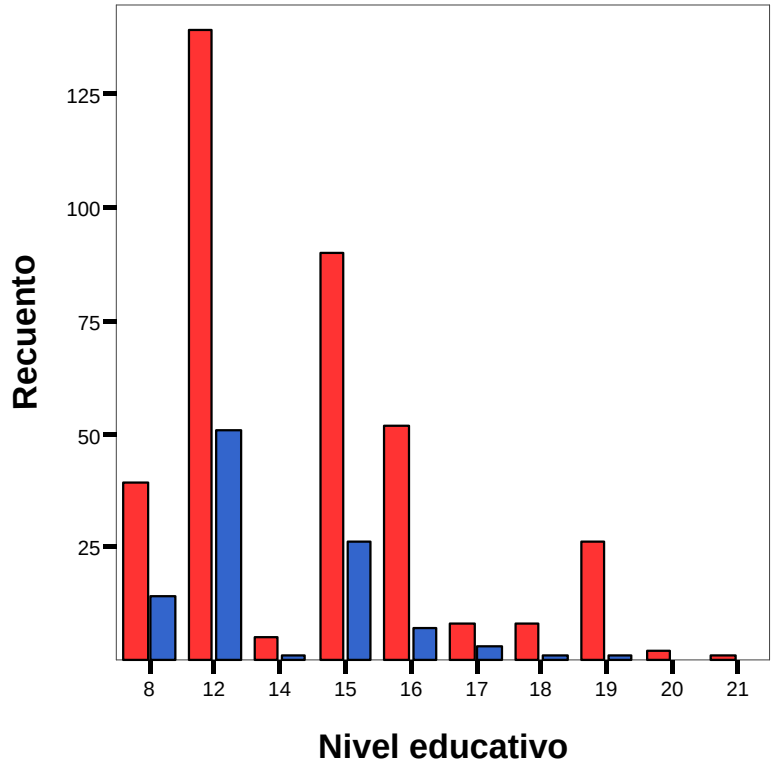
o dicho al revés,

¿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.



Las barras muestran frecuencias



Las barras muestran frecuencias

7 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 ~ 0.43
Proporción de hombres en los Directivos: 74 / 84 ~ 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:

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)

Si lo que vamos buscando es demostrar que existe dependencia, este es el tipo de test que teníamos que hacer, ya que, como ya dijimos, lo mejor desde un punto de vista estadístico, es poder rechazar una hipótesis H_0 (en este caso, rechazar independencia).

Supongamos una tabla de contingencia de dos entradas con n filas y m columnas. Si llamamos Oij 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

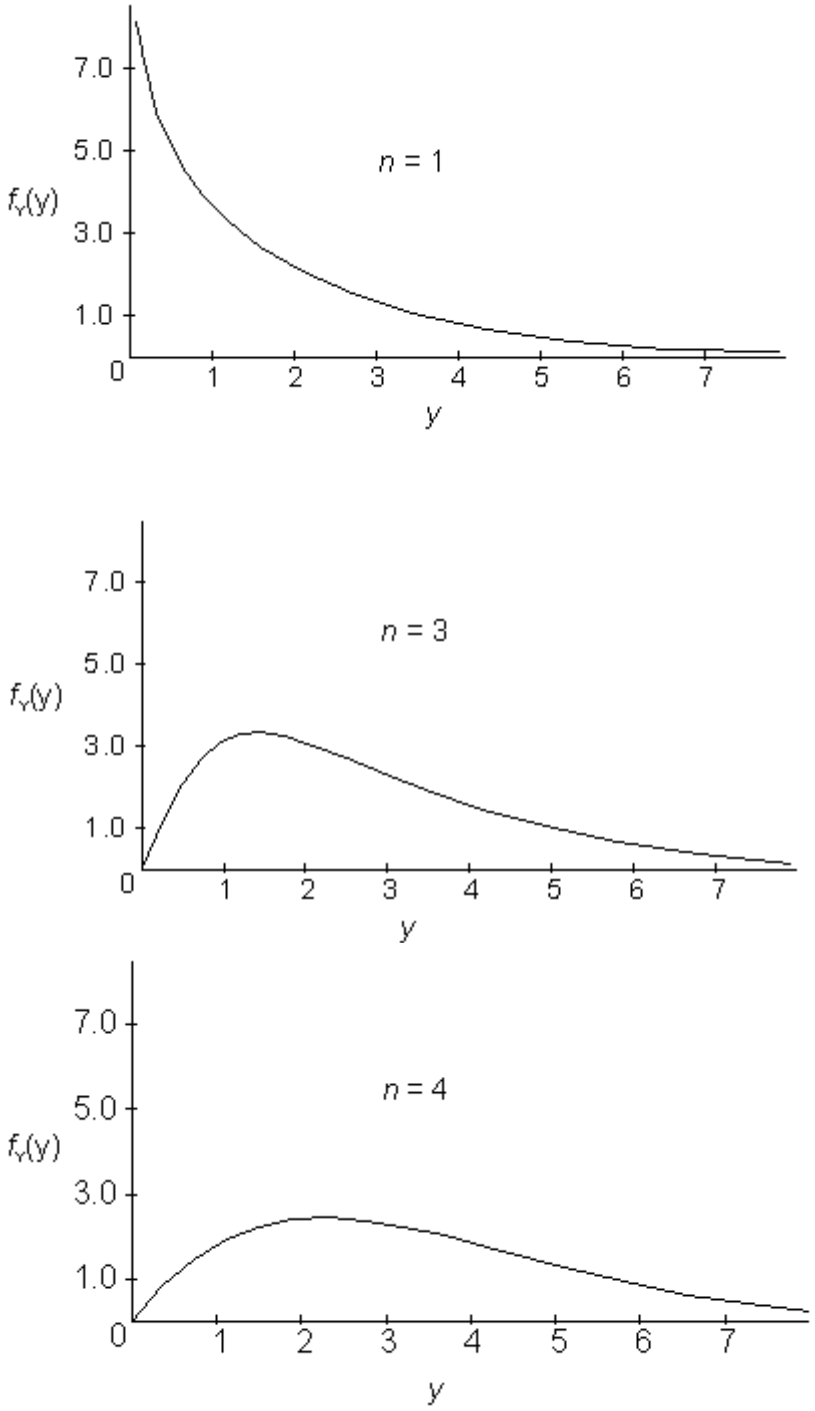
m_j es el número total de elementos de la columna j

Entonces, se construye el estadístico siguiente:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$$

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.

Por curiosidad, la distribución chi cuadrado tiene la siguiente forma (para distintos valores de n, que representa los grados de libertad)



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

- Ninguna *Eij* es inferior a 1

- No mas del 20% de las *Eij* son inferiores o iguales a 5

⇨ Analizar/EstadísticosDescriptivos/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[a]	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 para una interpretación de dichas medidas.
2. Analizar cuales son las casillas que más contribuyen a rechazar la independencia. SPSS no proporciona dichas medidas. En el curso 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.

8 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:

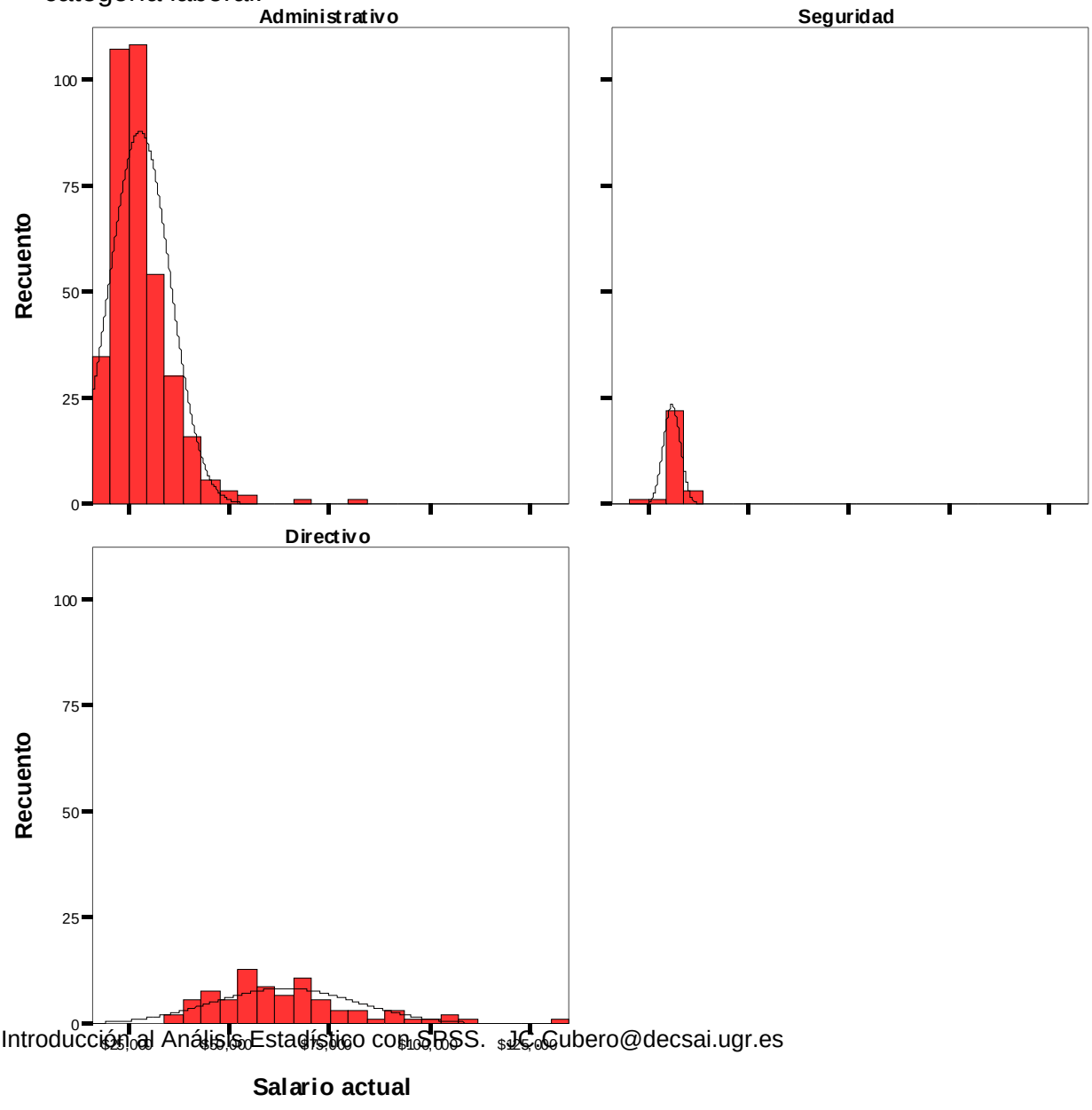
¿Cómo se distribuye el salario de los empleados en función de su categoría laboral?

En general:

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

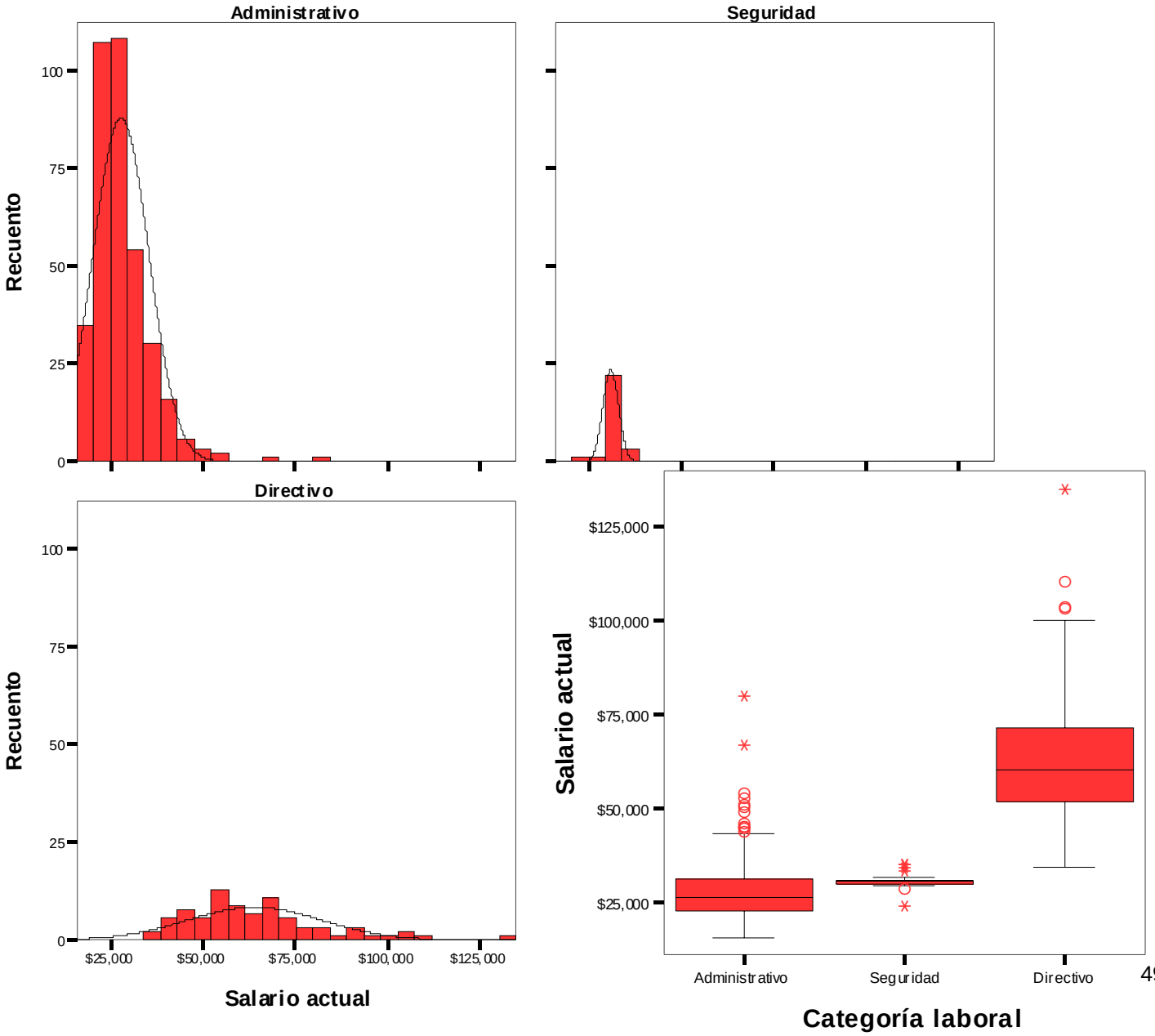
Como la variable dependiente (consecuente) es numérica, usamos un histograma.

⇨ Gráficos/Interactivos/Histogramas/ Seleccionamos Sal Actual y agrupamos por categoría laboral.



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/Interactivos/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". Incluid 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

⇨ 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:

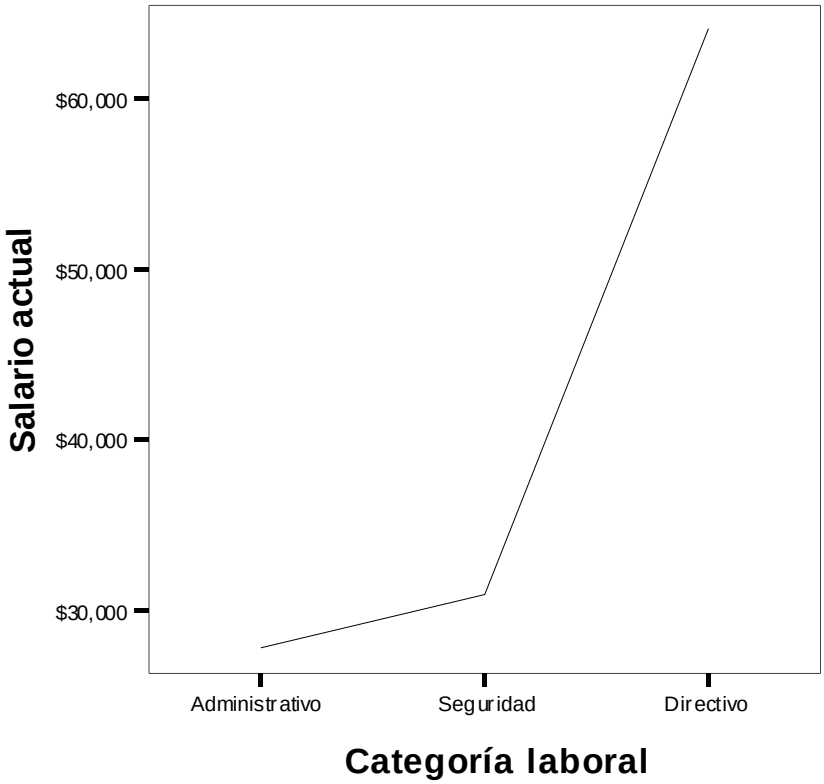
¿Depende el salario de los empleados de su categoría laboral?

es decir,

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

Para ello, vamos a ver si 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/Interactivos/Líneas/ Sustituid Recuento por Salario Actual y que las líneas representen medias. Seleccionad Categoría laboral en el antecedente.



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

⇨ Incluid la variable “Sexo” como variable de leyenda (para que aparezcan juntas las cajas) o como variable de panel (para que aparezcan separadas las cajas)

9 Análisis Estadísticos de dependencia. Numérica-Nominal

Vamos a cuantificar con métodos estadísticos, la dependencia entre "Salario" y "Categoría laboral" que se ha vislumbrado en el AED anterior. Siempre que tengamos una variable dependiente de escala y variables independientes nominales, usaremos un ANOVA. Por cada valor que tome la(s) variable(s) independiente(s) (o factor) se forma un *grupo.* El ANOVA plantea el siguiente test de hipótesis:

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

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.

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 $g =$ 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:

$$\sum_g \sum_i \left(x_{gi} - \bar{x}\right)^2 = \sum_g n_g \left(\bar{x}_g - \bar{x}\right)^2 + \sum_g \sum_i \left(x_{gi} - \bar{x}_g\right)^2$$

$$SS_T = SS_B + SS_W$$

Dónde *B* representa “Between groups” (inter-grupos), aunque estaría mejor dicho “Among groups”, es decir, la sumas 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 ($df_B =$ número de grupos -1 para SS_B y $df_W =$ número de valores – número de grupos para SS_W) se construye el estadístico *F*:

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

Si las medias fuesen distintas entre los grupos, tendríamos que (la desviación medida por) SS_B sería más grande que SS_W, es decir, que las sumas de cuadrados de desviaciones entre grupos sería mayor que las sumas de cuadrados de desviaciones intra-grupos. Es decir, el estadístico *F* sería mayor que 1.

Se puede demostrar que cuando la hipótesis nula de igualdad de medias es cierta, 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.

⇨ 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			

En cualquier caso, hay una gran parte de variabilidad que no queda explicada únicamente por la componente inter-grupos. Así pues, aunque las medias de Salario Actual son distintas entre las categorías laborales, hay otros factores que también intervienen a la hora de determinar el salario

Lamentablemente, desde este menú no se pueden elegir variables nominales como factores, por lo que debemos irnos a

Analizar/Modelo Lineal General/Univariante -> Factores Fijos

Ejercicio: ¿Hay diferencia de salarios atendiendo al sexo?

Podemos analizar qué grupos concretos son los que presentan unas mayores diferencias. Para ello, volvemos a hacer un ANOVA, pero en Post Hoc seleccionamos, por ejemplo, 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[a,b]

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.

a. Usa el tamaño muestral de la media armónica = 58,031.

b. 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.

Como siempre, desde SPSS se puede hacer lo mismo desde otros menús. Por ejemplo, podemos seleccionar Comparar Medias / Medias cuando sólo tengamos una variable independiente (si incluimos varias, se hace un estudio de un ANOVA de la v. dependiente cruzada de forma independiente con cada una de las independientes). O por ejemplo, Comparar Medias / Muestras Independientes, en el caso de que queramos fijar valores específicos de la v. independiente.

¿Y qué pasa cuando tenemos varias variables independientes? En el caso de que queramos agrupar considerando varias variables independientes, debemos seleccionar:
Analizar/Modelo Lineal General/Univariante/

Cargad la base de datos que se encuentra en Tutorial/Sample_files 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.

⇨ Analizar/Modelo Lineal General/Univariante/

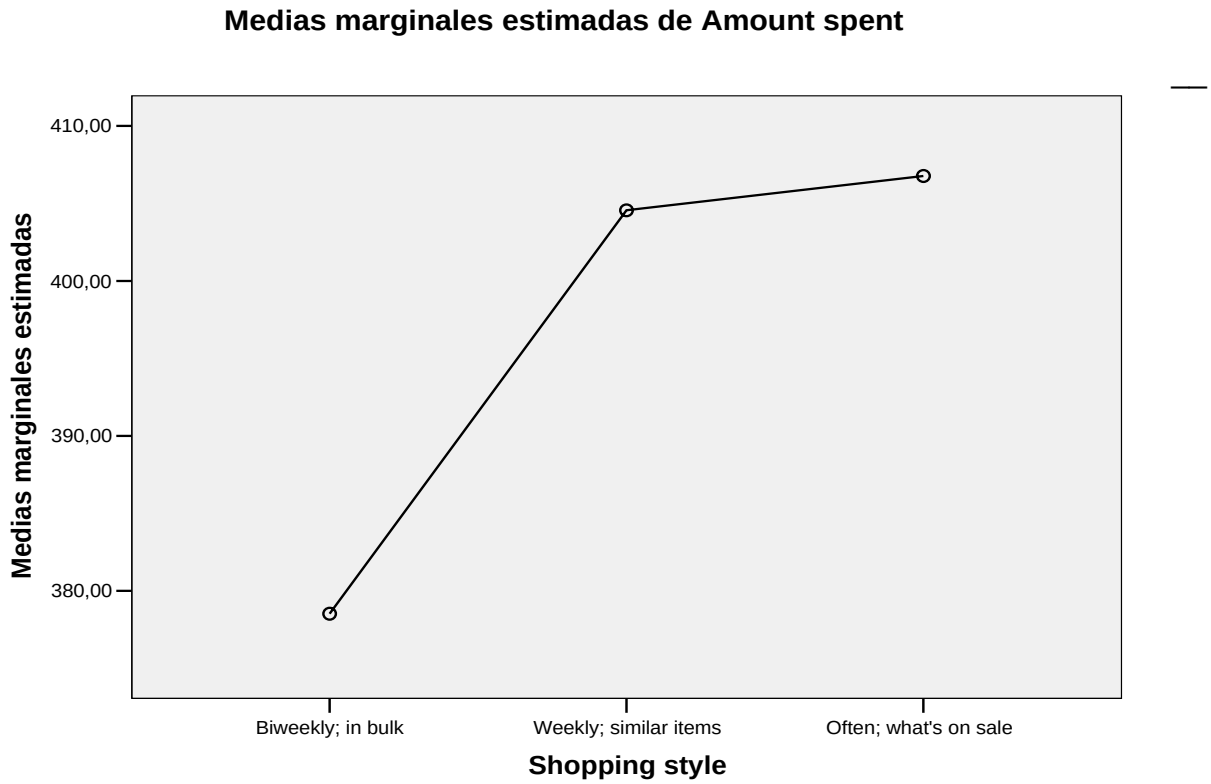
Seleccionad Amount Spent (amtspent) como variable dependiente y Shopping Style como factor fijo. En Gráficos, seleccionad la variable independiente (Shopping Style). Post-Hoc para style

Comparaciones múltiples

Variable dependiente: Amount spent
DHS de Tukey

(I) Shopping style	(J) Shopping style	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Biweekly; in bulk	Weekly; similar items	-26,0342	12,61108	,099	-55,7191	3,6507
	Often; what's on sale	-28,2471	16,25946	,193	-66,5198	10,0256
Weekly; similar items	Biweekly; in bulk	26,0342	12,61108	,099	-3,6507	55,7191
	Often; what's on sale	-2,2130	13,47525	,985	-33,9320	29,5061
Often; what's on sale	Biweekly; in bulk	28,2471	16,25946	,193	-10,0256	66,5198
	Weekly; similar items	2,2130	13,47525	,985	-29,5061	33,9320

Basado en las medias observadas.

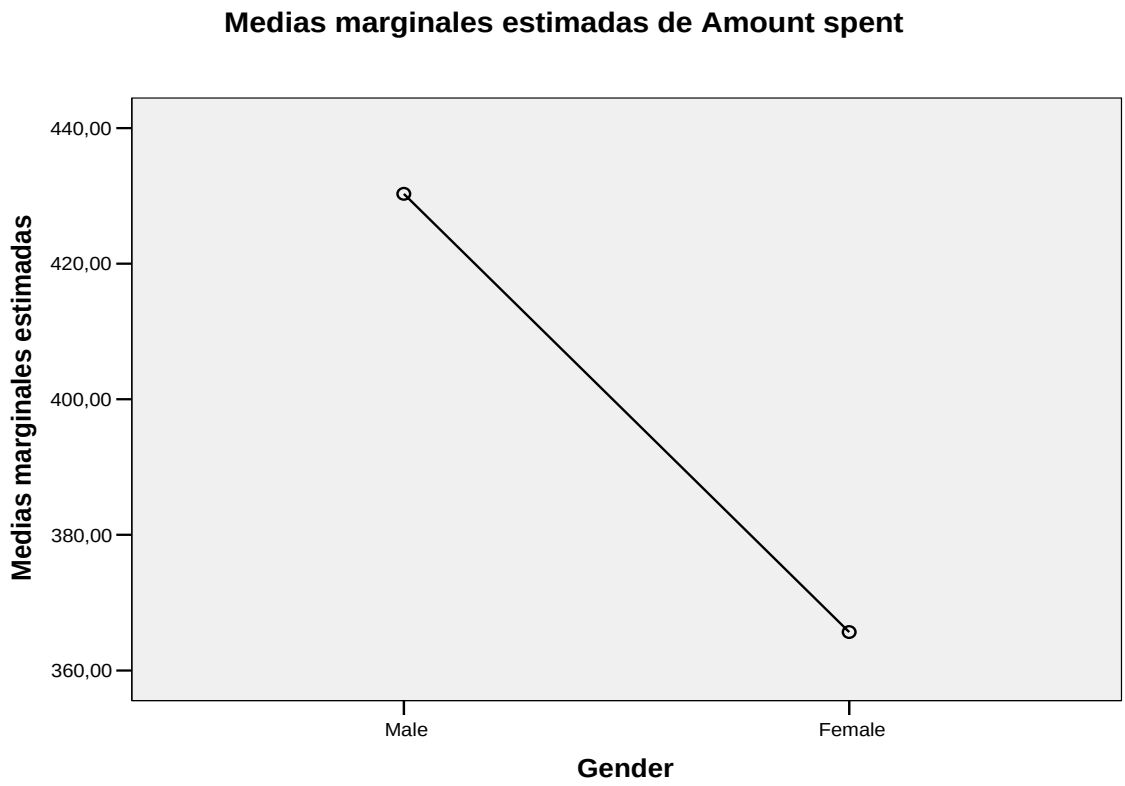


En el ANOVA resultante, no es significativa la diferencia entre las medias. Estos mismos resultados pueden obtenerse desde Analizar/Comparar Medias/Anova de un factor.

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.

—



Ahora, las diferencias sí son significativas.

Sin embargo, queremos ver si hay alguna interacción entre Shoppying Style y Gender.

Seleccionamos la misma dependiente y ambas como factores fijos.

En Gráficos, style en el Eje Horizontal y gender en "líneas distintas". Añadimos el gráfico style*gender.

Opciones. Mostrar medias para todas.

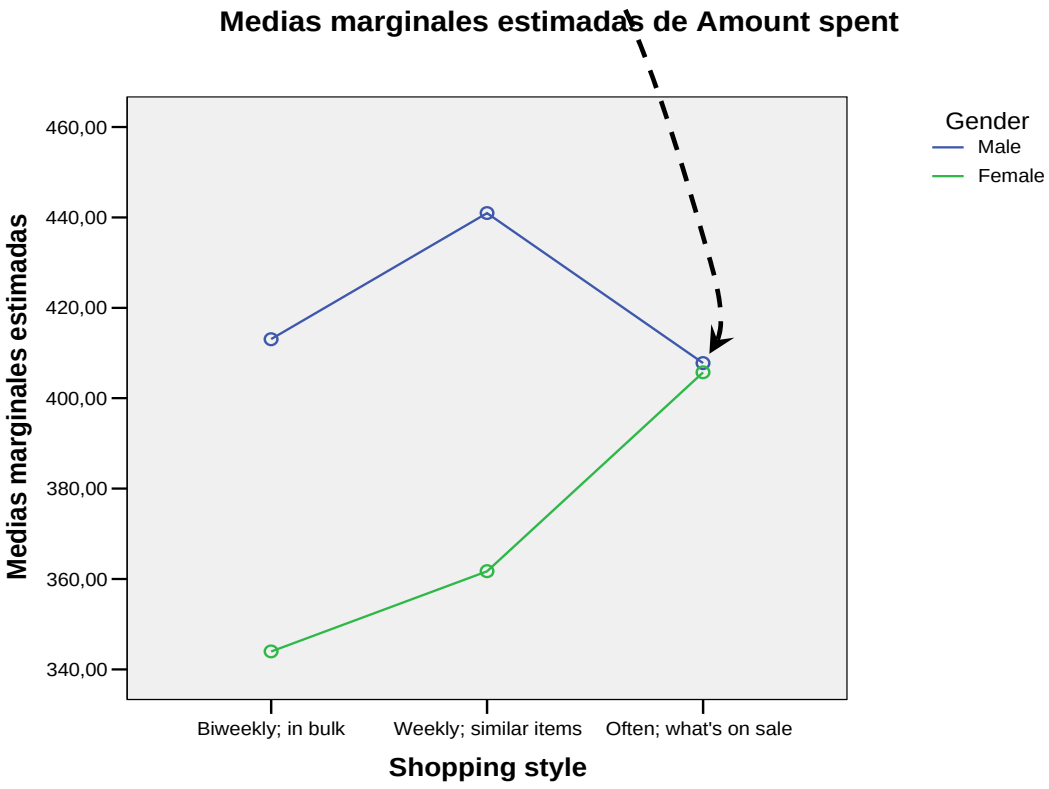
Pruebas de los efectos inter-sujetos

Variable dependiente: Amount spent

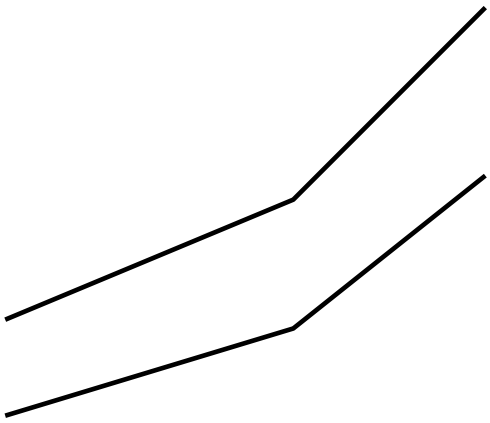
Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	469402,996[a]	5	93880,599	11,092	,000
Intersección	39359636,4	1	39359636	4650,274	,000
gender	158037,442	1	158037,442	18,672	,000
style	33506,210	2	16753,105	1,979	,140
gender * style	69858,325	2	34929,163	4,127	,017
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?

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

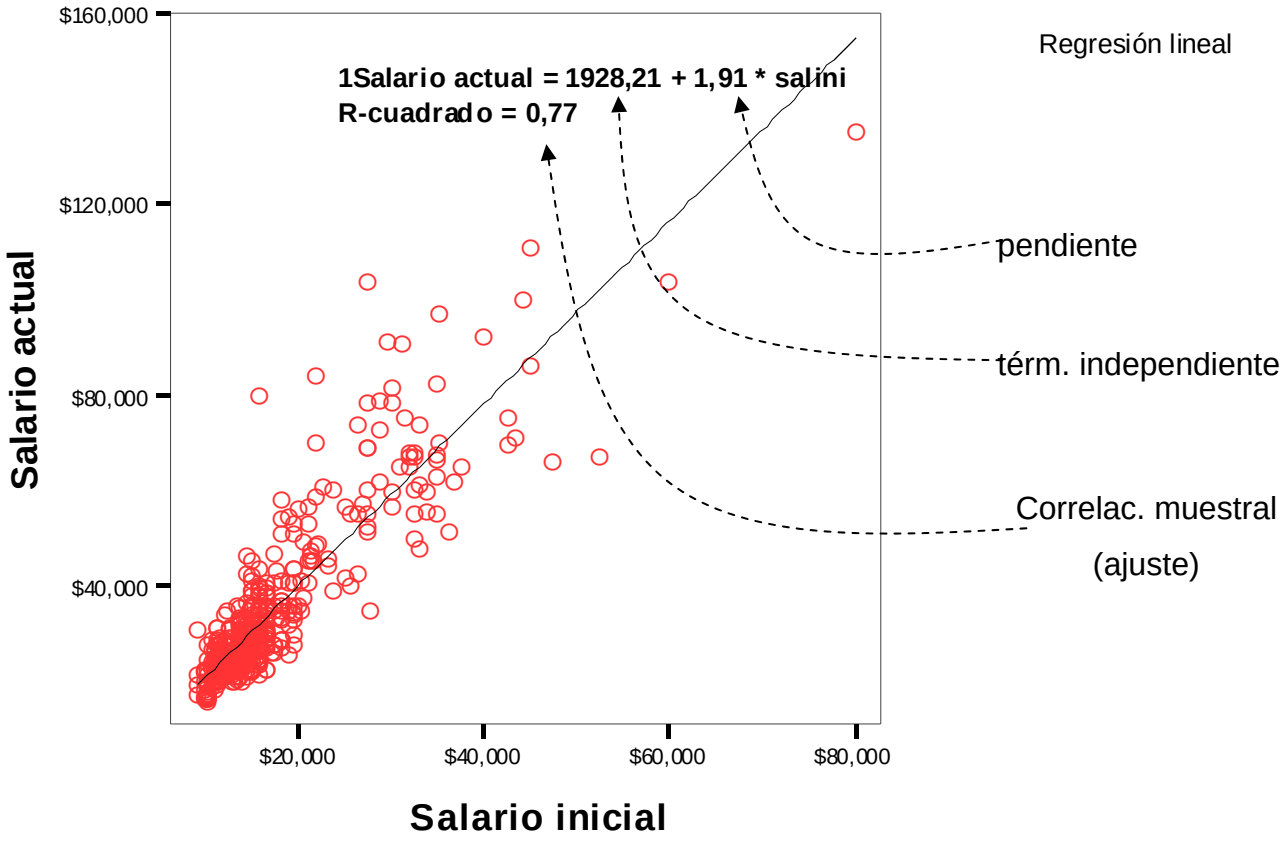
Suponemos que trabajamos con variables numéricas de escala y estamos interesados en ver si unas dependen de otras. Queremos responder a preguntas del tipo:

¿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.

⇨ Gráficos Interactivos/Dispersión. Seleccionad "Salario actual" como variable dependiente y "Salario Inicial" como variable independiente. En la pestaña Ajuste seleccionar Método: Regresión (e incluir cte en la ecuación)



Efectivamente, parece que el Salario actual depende del salario inicial.
R-cuadrado es el coeficiente de correlación muestral. Mide el ajuste de la recta de regresión a la nube de puntos.

Importante: Esta es la típica regresión que ajusta una recta. Pero podría haber otra dependencia explicada por otro tipo de ecuación (En SPSS se encuentra en el menú de Regresión)

11 Análisis Estadísticos de dependencia. Numérica-Numérica. Regresión.

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

⇨ 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[b]

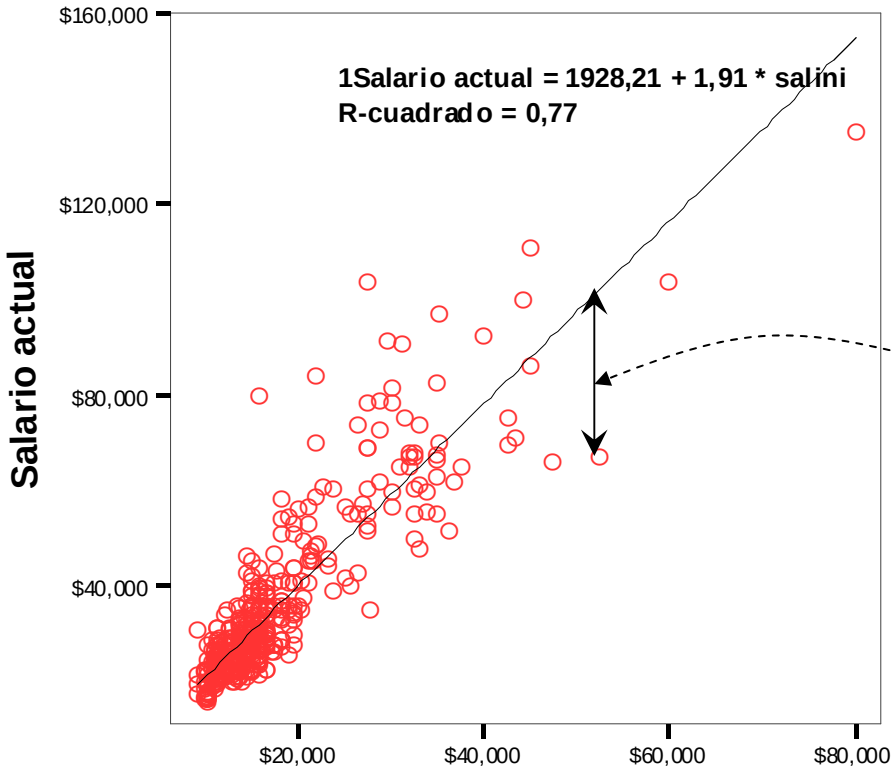
Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,068E+11	1	1,07E+11	1622,118	,000[a]
	Residual	3,109E+10	472	65858997		
	Total	1,379E+11	473			

a. Variables predictoras: (Constante), Salario inicial

b. Variable dependiente: Salario actual

Sin embargo, la SC Residual no es despreciable en comparación con SC Regresión. Esto nos dice que Sal Inic no explica por completo el comportamiento de Sal Act.

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



SC Residual representa la suma de las distancias (entre el valor real y el pronosticado con la recta) al cuadrado

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

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	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)

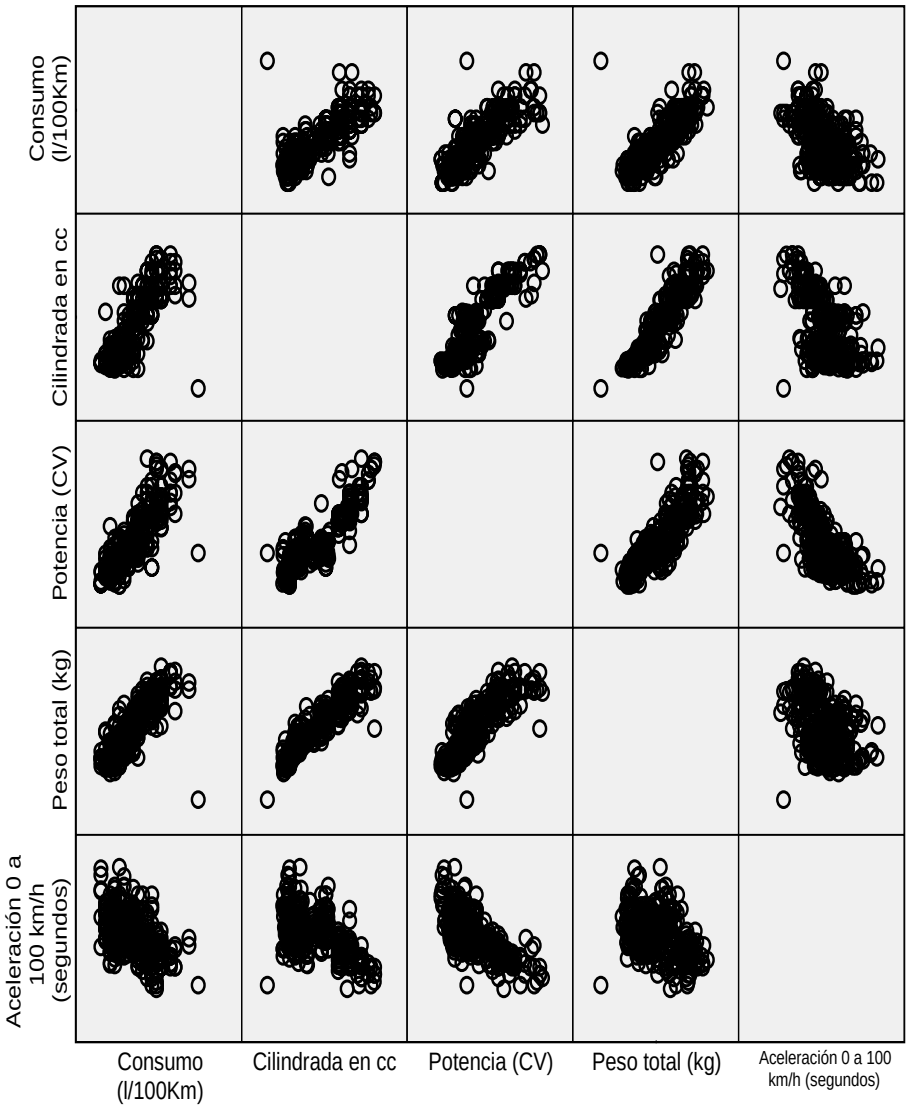
Valor muy significativo. La pendiente de la recta (1.909) es bastante fiable. El IC es pequeño.

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 doches

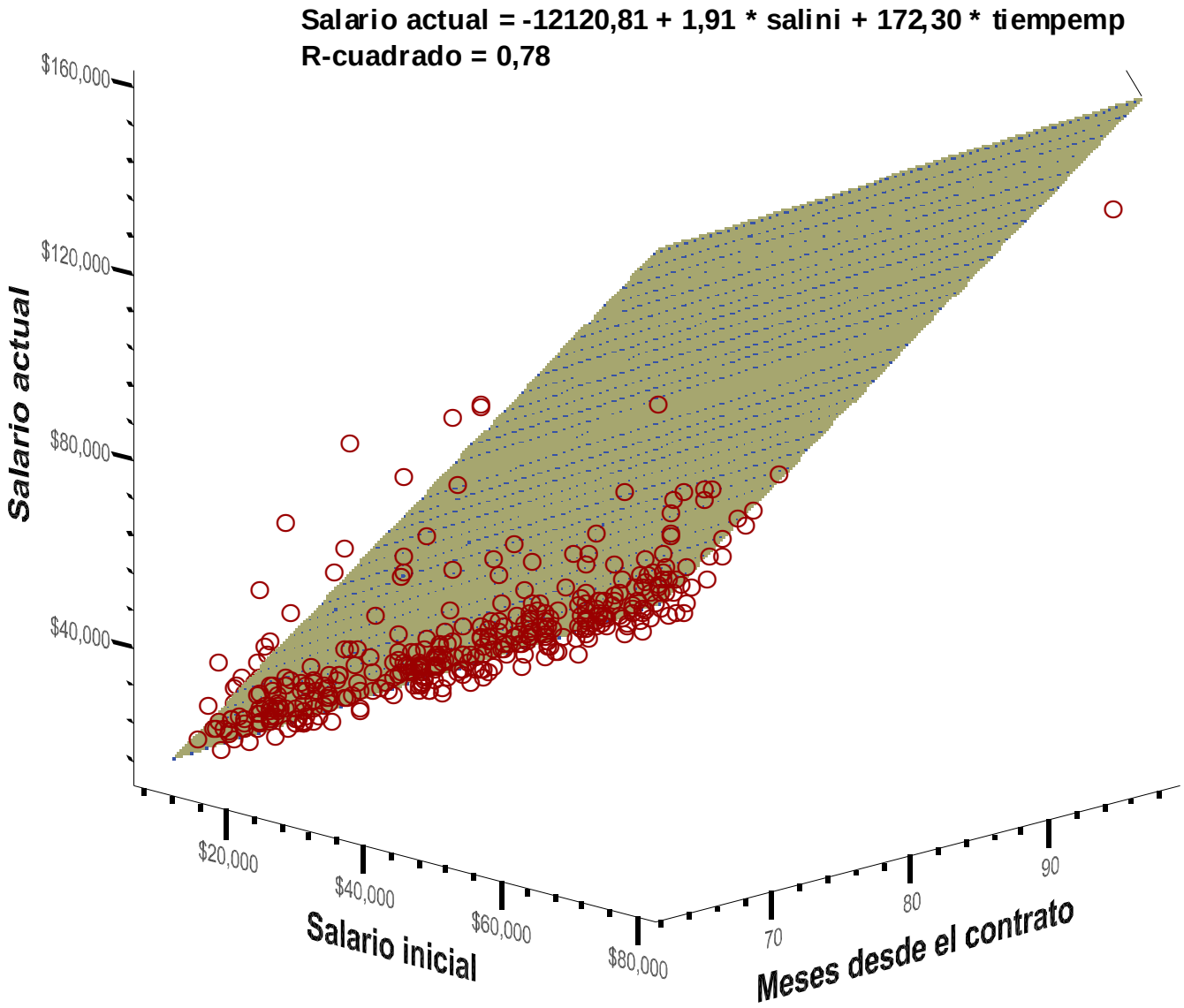
⇨ Gráficos/Dispersión/Dispersión Matricial. Seleccionad el consumo, cilindrada, potencia, peso, aceleración.



¿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[a]

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[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,083E+11	2	5,41E+10	859,383	,000[a]
	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

Sin embargo, hay bastante variabilidad no explicada (SC residual alta)

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.

Casi todos los modelos vistos anteriormente exigen que las variables numéricas tengan una distribución normal y con igualdad de varianzas. Muchas veces estos requisitos no se cumplen, por lo que es conveniente recurrir a métodos no paramétricos, que no exigen dichas restricciones.

Por ejemplo, para un ANOVA, seleccionaríamos Pruebas no paramétricas -> k muestras independientes