CURSOFUNDAMENTOS ESTADÍSTICOS PARA INVESTIGACIÓN. (NIVEL INICIAL)

Con SPPS , EXCEL y R

PLAN DE FORMACIÓN DEL PERSONAL DE ADMINISTRACIÓN Y SERVICIOS
Universidad de Murcia.

Antonio Maurandi López (amaurandi@um.es)
Sección de Cálculo Científico y Apoyo Estadístico.
Servicio de Apoyo a la Investigación (SAI)
www.um.es/sai www.um.es/ae

Índice de contenidos:

CURSO
FUNDAMENTOS ESTADÍSTICOS PARA
INVESTIGACIÓN. (NIVEL INICIAL)

PLAN DE FORMACIÓN DEL PERSONAL DE ADMINISTRACIÓN Y SERVICIOS

FUNDAMENTOS ESTADÍSTICOS PARA INVESTIGACIÓN. (NIVEL INICIAL)
Con SPPS , EXCEL y R

Antonio Maurandi López (amaurandi@um.es)
Cálculo Científico y Apoyo Estadístico. SAI
www.um.es/sai www.um.es/ae

Propósito general del curso

- *Proporcionar conocimientos de estadística básica. Empleo básico de SPSS y MS-excel en resultados de laboratorio.*
- Con este fin:
  - Intentaremos usar Excel hasta donde sea posible y siempre que sea posible de forma natural o sencilla.
  - Nos centraremos en practicas usuales de un laboratorio experimental.
  - Buscaremos la sencillez, el curso es básico, pero sin dejar de lado cierto rigor matemático.
  - Procuraremos cubrir las técnicas mas universalmente usadas.

De donde partimos.

- Partimos de un cierto dominio de un entorno informático estándar.
- Partimos de por lo menos un conocimiento básico del programa Excel.

TEMAS

1. Entorno de trabajo SPSS.
2. Estadística Descriptiva con SPSS y MS-Excel.
3. Contrastes. Fundamentos y Generalidades
4. Supuestos de Normalidad y Homocedasticidad.
5. Correlación lineal.
6. Regresión lineal.
7. T-test. Comparación de medias.

Objetivos concretos

- Alcance de la estadística en la práctica científica.
- Cuando estamos haciendo buena estadística.
- Que puedo afirmar y que no.
- Clarificar algunas dudas que existen sobre la estadística.
- Adquirir buenas maneras en la práctica estadística.
- Ser capaces de realizar algunos 'análisis' e interpretarlos correctamente.

TEMA 1

Entorno de trabajo SPSS

Índice

- Que es SPSS
- Entorno SPSS
- Variables.

Tema 1

- *"La meta más simple a la hora de analizar datos es sacar las conclusiones más fuertes con la cantidad limitada de datos de que disponemos".*

Definiciones

- Definiciones: Variable, Dato, Caso.
  - **Variable** es una característica (magnitud, vector o número) que puede ser medida, adoptando diferentes valores en cada uno de los casos de un estudio.
  - **Dato** es la unidad mínima de información referida a un objeto.
  - **Caso**. 'Cada una de las observaciones'.

Estructura de la información

Los datos en estadística se presentan en matrices, filas y columnas.

Cada **Fila** un Caso (un individuo)

Cada **Columna** una Variable

	id	cliente	ingresos
1	1	Cliente habitual	$3,787
2	2	Cliente habitual	$1,734
3	3	Cliente preferente	$2,126
4	4	Cliente preferente	$2,259
5	5	Cliente preferente	$1,587
6	6	Cliente habitual	$0
7	7	Cliente preferente	$1,838
8	8	Cliente habitual	$1,847
9	9	Cliente habitual	$1,714
10	10	Cliente preferente	$1,718
11	11	Cliente preferente	$4,388
12	12	Cliente preferente	$3,155
13	13	Cliente preferente	$3,834
14	14	Cliente preferente	$2,291
15	15	Cliente preferente	$4,140
16	16	Cliente preferente	$2,194
17	17	Cliente preferente	$2,938
18	18	Cliente habitual	$3,313

Tipos de variables

Nivel de medida	Tipo
Nominal	Cualitativa/Categórica
Ordinal	Ordinal o cuasicuantitativa
Intervalo	Cuantitativa
Razón	Cuantitativa

SPSS

- Originalmente SPSS era el acrónimo de (**S**tatistical **P**ackage for the **S**ocial **S**ciences). Después la sigla designo tanto el programa como la empresa que lo producia SPSS (**S**tatistical **P**roduct and **S**ervice **S**olutions) .
- Actualmente SPSS ha sido comprada por IBM y el programa paso a llamarse **pasw statistics** y posteriormente "**IBM SPSS Statistics**"

La licencia de SPSS de la UMU

- SPSS 15.0 Family
  - SPSS Base +6 mod
    - SPSS Modelos Avanzados.
    - SPSS Modelos de Regresión.
    - SPSS Tablas.
    - SPSS Tendencias.
    - SPSS Categorías.
    - SPSS Conjoint .
  - SPSS Pruebas Exactas

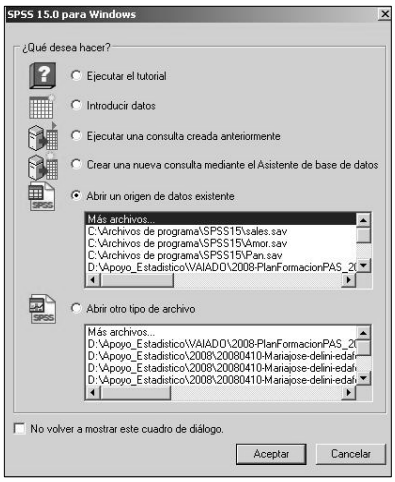
SPSS (2)

- El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos.
- SPSS for Macintosh 17 (2008) (OS X 10.4, OSX 10.5)
- SPSS para Windows 15 (2009) (**win XP** y parche para win Vista)
- SPSS 16, para Linux 17 (2008) (Kernel 2.9.9.42)
  - No hay versión para Linux de ciertos módulos como el de pruebas exactas.
- IBM SPSS Statistics 19 (2010). (Linux, Mac OS X, Win)

SPSS (3)

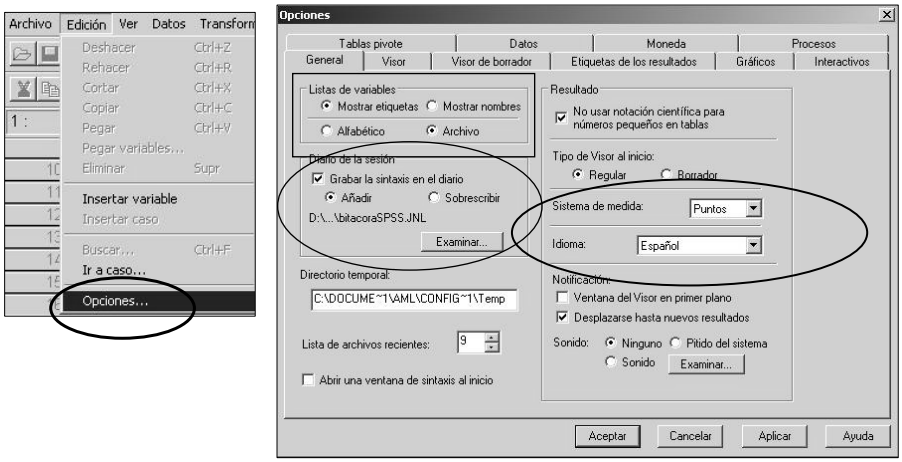
- SPSS 15 tiene un sistema de ficheros en el cual el principal son los archivos de datos (extensión .SAV). A parte de este tipo existen otros dos tipos de uso frecuente:
  - Archivos de salida (output, extensión .SPO): en estos se despliega toda la información de manipulación de los datos que realizan los usuarios mediante las ventanas de comandos. Son susceptibles de ser exportados con varios formatos (originalmente HTML, RTF, TXT, XLS y DOC, actualmente la versión 15 incorpora la exportación a PDF.
  - Archivos de sintaxis (extensión .SPS).

SPSS (4). Inicio

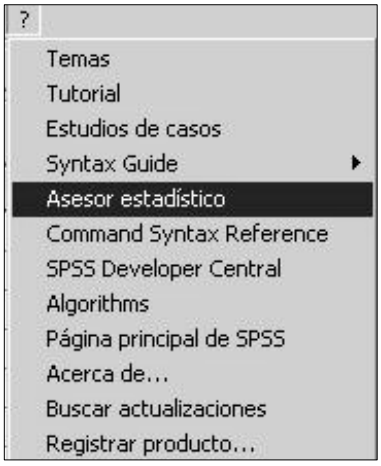


SPSS (5). Ventanas

SPSS (6). Opciones

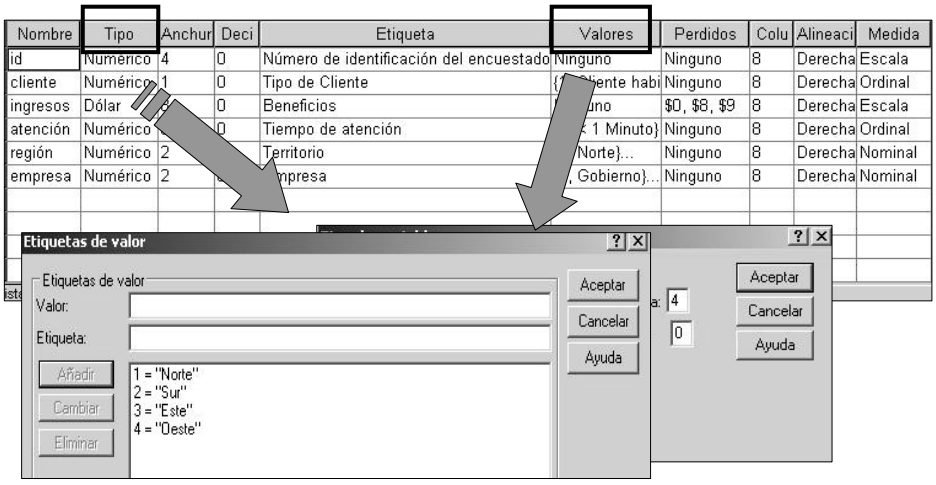


SPSS (7). Ayuda/s



- Temas
- Tutorial
- Asesor estadístico

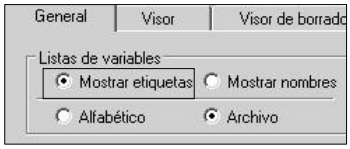
Variables (1)



Variables (2)

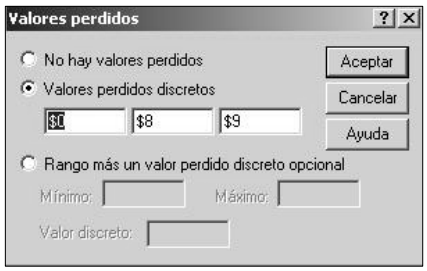
Nivel de medida	Tipo	SPSS
Nominal	Cualitativa/Categórica	Nominal
Ordinal	Ordinal o cuasicuantitativa	Ordinal
Intervalo	Cuantitativa	Escala
Razón	Cuantitativa	Escala

Menú opciones



Ausentes (missings)

- Definidos por el sistema
  - Valores que se dejaron en blanco en la introducción de datos
- Definidos por el usuario
  - *Trucos para hacer ciertos análisis.*
  - Exploración de datos.



Variables (3)

- **Compatibilidad de los ficheros de SPSS entre las diferentes versiones.**
  - Los ficheros de datos de SPSS (**.sav**) son compatibles entre todas las versiones. Solo hay que tener una cosa en cuenta, SPSS 12 y las versiones posteriores permiten variables con nombres de más de 8 caracteres, mientras que las versiones anteriores a la 13 solo permiten un máximo de 8 caracteres.
  - Si queremos utilizar una versión anterior debemos cerciorarnos de que los nombres de las variables no exceden los 8 caracteres. En todo caso spss v.X abrirá el fichero .sav pero truncara las variables con nombres largos.

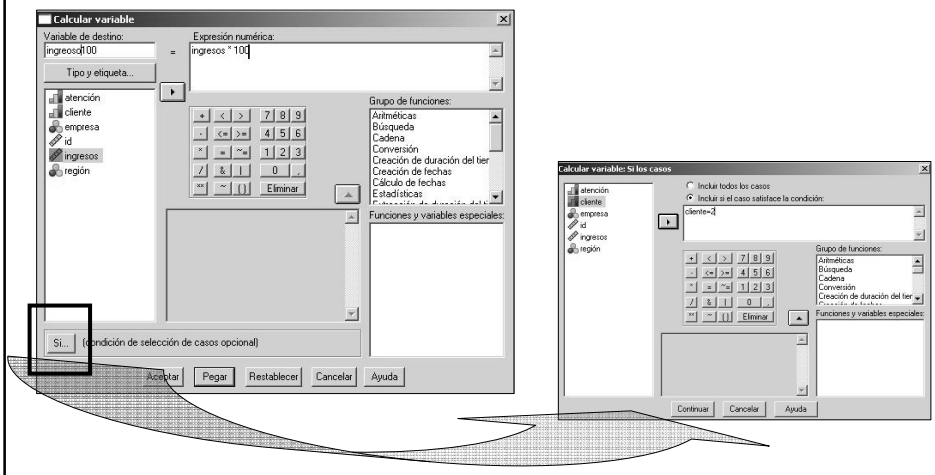
Variables (4)

- El primer carácter ha de ser una letra.
- No se pueden usar espacios en blanco
- Mejor si solo tiene 8 caracteres para la compatibilidad con versiones anteriores (v12).
- No diferencia mayúsculas y minúsculas.
- Hay ciertas palabras clave que no se pueden utilizar (ALL, AND, BY,…)

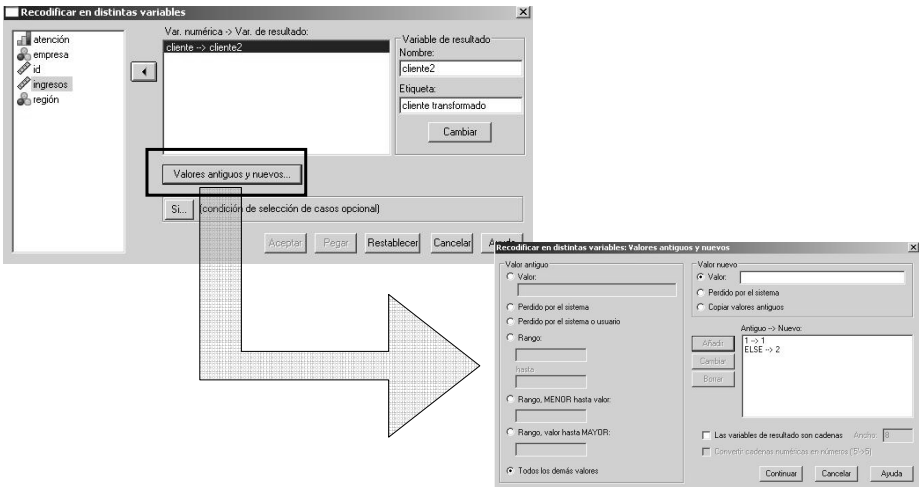
Transformación de variables

- Calculada a partir de otra variable: Calcular
- Recodificación
  - En la misma variable
  - En otra variable
- Comando '$CASENUM'

Calcular

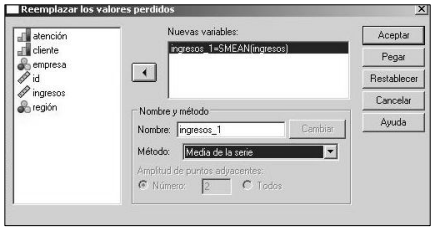


Transformar



Tratamiento de missings

- Media de la Serie
- Media de puntos adyacentes.
- Mediana de puntos adyacentes
- Interpolación lineal
- Tendencia lineal en el punto.



Ponderación de casos (1)

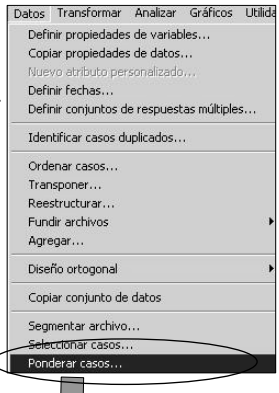
- En ciertas ocasiones los datos se presentan resumidos, e.d. un valor de una variable representa más de un caso o representa una importancia relativa.

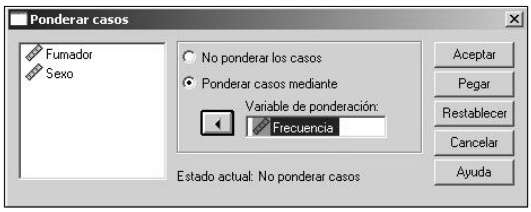
	Hombre	Mujer	Total
Fumador	40	42	82
No Fumador	80	110	190
Total	120	152	272

Fuma	Sexo	Frec
SI	H	40
No	H	80
SI	M	42
NO	M	110

Ponderación de casos (2)

	Fumador	Sexo	Frecuencia
1	Si	Hombre	40
2	No	Hombre	80
3	Si	Mujer	42
4	No	Mujer	110
5			





Segmentación de ficheros

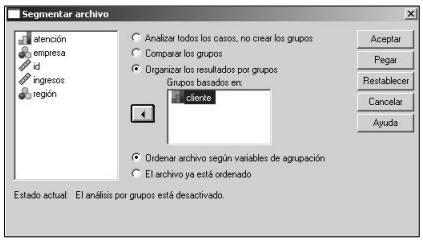
- En muchas ocasiones conviene realizar operaciones según ciertos grupos y no al cto total de datos. Decimos que 'Segmentamos el fichero según una variable'.

Tipo de Cliente = Cliente habitual

Estadísticos descriptivos[a]

	N	Media	Desv. típ.
Beneficios	635	$2,583.63	$947.593
N válido (según lista)	635		

a. Tipo de Cliente = Cliente habitual



Tipo de Cliente = Cliente preferente

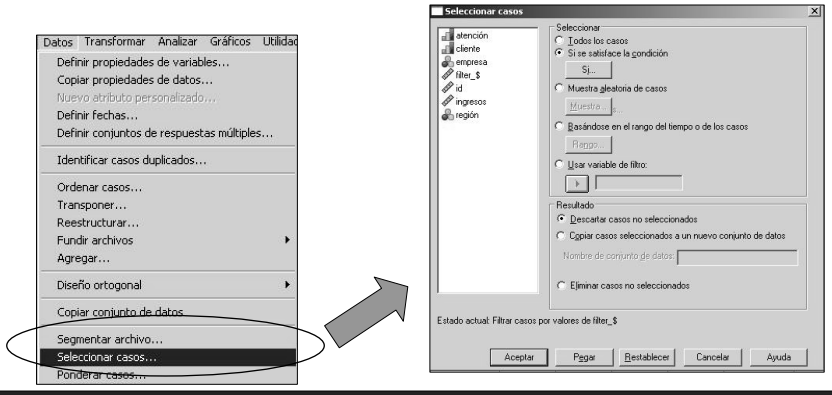
Estadísticos descriptivos[a]

	N	Media	Desv. típ.
Beneficios	853	$2,466.66	*********
N válido (según lista)	853		

a. Tipo de Cliente = Cliente preferente

Selección de casos

- En ocasiones conviene eliminar ciertos caso que satisfacen cierta condición.
  - Por ejemplo No queremos trabajar con 'Clientes habituales' de la 'región Este' que tengan un salario inferior a $2000.



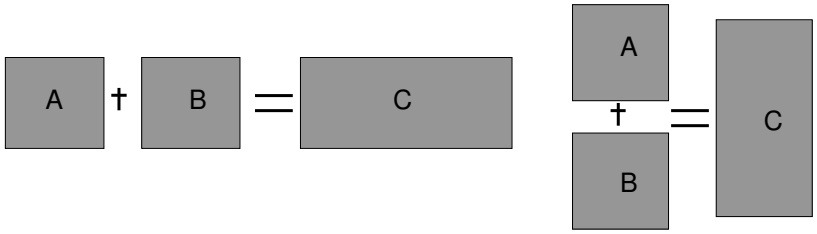
Selección de casos (2)

~ (cliente=1 & región=3 &
(ingresos <= 2000))

	id	cliente	ingresos	atención	región	empresa	filter_$
7	7	Cliente preferente	$1,838	> 4 Minuto	Sur	Gobierno	Seleccionado
8	8	Cliente habitual	$1,847	> 4 Minuto	Oeste	Gobierno	Seleccionado
9	9	Cliente habitual	$1,714	1-2 Minuto	Este	Educación	No selecciona
10	10	Cliente preferente	$1,718	> 4 Minuto	Sur	Educación	Seleccionado
11	11	Cliente preferente	$4,388	< 1 Minuto	Oeste	Gobierno	Seleccionado
12	12	Cliente preferente	$3,155	< 1 Minuto	Norte	Cia. Comer	Seleccionado
13	13	Cliente preferente	$3,834	> 4 Minuto	Oeste	Gobierno	Seleccionado
14	14	Cliente preferente	$2,291	> 4 Minuto	Norte	Cia. Comer	Seleccionado
15	15	Cliente preferente	$4,140	1-2 Minuto	Sur	Cia. Comer	Seleccionado
16	16	Cliente preferente	$2,194	1-2 Minuto	Sur	Educación	Seleccionado
17	17	Cliente preferente	$2,938	1-2 Minuto	Sur	Gobierno	Seleccionado
18	18	Cliente habitual	$3,313	> 4 Minuto	Este	Cia. Comer	Seleccionado
19	19	Cliente preferente	$3,327	> 4 Minuto	Este	Educación	Seleccionado
20	20	Cliente habitual	$1,449	1-2 Minuto	Este	Gobierno	No selecciona
21	21	Cliente preferente	$2,696	1-2 Minuto	Norte	Educación	Seleccionado

Unir (fundir) ficheros

- Unión simple Horizontal
- Unión mediante variables clave Horizontal
- Unión simple vertical
- Unión mediante variables clave Vertical



Otras funcionalidades

- Ordenar datos (fichero .sav)
- Recuento de valores
- Trasponer casos a variables
- Sintaxis (botón ‘Pegar’)

TEMA 2

Estadística Descriptiva con SPSS y MS-EXCEL

Tema 2
Estadística con SPSS y MS-EXCEL

- Estadística descriptiva.
  - Algunas definiciones
  - Parámetros y estadísticos.
  - Medidas de Posición, Centralización, Dispersión, Forma
  - Box Plots
- Excel. Análisis de datos.
  - Instalación.
  - Estadística descriptiva.
  - Histograma, Frecuencias

Algunas definiciones

- **Población**: Llamamos población estadística, universo o colectivo al conjunto de referencia sobre el cual van a recaer las observaciones.
- **Individuos**: Se llama unidad estadística o individuo a cada uno de los elementos que componen la población estadística. El individuo es un ente observable que no tiene por qué ser una persona, puede ser un objeto, un ser vivo, o incluso algo abstracto.
- **Muestra**: Es un subconjunto de elementos de la población. Se suelen tomar muestras cuando es difícil o costosa la observación de todos los elementos de la población estadística.
- **Censo**: Decimos que realizamos un censo cuando se observan todos los elementos de la población estadística.
- **Caracteres**: La observación del individuo la describimos mediante uno o más caracteres. El carácter es, por tanto una cualidad o propiedad inherente en el individuo.

  tipos de caracteres :

  Cualitativos : aquellos que son categóricos, pero no son numéricos.
  p. ej. <color de los ojos>, <profesión>, <marca de coche>,...

  Ordinales : aquellos que pueden ordenarse, pero no son numéricos.
  p. ej. <preguntas de encuesta sobre el grado de satisfacción de algo>
  Mucho, poco, nada. Bueno, regular, malo, ...

  Cuantitativos : son numéricos. p. ej. <peso>, <talla>, <núm. de hijos>, <núm. de libros leídos al mes>,...

Algunas definiciones (2)

- **Modalidad Valor:** Un carácter puede mostrar distintas modalidades o valores, es decir, son distintas manifestaciones o situaciones posibles que puede presentar un carácter estadístico. Las modalidades o valores son incompatibles y exhaustivos.
  - Generalmente se utiliza el término modalidad cuando hablamos de caracteres cualitativos y el término valor cuando estudiamos caracteres cuantitativos.
  - p. ej. el carácter cualitativo <Estado Civil> puede adoptar las modalidades : casado, soltero, viudo. El carácter cuantitativo <Edad> puede tomar los valores : diez, once, doce, quince años, ...
- **Variable Estadística:** Al conjunto de los distintos valores numéricos que adopta un carácter cuantitativo se llama variable estadística.

  Tipos de variables estadísticas :

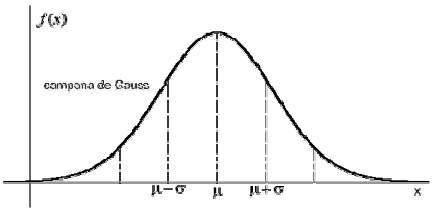
  Discretas : Aquellas que toman valores aislados (números naturales), y que no pueden tomar ningún valor intermedio entre dos consecutivos fijados. p. ej. <núm. de goles marcados>, <núm. de hijos>, <núm., de discos comprados>, <núm. de pulsaciones>,...

  Continuas : Aquellas que toman infinitos valores (números reales) en un intervalo dado, de forma que pueden tomar cualquier valor intermedio, al menos teóricamente, en su rango de variación. p. ej. <talla>, <peso>, <presión sanguínea>, <temperatura>, ...
- **Observación**: Una observación es el conjunto de modalidades o valores de cada variable estadística medidos en un mismo individuo.

  p. ej. en una población de 100 individuos podemos estudiar, de forma individual, tres caracteres : <edad : 18, 19, ...>, <sexo : Hombre, Mujer> y <si ha votado en las elecciones : Si, No>. Realizamos 100 observaciones con tres datos cada una, es decir, una de las observaciones podría ser (43, H, S).

Algunas definiciones (3)

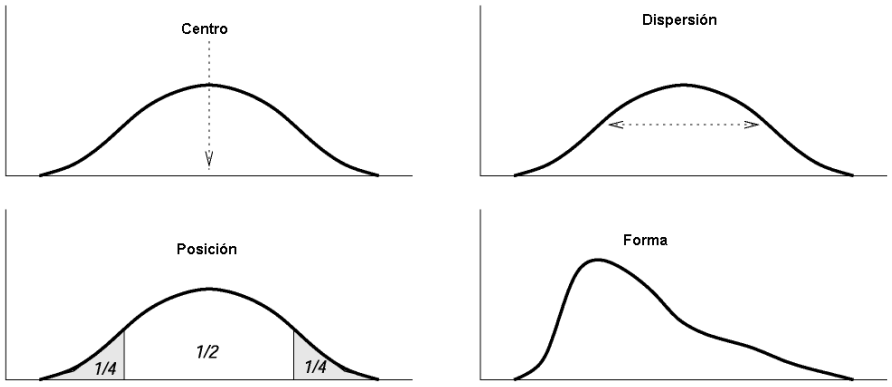
- La **función de densidad** de probabilidad, función de densidad, o, simplemente, densidad de una variable aleatoria continua es una función, usualmente denominada f(x) que describe la densidad de la probabilidad en cada punto del espacio de tal manera que la probabilidad de que la variable aleatoria tome un valor dentro de un determinado conjunto sea la integral de la función de densidad sobre dicho conjunto.



Parámetros y estadísticos

- Parámetro: Es una cantidad numérica calculada sobre una población
  - La altura media de los individuos de un país
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).
- Estadístico: Ídem (cambiar población por muestra)
  - La altura media de los que estamos en este aula.
    - Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

Tipos de estadísticos.



Tipos de estadísticos (2)

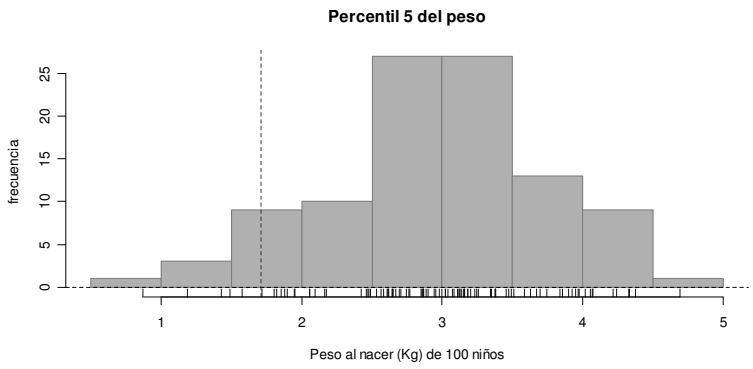
- Posición
- Centralización
- Dispersión
- Forma

Estadísticos de posición

- Posición
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
  - Cuantiles, percentiles, cuartiles, deciles,...
  - *Se define el cuantil de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.*
    - *El Percentil de orden k es el cuantil de ornden k/100*.

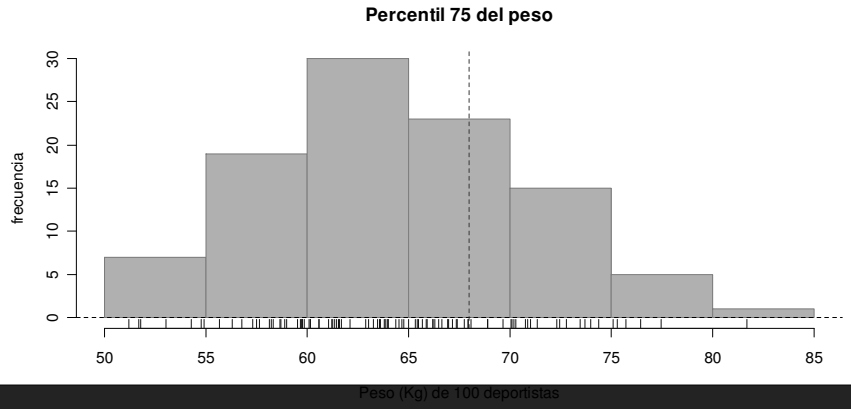
Estadísticos de posición (2)

- El 5% de los recién nacidos tiene un peso demasiado bajo. ¿Qué peso se considera "demasiado bajo"?
  - Percentil 5 o cuantil 0,05



Estadísticos de posición (3)

- ¿Qué peso es superado sólo por el 25% de los individuos?
  - Percentil 75, tercer cuartil o cuantil 0.75



Estadísticos de Centralización

- Centralización
  - Indican valores con respecto a los que los datos '*parecen'* agruparse.
  - Media, mediana y moda,…

Estadísticos de Centralización (2).

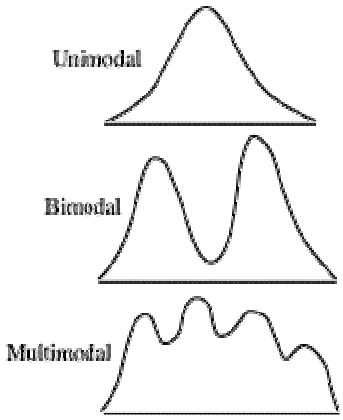
- Media (‘mean’) Es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de 2,2,3,7 es (2+2+3+7)/4=3,5
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos.

Estadísticos de Centralización (3).

- Mediana (‘median’) Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1,2,4,**5**,6,6,8 es 5
  - Mediana de 1,2,4,**5**,**6**,6,8,9 es (5+6)/2=5,5
  - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
    - Mediana de 1,2,4,**5**,6,6,**800** es 5. ¡La media es 117,7!

Estadísticos de Centralización (4).

- Moda ('mode') Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.



Estadísticos de Centralización (5).

- Datos sin agrupar: $x_1, x_2, ..., x_n$
  - Media $\bar{x} = \frac{\sum_i x_i}{n}$
- Datos organizados en tabla
  - si está en intervalos usar como x_i las marcas de clase. Si no ignorar la columna de intervalos.

Variable		fr.	fr. ac.
$L_0 - L_1$	x_1	n_1	N_1
$L_1 - L_2$	x_2	n_2	N_2
...			
$L_{k-1} - L_k$	x_k	n_k	N_k
n			

  - Media

$$\bar{x} = \frac{\sum_i x_i n_i}{n}$$

Estadísticos de dispersión.

- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
  - Desviación típica, coeficiente de variación, rango, varianza,..

Estadísticos de dispersión (2)

- Los estudiantes de Bioestadística reciben diferentes calificaciones en la asignatura (variabilidad). ¿A qué puede deberse?
  - Diferencias individuales en el conocimiento de la materia.
- ¿Podría haber otras razones (fuentes de variabilidad)?

Estadísticos de dispersión (3)

- Por ejemplo supongamos que todos los alumnos poseen el mismo nivel de conocimiento. ¿Las notas serían las mismas en todos? –
  → Seguramente No.
  - Dormir poco el día del examen.
    - Diferencias individuales en la habilidad para hacer un examen.
  - El examen no es una medida perfecta del conocimiento.
    - Variabilidad por error de medida.
  - En alguna pregunta difícil, se duda entre varias opciones, y al azar se elige la mala
    - Variabilidad por azar, aleatoriedad.

Estadísticos de dispersión.(4)

- Amplitud o Rango ('*range*'):
  Diferencia entre observaciones extremas.
- Rango intercuartílico ('*interquartile range'=IQR*):
  - Es la distancia entre primer y tercer cuartil.
    - Rango intercuartílico = $P_{75} - P_{25} = Q_3 - Q_1$

Estadísticos de dispersión.(5)

- Varianza S^2 ('*Variance*'): Mide el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a la media. $$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$
  - Es sensible a valores extremos (alejados de la media).
  - Sus unidades son el cuadrado de las de la variable.

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \qquad \hat{S}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

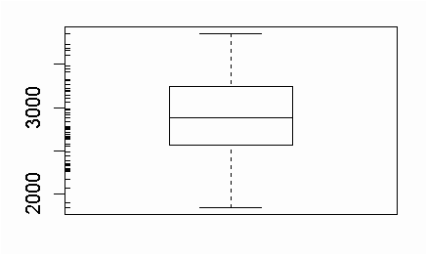
Estadísticos de dispersión.(6)

Desviación típica ('*standard deviation', SD*)
Es la raíz cuadrada de la varianza. $$S = \sqrt{S^2}$$

- Tiene las misma dimensionalidad (unidades) que la variable. Versión 'estética' de la varianza.

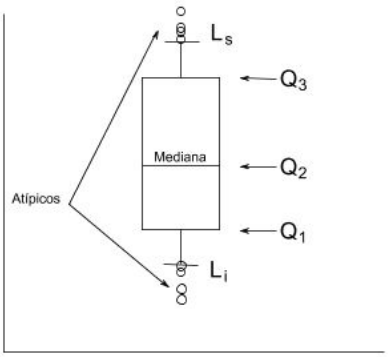
Box plots (Diagramas de caja)

- Un **diagrama de caja,** John Tukey (1977), es un gráfico, basado en cuartiles, mediante el cual se visualiza un conjunto de datos. Está compuesto por un rectángulo, la caja, y dos brazos, los bigotes.
  - También llamados 'diagramas de caja y bigotes'.



Box plots (2)

- Es un gráfico que se suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y simetría de la distribución.



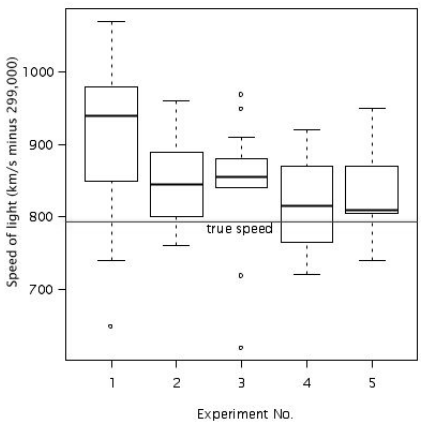
L_i=Q1-1.5*IQR L_s=Q3+1.5*IQR. (IQR=Q3-Q1)

Los valores atipicos son los inferiores a Li y los superiores a Ls

Box plots (3)

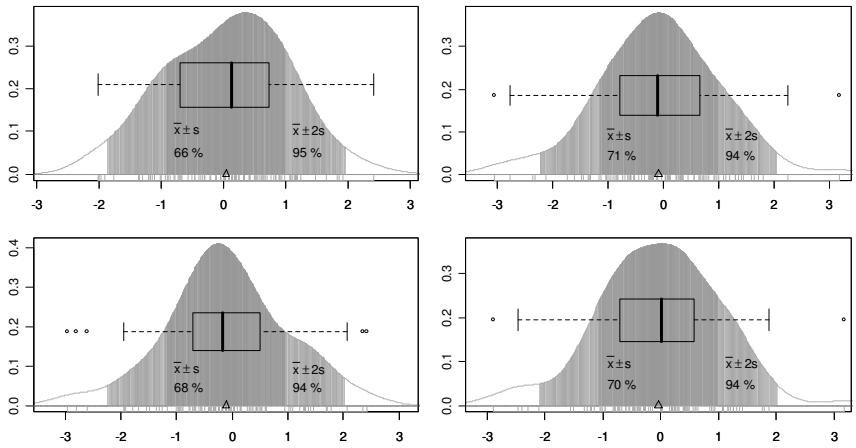
- Proporcionan una visión general de la simetría de la distribución de los datos, si la media no está en el centro del rectángulo, la distribución no es simétrica.
- Son útiles para ver la presencia de valores atípicos.
- Muy útiles para comparar distribuciones.

Box plots (4)

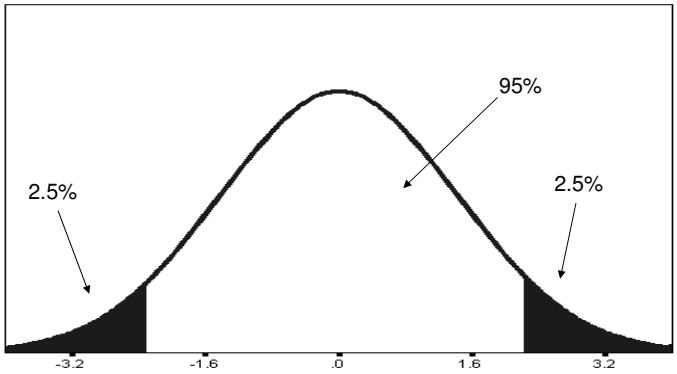


Box plots (5)

Datos 'casi normales'. Eje 'x' medido en desviaciones típicas



Estadísticos de dispersión.



La probabilidad de que una variable normal tipificada tome valores en el intervalo [-1.96,1.96] es del 95%.

Estadísticos de dispersión.(9)

Coeficiente de variación

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

Es la razón entre la desviación típica y la media.

- Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media"
- También se la denomina variabilidad relativa.
- Es una cantidad **adimensional**. Interesante para comparar la variabilidad de diferentes variables.
  - Si el peso tiene CV=30% y la altura tiene CV=10%, los individuos presentan más dispersión en peso que en altura.

Estadísticos de forma.

- Forma
  - Asimetría.
  - Apuntamiento o curtosis.

Estadísticos de forma. (2)

- Las discrepancias entre las medidas de centralización son indicación de asimetría.
- Coeficiente de asimetría. (positiva o negativa).
- Distribución simétrica ➔ asimetría nula.

Estadísticos de forma. (3)

La curtosis nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.

- Platicúrtica (aplanada): curtosis < 0
- Mesocúrtica (como la normal): curtosis = 0
- Leptocúrtica (apuntada): curtosis > 0

Estadísticos de forma. (3)

- Regla aproximativa (para ambos estadísticos).
  - Curtosis y/o Coef de asimetría entre -1 y 1, es generalmente considerada una muy ligera desviación de la normalidad.
  - Entre -2 y 2 tampoco es malo del todo, según el caso.

¿Que hemos visto hasta ahora?

- Parámetros
- Estadísticos y estimadores
- Clasificación
  - Posición (*cuantiles, percentiles*,...)
  - Medidas de centralización: *Media, mediana y moda*
  - Medidas de dispersión: *Rango, Rango Intercuartilico (RI), Var y* ***SD***
  - Asimetría (*coef de asimetría*)
  - Medidas de apuntamiento (*curtosis*)

De Forma

SEM

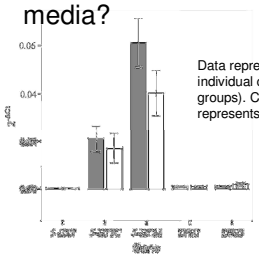
- **Error típico de la media**: Cuando estimamos la media a partir de una muestra de un determinado tamaño (n) los valores que toma la media en las diferentes muestras varía. A la desviación típica de los valores que toma el estadístico se le denomina error típico de la media. Da una idea de la variabilidad del estadístico.
  - Ojo!: No de la distribución de la variable.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$S_{\bar{x}} = \frac{\hat{S}}{\sqrt{n}}$$

SEM vs SD

- Un 'error' muy típico es mostrar medias y errores típicos de la media en lugar de la Desviación típica y además poner bar plots.
- ¿Es un error o es intencionado?
- ¿PQ solemos mostrar Gráficos de barras y Errores típicos de la media?



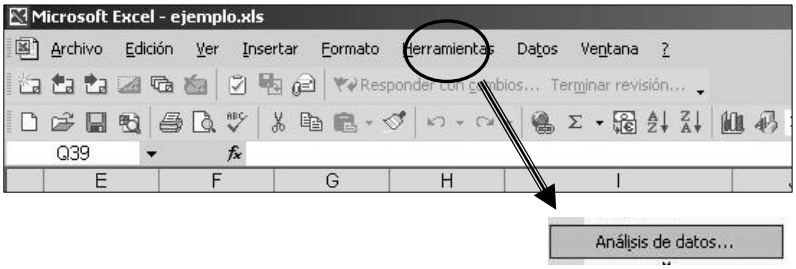
Data represents the **mean +SEM** for the individual cells studied (N=7 for both groups). Current trace calibration bar represents 1000 pA.....etc

Excel

- Hoja de calculo.
  - Permite manipular datos numéricos y alfanuméricos dispuestos en forma de tablas.
  - Es posible realizar cálculos complejos con fórmulas y funciones y dibujar distintos tipos de gráficas.
- Además ofrece un conjunto de herramientas para el análisis de datos con el que se puede ahorrar tiempo y esfuerzo en un análisis estadístico.
  - Se proporcionan los datos o los rangos de datos y parámetros y MS-Excel utilizará las funciones macro apropiadas para mostrar lo resultados en tablas de resultados.
  - También genera gráficos.

Análisis de datos

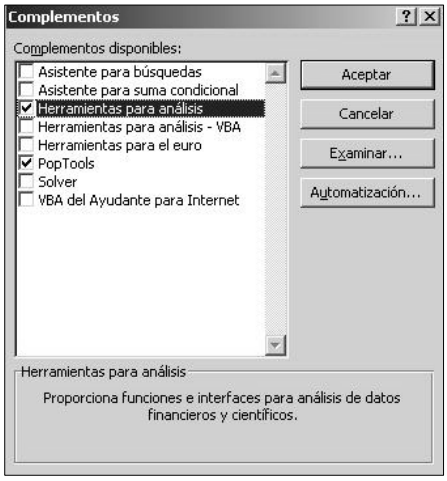
- Análisis de datos en el menú Herramientas.



• Si este comando no está en el menú, ejecute el programa de instalación para instalar las Herramientas para análisis de la forma siguiente:

Análisis de datos. (2)

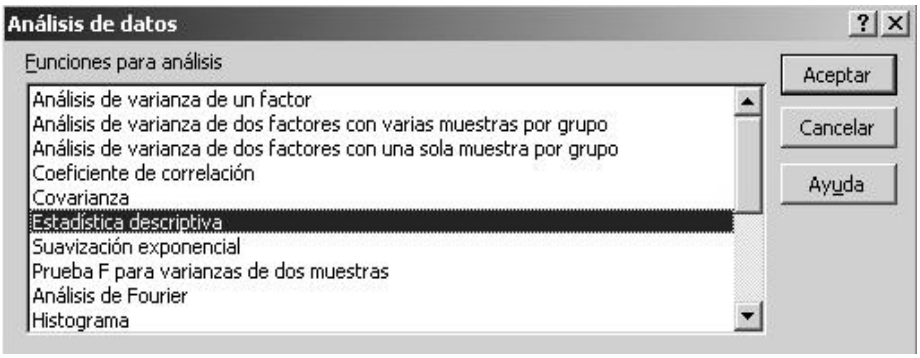
- Seleccionar el menú Herramientas | Complementos…
- y seleccionar las casillas de verificación de los complementos a agregar.



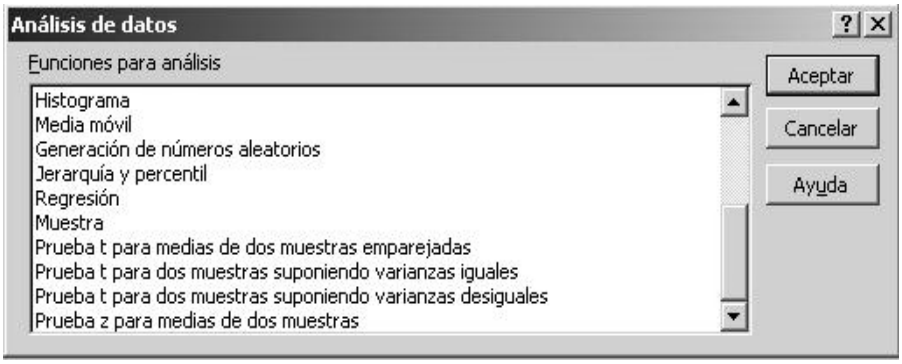
Análisis de datos. (3)

- Herramientas para análisis: *Agrega funciones y herramientas de análisis financiero, estadístico y técnico.*
- Hacer clic en Aceptar.
- Aparecerán en el menú Herramientas los complementos que se seleccionaron para poder utilizarse.

Análisis de datos. (4)



Análisis de datos. (5)

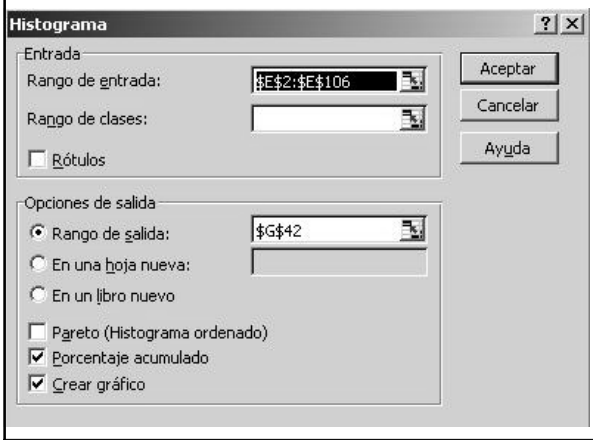


Análisis de datos. (6)

- Nosotros nos vamos a centrar de momento en:
  - Histograma.
  - Estadística descriptiva.

Análisis de datos. (7)

- Histograma

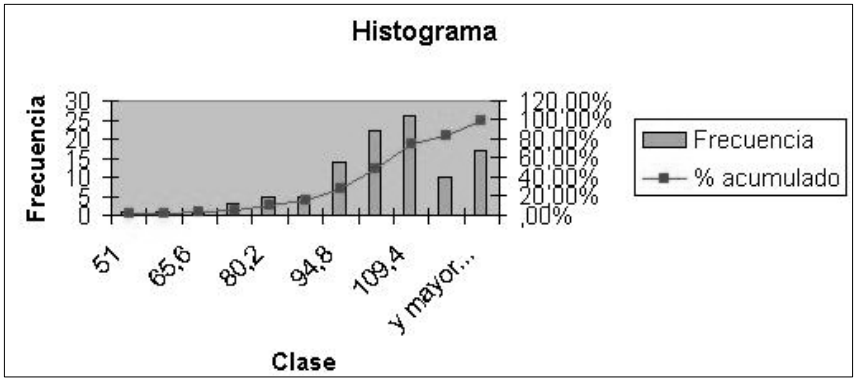


	A	B	C	D	E
1	id	Gender	Ethnicity	Grade	Total
2	106484	2	2	D	80
3	108642	2	4	C	96
4	127285	1	4	C	98
5	132931	1	3	B	103
6	140219	1	2	B	108
7	142630	1	4	A	122
8	153964	2	2	A	112
9	154441	1	5	A	120
10	157147	2	4	A	123
11	164605	1	3	A	124
12	164842	1	1	C	97
13	167664	2	4	A	118
14	175325	2	4	B	111
15	192627	1	4	D	84
16	211239	2	4	D	79
17	219593	1	5	C	94
18	237983	2	2	C	92
19	245473	2	4	C	88
20	249586	2	4	C	98
21	260983	2	4	B	106

Análisis de datos. (8)

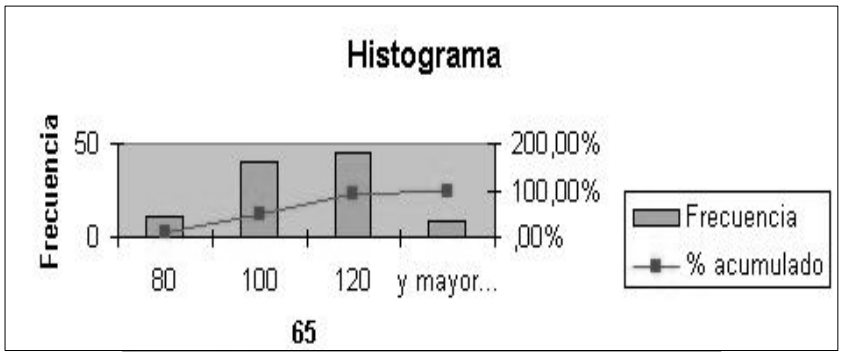
Clase	*Frecuencia*	*% acumulado*
51	1	,95%
58,3	1	1,90%
65,6	1	2,86%
72,9	3	5,71%
80,2	5	10,48%
87,5	5	15,24%
94,8	14	28,57%
102,1	22	49,52%
109,4	26	74,29%
116,7	10	83,81%
y mayor...	17	100,00%

Análisis de datos. (9)



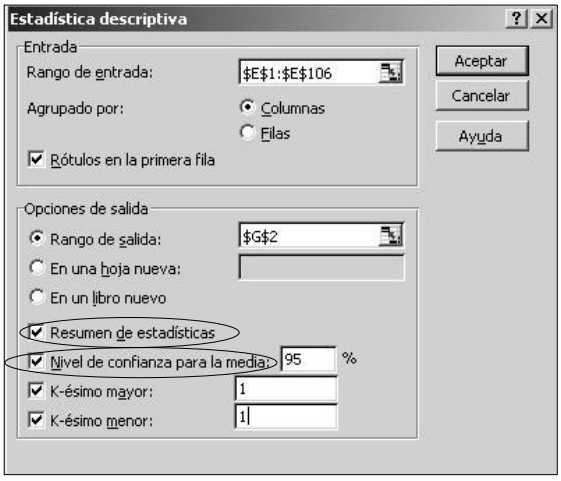
Análisis de datos. (10)

- También podemos indicar las 'clases', indicando sus extremos superiores.



Análisis de datos. (11)

- Opción Estadística descriptiva.

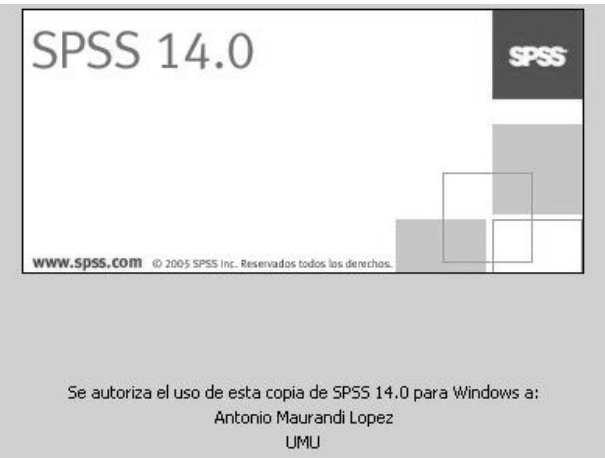


Análisis de datos. (12)

Total	
Media	100,5714286
Error típico	1,493076644
Mediana	103
Moda	98
Desviación estándar	15,29948286
Varianza de la muestra	234,0741758
Curtosis	0,942541816
Coeficiente de asimetría	-0,836688049
Rango	73
Mínimo	51
Máximo	124
Suma	10560
Cuenta	105
Mayor (1)	124
Menor(1)	51
Nivel de confianza(95,0%)	2,960823187

Media
Error típico de la media
Mediana
Desviación Estándar
Cutosis
Coef de Asimetría
Cuenta = tamaño Muestral, n.
Nivel de confianza, 'margen de error'

SPSS. (1)



SPSS (2)

- Tres etapas fundamentales en el análisis de datos:
  - Validación de datos.
  - Descripción de los datos.
  - Análisis estadístico.

Introducción a SPSS (4)

- Barra de menú.
- Barra de herramientas.
- Línea de edición de datos.
- Vista de datos.
- Vista de variables.

Introducción a SPSS (5)

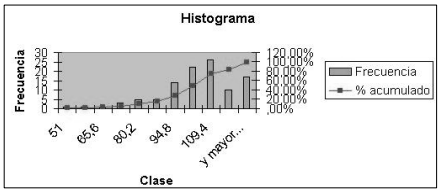
- Analizar/Informes/Resúmenes de casos
- Analizar/Estadísticos descriptivos/Frecuencias.
- Analizar/Estadísticos descriptivos/Descriptivos.
- Gráficos
  - Histograma.
  - Diagrama de barras para dos variables.
  - Tallo y hojas.
  - Diagramas de caja.
  - Diagrama de dispersión.
- Correlación entre dos variables.
- Tablas de contingencia.

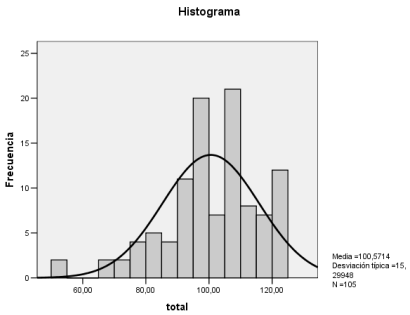
Introducción a SPSS (6)

- Menú Datos
  - Ordenar casos…
  - Fundir archivos
  - Segmentar archivo…
  - Ponderar casos…
- Transformar
  - Calcular
  - Recodificar

Descriptiva con SPSS (1)

- En el ejemplo anterior, con SPSS
- Es más informativo, y sin 'trastear' es más informativo que el de Excel





Descriptiva con SPSS (2)

Descriptivos

			Estadístico	Error típ.
total	Media		100,5714	1,49308
	Intervalo de confianza para la media al 95%	Límite inferior	97,6106	
		Límite superior	103,5323	
	Media recortada al 5%		101,4921	
	Mediana		103,0000	
	Varianza		234,074	
	Desv. típ.		15,29948	
	Mínimo		51,00	
	Máximo		124,00	
	Rango		73,00	
	Amplitud intercuartil		19,00	
	Asimetría		-,837	,236
	Curtosis		,943	,467

Florence Nightingale (1820-1910), Orden del Merito del Reino Unido, británica, es considerada una de las pioneras en la práctica de la enfermería. Se la considera la madre de la enfermería moderna y creadora del primer modelo conceptual de enfermería. Destacó desde muy joven en la matemática, aplicando después sus conocimientos de estadística a la epidemiología y a la estadística sanitaria**. Inventó los gráficos de sectores o histogramas** para exponer los resultados de sus reformas. En 1858, fue la primera mujer miembro de la Statistical Society. Tb fue miembro honorario de la American Statistical Association.

Durante la guerra de Secesión en 1861 fue llamada por el gobierno de la Unión para que organizara sus hospitales de campaña durante la guerra la cual redujo del 44% de heridos al 2.2 %.

TEMA 3

Introducción a los contrastes estadísticos.

Tema 3
Introducción a los contrastes estadísticos.

- Saber lo que hago.
- Cuándo necesito el apoyo de la estadística.
- Población y muestra.
- La importancia de la independencia de las observaciones.
- Cómo usamos la estadística para extrapolar datos de nuestra muestra a la población.
- Limitaciones de la estadística y sentido común.
- La distribución normal (TCL).
- Intervalos de confianza.
- ¿Por qué el 95%?

Tema 3
Introducción a los contrastes estadísticos. (2)

- ¿Que es un P-valor? (p-valor, Hip nula).
- Resultados significativos en la ciencia.
- Potencia estadística.
- El problema de las comparaciones múltiples.
- Valores atípicos (Normalización).

Saber lo que hago.

- Las computadoras y los paquetes estadísticos en general son excelentes herramientas para analizar datos.
- Problema. El mal uso:
  - Introducir incorrectamente los datos.
  - Seleccionar un test inapropiado.

Cuándo necesito el apoyo de la estadística.

- La meta más simple a la hora de analizar datos es sacar las conclusiones más fuertes con la cantidad limitada de datos de que disponemos.
- Dos problemas:
  - Diferencias importantes pueden ser oscurecidas por variabilidad biológica e imprecisión experimental. Se hace complicado distinguir entre **diferencias reales** y **variabilidad aleatoria**.
  - El ser humano es capaz de encontrar modelos. Nuestra inclinación natural (*especialmente con nuestros propios datos*) es concluir que las diferencias observadas son REALES, *tendemos a minimizar los efectos de la variabilidad aleatoria*.
- El rigor estadístico nos previene de cometer estos errores.

Variabilidad (1)

- Los estudiantes de Bioestadística reciben diferentes calificaciones en la asignatura (variabilidad). ¿A qué puede deberse?
  - Diferencias individuales en el conocimiento de la materia.
- ¿Podría haber otras razones (fuentes de variabilidad)?

Variabilidad (2)

- Por ejemplo supongamos que todos los alumnos poseen el mismo nivel de conocimiento. ¿Las notas serían las mismas en todos? Seguramente No.
  - Dormir poco el día del examen.
    - Diferencias individuales en la habilidad para hacer un examen.
  - El examen no es una medida perfecta del conocimiento.
    - Variabilidad por error de medida.
  - En alguna pregunta difícil, se duda entre varias opciones, y al azar se elige la mala
    - Variabilidad por azar, aleatoriedad.

Población y muestra

- La idea básica de las estadística es extrapolar, desde los datos recogidos, para llegar a conclusiones más generales sobre la población de la que se han recogido los datos.

Población y muestra (2)

- Los estadísticos han desarrollado métodos basados en un modelo simple:
  - Si *razonablemente asumimos* que los datos han sido obtenidos mediante un muestreo aleatorio de una población infinita. Analizamos estos datos y hacemos inferencias sobre la población.

Población y muestra (3)

- No siempre es tan ideal.
  - En un experimento típico no siempre tomamos una muestra de una población, pero queremos extrapolar desde nuestra muestra a una situación más general.
  - En esta situación aún podemos usar el concepto de población y muestra si definimos la muestra como los 'datos recogidos' y la población como los datos que habríamos recogido si repitiéramos el experimento un número infinito de veces.

La importancia de la independencia de las observaciones.

- No solo es necesario que los datos provengan de una población. También es necesario que cada sujeto, (cada observación) sea 'escogido' independientemente del resto.
- Ejemplos:
  - Si realizas un experimento biomédico 3 veces, y cada vez por triplicado, no tenemos 9 valores independientes. Si promediamos los triplicados, entonces tenemos 3 valores medios independientes.
  - Si en un estudio clínico muestreamos 10 pacientes de una clínica y otros 10 de un Hospital. No hemos muestreado 20 individuos independientes de la población. Probablemente hemos muestreado dos poblaciones distintas.

Cómo usamos la estadística para extrapolar datos de nuestra muestra a la población.

- Hay tres enfoques básicos.

1. El primer método consiste en asumir que la población sigue un distribución especial conocida como **Normal** o Gaussiana (campana de Gauss).
   - Los tests te permiten hacer inferencias sobre la media (y tb sobre otras propiedades).
   - Los tests más conocidos pertenecen a este enfoque.
   - También se conoce como enfoque paramétrico.

Cómo usamos la estadística para extrapolar datos de nuestra muestra a la población (2)

2. El segundo enfoque consiste en ordenar los valores y ordenarlos de mayor a menor (rangos) y comparar distribuciones de rangos.
   - Es el principio básico de los tests no-paramétricos.
3. El tercer enfoque es conocido como 'Resampling'.
   - *Se escapa a los objetivos de este curso*.

Limitaciones de la estadística y sentido común.

- La idea básica de las estadística es extrapolar desde los datos recogidos para llegar a conclusiones más generales sobre la población de la que proceden los datos.
  - El problema es que solo podemos aplicar las inferencias a la población de la que hemos obtenido las muestras .
  - Generalmente queremos ir mas lejos .

Limitaciones de la estadística y sentido común. (2)

- Desafortunadamente la estadística no nos puede ayudar con estas extrapolaciones.
  - En este caso se necesita juicio científico y sentido común para hacer inferencias más allá de las limitaciones de la estadística.
  - El análisis estadístico es solo parte de la interpretación de nuestros datos.

La distribución Normal

- La distribución normal tiene características matemáticas especiales que hacen la base de la mayoría de los test estadísticos. La razón, el TCL.
- TCL: Teorema central de limite
  - Si tus muestras son suficientemente grandes, la distribución de las medias seguirá una distribución Normal.

La distribución Normal (2)

- TCL, de una forma más sencilla
  - Dada una población con una distribución cualquiera.
  - Aleatoriamente obtenemos varias muestras de esa población. Calculamos sus medias.
  - Construimos un histograma de la distribución de frecuencias de las medias.
- Esta dist. de medias sigue una 'Normal'.
- suficientemente grandes: Depende de cuan distinta sea la distribución de una normal.

Intervalos de Confianza.

- La media que calculamos de una muestra probablemente no sea igual a la media de la población.
- El tamaño de la diferencia dependerá del tamaño y de la variabilidad de la muestra. (n, σ).
- Combinamos estos dos factores, tamaño de la muestra, n, y variabilidad, σ, para calcular intervalos de confianza a un determinado nivel de confianza, generalmente 95%.

$$IC(95\%) \sim n, \sigma$$

Intervalos de Confianza (2).

- Si asumimos que tenemos una muestra aleatoria de una población. Podemos estar seguros al 95% de que el IC (95%) contiene a la media poblacional.
- De otra manera dicho: Si generamos 100 IC(95%) de una población, se espera que contengan la media poblacional en 95 casos y no lo haga en 5.
- No sabemos cuál es la media poblacional, así que no sabemos cuándo esto ocurre.

Intervalos de Confianza (3).

- Que asumimos cuando interpretamos un IC para una media:
  - Muestreo aleatorio de una población.
  - La población se distribuye, aproximadamente, como una Normal.
    - Si n es grande no es tan importante está condición.
  - Existe independencia de las observaciones.
  - Si se viola alguna de estas condiciones, el IC real es más 'ancho' que el calculado.

Intervalos de Confianza. (4)

- Un intervalo de confianza es un rango de valores (calculado en una muestra) en el cual se encuentra el verdadero valor del parámetro, con una probabilidad determinada.
- Nivel de confianza (1-α): probabilidad de que el verdadero valor del parámetro se encuentre en el intervalo.
- Nivel de significación (α): prob de equivocarnos.
- Normalmente 1-α =95% (α =5%)

¿Por que el 95%?

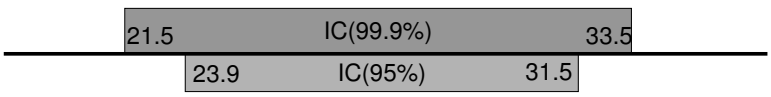
- No hay nada especial con el 95%, se pueden calcular IC para cualquier grado de confianza.
- A más confianza más grande el intervalo.
  - "*Lo que se gana en seguridad se pierde en precisión*".

¿Por que el 95%? (2)

- Ejemplo: De 720 estudiantes entrevistados, 198 dijeron 'si'. Así que tenemos 198/720=0.275, ed un 27.5%.
- La proporción está en torno al 27.5%. Para cuantificar ese entorno se añaden los IC.

  IC(95%): 27.5%±3.6%, ed (23.9%, 31.1%).
- Hablamos pues de estimación puntual y de margen de error.

  IC(99.9%): 27.5% ± 6%, ed (21.5%, 33.5%).



¿Que es un P-valor?

- Observar medias muestrales diferentes no es suficiente evidencia de que sean diferentes las medias poblacionales.
- Es posible que sean iguales y que la diferencia observada se deba a una coincidencia.
- Nunca se puede estar seguro. Lo único que se puede hacer es calcular las probabilidades de equivocarnos.

¿Que es un P-valor? (2)

- Si las poblaciones tienen la misma media realmente: ¿Cuál es la probabilidad de observar una diferencia tan grande o mayor entre las medias muestrales?.
- El P valor es una probabilidad de 0 a 1.
- Si p es pequeño, podemos concluir que la diferencia entre muestras ('*probablemente*') no es debida al azar.
  - Concluiríamos que tiene distintas medias.

¿Que es un P-valor? (3)

- Otra forma de ver lo mismo.
  - Los estadísticos hablan de hipótesis nula (H_o).
  - La hipótesis nula dice/es que 'no hay diferencia entre las medias'.
  - La hipótesis nula es lo contrario que la hipótesis experimental.
  - Así podemos definir p-valor como *la probabilidad de observar una diferencia tan grande o mayor que la observada si la hipótesis nula fuera cierta*.

¿Que es un P-valor? (4)

- Se dice que rechazamos la hipótesis nula si $p<0.05$, la diferencia es estadísticamente significativa (ss).
- Si $p>0.05$, no rechazamos la hipótesis nula y decimos que la diferencia no es estadísticamente significativa (ns).
- No podemos decir que la H_0 sea verdad, simplemente no la rechazamos, es decir: No tenemos suficiente evidencia para rechazar la hipótesis de igualdad (*ed la* H_0).

Resultados significativos en la ciencia.

- El termino Significativo es muy atractivo para los científicos.
- Es muchas veces malinterpretado.
- El significado el lenguaje natural no es el mismo.
  - Significativo en estadística ≠ importante.
  - Significativo en estadística ≠ interesante.
- Si α=0.05, un resultado se dice significativo cuando ocurre menos del 5% de las veces si la poblaciones fueran realmente idénticas.

Resultados significativos en la ciencia. (2)

- Si un resultado es estadísticamente significativo hay dos posibles explicaciones:
  1. Las poblaciones son idénticas (!!). Por casualidad has obtenido valores mayores en una muestra que en la otra. Encontrar un resultado sig aquí se llama cometer error de tipo I.
     - Si el nivel de significación es del 5% ed P<0.05, se comete este error en el 5% de los experimentos donde no hay diferencia real.

Resultados significativos en la ciencia. (3)

2. Las poblaciones son realmente diferentes. La conclusión es correcta.
   - La diferencia puede ser suficientemente grande para ser científicamente significativa o importante o pequeña y es trivial.

- Se llega así al concepto de 'Potencia Estadística'.

Potencia estadística.

- Cuando un experimento concluye diciendo que no se ha encontrado una 'diferencia significativa', no implica que no haya diferencia!!
- Simplemente no la hemos encontrado.
- Posibles causas:
  - Tamaño de la muestra.
  - Alta variabilidad.
- Si esta es la razón estamos cometiendo **Error de tipo II**.

Potencia estadística.(2)

- ¿Cuánta potencia tiene nuestro análisis para encontrar hipotéticas diferencias en el caso de existir estas?.
- La potencia depende del tamaño (n) y de la cantidad de variación de la muestra (desviación estándar o típica).
- También influye el tamaño de lo que es para nosotros una diferencia, o *¿Cuánto han de diferir para considerarlos distintos?*.
  - $n=(\sigma*1.96/d)^2$, $z_{1-\alpha/2}=1.96$, $\alpha=0.05$

El problema de las comparaciones múltiples

- Interpretar un p-valor es sencillo.
- Si la hip nula es cierta, hay un 5% de posibilidades de que una selección aleatoria de sujetos muestre una diferencia tan grande (o mayor) como la que muestran.

El problema de las comparaciones múltiples. (2)

- Interpretar muchos p-valores a la vez puede no ser tan sencillo.
- Si testeamos diferentes hip nulas (independientes) a la vez, con un nivel de significación del 0.05 hay más de un 5% de probabilidades obtener un resultado significativo por azar.

El problema de las comparaciones múltiples. (3)

- SI hacemos 3 test con $\alpha=0.05$.
- La probabilidad de no cometer error de tipo I es 0.95 (95%) para cada test.
- Suponemos que los tests son independientes.
- La probabilidad General de **NO** cometer Error de tipo I es: $(0.95)^3=0.875$.
- Así la probabilidad General de cometer Error de tipo I es $1-(0.95)^3=1-0.875=0.143$, ed 14.3%
- Se ha incrementado de 5% a 14.3% !!

El problema de las comparaciones múltiples. (4)

Nº hip nulas Independientes	Prob de obtener al menos un p valor<0.05 por azar	Significación para mantener error tipo I=0.05
1	5%	0.05
2	10%	0.0253
3	14%	0.0170
4	19%	0.0127
….	….	….
100	99%	0.0005
N	$100(1-0.95^{N})$	$1-0.95^{(1/N)}$

Valores atípicos (Normalización).

- Cuando analizamos los datos encontramos valores que están alejados del resto, los llamamos outlieres o atípicos.
- Cuando se encuentran outliers se tiende a eliminarlos. Esto es un error.
- Dos posibles orígenes:
  - Se debe al azar. Por lo tanto hay que respetarlo ya que proviene de la misma distribución que los otros valores.
  - Se debe a un error. Hay que eliminarlo.

Valores atípicos (Normalización).(2)

- Nunca podemos estar seguros, pero si podemos calcular la probabilidad de obtener un valor tan distinto de los otros por casualidad. Suponiendo que la población provenga de una normal.
  - Probabilidad es pequeña: podemos concluir que se debe a un error. Estamos justificados para eliminarlo.
  - Cómo encontramos outliers: 'Normalizando'.

Valores atípicos (Normalización).(3)

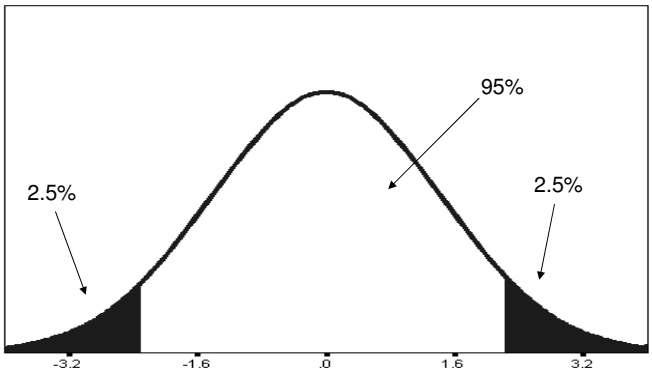
- Cuantifica 'como' de lejos está un valor de la media de su muestra.

$$z = \frac{x - \mu}{\sigma}$$

- El *z-score* expresa la divergencia del resultado experimental *x* con el valor mas probable μ como un múltiplo de las desviación típica.
- Cuando mas grande el *z-score*, menos probable es que el valor experimental se deba al azar.

Valores atípicos (Normalización).(4)

Gráficamente: para una normal tipificada, un intervalo de confianza del 95% se puede representar como:



La probabilidad de que una variable normal tipificada tome valores en el intervalo [-1.96,1.96] es del 95%.

Valores atípicos (Normalización). (5)

- Si conoces la media y la desviación típica de la población. Podemos sospechar de los z-scores mayores de 1.96 (esto es la base del control de calidad).

Apendice A (tema 3)

Un ejemplo de Análisis

Ejemplo Análisis

- Pasos en un análisis.
  - Validación de datos.
  - Descripción de los datos.
  - Formulación de hipótesis.

Ejemplo de análisis (1)

Tratamiento	id	N	var_f	var_num	var_e	var_ef	var_g
1	6	31	1	16	20,700	12,330	,291
1	3	32	1	15	18,100	10,120	,299
1	5	35	7	14	17,200	9,800	,301
1	2	37	0	15	16,700	9,070	,305
1	8	37	1	20	27,600	10,840	,289
1	2	37	4	18	23,700	13,440	,287
1	4	37	12	12	14,300	7,380	,318
1	3	38	6	21	28,100	17,070	,279
1	7	38	2	11	11,500	7,250	,319
1	7	40	5	8	9,200	4,730	,356
2	1	40	6	26	54,560	17,390	,279
2	2	40	6	27	62,850	19,510	,276
1	4	41	0	12	12,400	8,360	,310
1	5	42	0	25	36,100	20,050	,275
2	2	43	2	22	34,000	15,280	,283
2	10	43	2	18	22,690	12,750	,289
1	1	44	12	12	14,400	8,330	,310
2	A2	44	4	15	16,300	10,080	,300
2	7	44	0	11	11,470	5,830	,336
2	3	44	3	17	20,610	12,490	,290

Ejemplo de análisis (2)

- Una primera aproximación a los parámetros de la población son los estadísticos

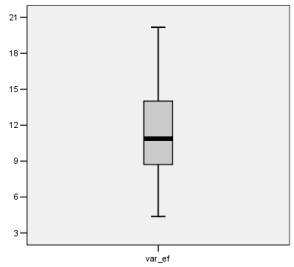
Estadísticos

var_ef

N	Válidos	48
	Perdidos	0
Media		11,75667
Mediana		10,87500
Desv. típ.		4,173449
Asimetría		,444
Error típ. de asimetría		,343
Curtosis		-,446
Error típ. de curtosis		,674
Mínimo		4,390
Máximo		20,170

Ejemplo de análisis (3)

- SPSS nos permite ver lo valores extremos

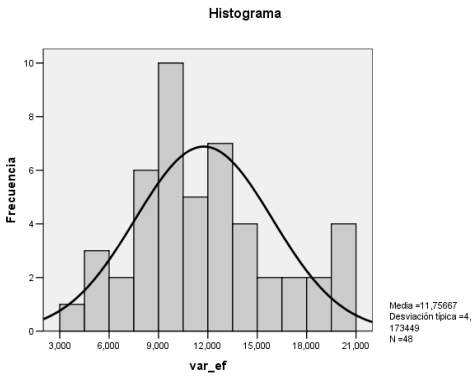


Valores extremos

			Número del caso	Valor
var_ef	Mayores	1	23	20,170
		2	14	20,050
		3	47	19,860
		4	12	19,510
		5	42	18,770
	Menores	1	25	4,390
		2	10	4,730
		3	40	5,260
		4	19	5,830
		5	9	7,250

Ejemplo de análisis (4)

- La siguiente aproximación interesante podría ser un histograma.



Ejemplo de análisis (5)

- Para rematar podríamos hacer un test sobre la normalidad.

Sobre este test volveremos a hablar

Prueba de Kolmogorov-Smirnov para una muestra

		var_ef
N		48
Parámetros normales [a,b]	Media	11,75667
	Desviación típica	4,173449
Diferencias más extremas	Absoluta	,101
	Positiva	,101
	Negativa	-,065
Z de Kolmogorov-Smirnov		,701
Sig. asintót. (bilateral)		,709

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

P-valor

p=0.709>0.05 !!

No es significativo

Ejemplo de análisis (6)

- Un p-valor representa la probabilidad de encontrar una diferencia tan grande o mayor asumiendo la hipótesis nula (H_0).
- Cual es nuestra hipótesis nula en el test K-S?
  - H_0 = No hay diferencia entre nuestra distribución y una distribución normal con esa media y esa SD.
  - **P>0.05**, lo que significa que es probable lo que se observa suponiendo H_0 cierta.
  - Por lo que, No rechazamos la hipótesis de igualdad.

Ejemplo de análisis (7)

- Recapitulamos.
  - No hay indicios de errores de medición, no creemos que haya valores atípicos.
  - No rechazamos la hipótesis nula de normalidad.

Ahora debemos plantear nuestra hipótesis

Ejemplo de análisis (8)

- Queremos estudiar posibles diferencias entre los tratamientos.
  - Descriptivos según tratamiento.

Informe

var_ef

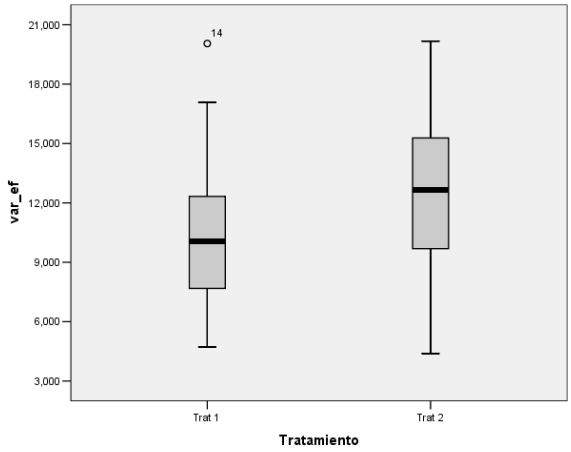
Tratamiento	Media	N	Desv. típ.	Mediana
Trat 1	10,42500	22	3,662014	10,05000
Trat 2	12,88346	26	4,312720	12,66000
Total	11,75667	48	4,173449	10,87500

Ejemplo de análisis (9)

- Ojo!: Como queremos estudiar posibles diferencias entre las medias de los tratamientos, nuestra hipótesis nula es la igualdad de medias.

- H_0= las medias son iguales.
- Hipótesis alternativa o experimental (H_a)= las medias no son iguales.

Ejemplo de análisis (10)



Ejemplo de análisis (11)

Sobre estos test volveremos a hablar

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Tratamiento	Estadístico	gl	Sig.	Estadístico	gl	Sig.
var_ef	Trat 1	,137	22	,200*	,943	22	,223
	Trat 2	,086	26	,200*	,969	26	,608

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

p>0.05, no sig

No rechazamos Ho de normalidad

Prueba de homogeneidad de la varianza

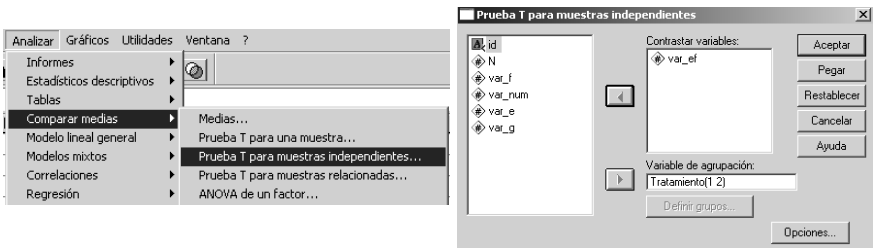
		Estadístico de Levene	gl1	gl2	Sig.
var_ef	Basándose en la media	1,141	1	46	,291
	Basándose en la mediana.	1,167	1	46	,286
	Basándose en la mediana y con gl corregido	1,167	1	45,991	,286
	Basándose en la media recortada	1,212	1	46	,277

p>0.05, no sig

No rechazamos Ho de HOV

Ejemplo de análisis (12)

- Estamos en las hipótesis para aplicar u t-test para muestras independientes.



Ejemplo de análisis (13)

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
var_ef	Se han asumido varianzas iguales	1,141	,291	-2,107	46	,041	-2,458462	1,167050	-4,807611	-,109312
	No se han asumido varianzas iguales			-2,136	45,998	,038	-2,458462	1,151055	-4,775419	-,141504

p-valor<0.05, es Significativo, podemos afirmar, con un riesgo de equivocarnos de un 5%, que las medias son diferentes entre los tratamientos

IC (95%), rango de valores en el cual se encuentra el verdadero valor del parámetro, diferencia de las medias, con un 95% de probabilidad.

Rechazamos la hipótesis nula de igualdad

TEMA 4

Normalidad y Homocedasticidad

Tema 4
Normalidad y Homocedasticidad

- Normalidad
- Homocedasticidad
- Hipótesis iniciales.
  - KS y Levene con SPSS
- Transformación de variables.

Normalidad y Homocedasticidad

- Normalidad

Normalidad.

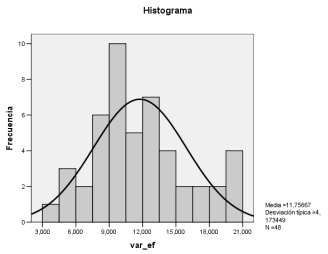
- La hipótesis de normalidad.
  - No hay diferencias entre nuestra distribución y una distribución normal con esa media y esa SD.

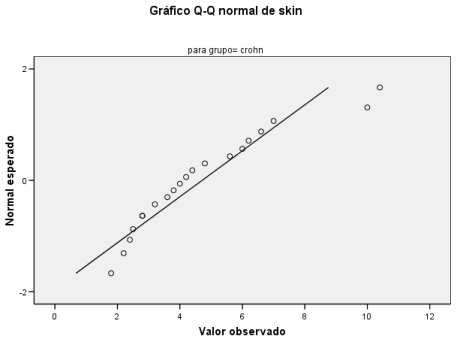
Normalidad (2)

- La hipótesis de Normalidad puede testearse de forma gráfica y/o de forma analítica.

Normalidad. (3)

- Gráficos para observar la normalidad son:
  - el histograma.
  - gráfico q-q.
  - Otros gráficos.





Contraste de asimetría y curtosis

- Contrastes de normalidad son:
  - contraste de asimetría y curtosis,
  - contraste chi-cuadrado,
  - contraste de Kolmogorov-Smirnov-Lilliefors

Contraste de asimetría y curtosis. (2)

- Sobre la Asimetría y la Curtosis.
  - Regla aproximativa.
    - Curtosis y/o Coeficiente de asimetría entre -1 y 1, es generalmente considerada una muy ligera desviación de la normalidad.
    - Entre -2 y 2 tampoco es malo del todo, según el caso.

Contraste de asimetría y curtosis (3)

- Contraste sobre Asimetría y curtosis
  - Primero normalizamos los coeficientes:
    - $Z_s=S/SE_s$
    - $Z_k=K/SE_k$
      - Ojo, para ambas la media es 0.
    - Si alguno de los valores es mayor en valor absoluto que 1.96 se considera significativo.
    - $|Z_s|>1.96$, significativo. Y consideramos que hay demasiada Asimetría,(en su caso Curtosis.)
  - En muestras grandes, es más interesante mirar a la distribución de forma visual (histograma).

Contraste de asimetría y curtosis (4).

Estadísticos

		Hygiene (Day 1 of Glastonbury Festival)	Hygiene (Day 2 of Glastonbury Festival)
N	Válidos	810	264
	Perdidos	0	546
Media		1,7934	,9609
Desv. típ.		,94449	,72078
Asimetría		8,865	1,095
Error típ. de asimetría		,086	,150
Curtosis		170,450	,822
Error típ. de curtosis		,172	,299

Z_S=1.095/0.15=7.3 >1.96, significativo, hay asimetría.

Kolmogorov-Smirnov-Lilliefors

- Mucho más sencillo es hacer un test de K-S

P<0.05, Significativo, Rechazamos H0

No hay normalidad

Prueba de Kolmogorov-Smirnov para una muestra

		Hygiene (Day 1 of Glastonbury Festival)
N		810
Parámetros normales[a,b]	Media	1,7934
	Desviación típica	,94449
Diferencias más extremas	Absoluta	,083
	Positiva	,083
	Negativa	-,063
Z de Kolmogorov-Smirnov		2,354
Sig. asintót. (bilateral)		,000

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

Kolmogorov-Smirnov-Lilliefors (2)

- En SPSS tb nos encontramos con esta salida.

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	transformacion	Estadístico	gl	Sig.	Estadístico	gl	Sig.
variable	crudo	,441	7	,000	,526	7	,000
	ln	,256	7	,183	,842	7	,103
	log	,256	7	,183	,842	7	,103

a. Corrección de la significación de Lilliefors

EL test de Shapiro-Wilk no se debe aplicar con n>50
Con muestras grandes es preferible el test de Kolmogorov-Smirnov

Normalidad. (3)

- La falta de normalidad influye en el modelo en:
  - Los estimadores mínimo-cuadráticos no son eficientes (de mínima varianza).
  - Los intervalos de confianza de los parámetros del modelo y los contrastes de significación son solamente aproximados y no exactos.
  - En otras palabras: Se pierde precisión.
  - EL TCL, nos permite '*saltarnos*' esta hipótesis para muestras suficientemente grandes..

Normalidad. (4)

- Causas que originan falta de normalidad:
  - Existen observaciones heterogéneas:
    - errores en la recogida de datos;
    - el modelo especificado no es correcto (por ejemplo, no se ha tenido en cuenta una variable de clasificación cuando las observaciones proceden de diferentes poblaciones).
  - Hay observaciones atípicas. (estimadores robustos)
  - Existe asimetría en la distribución. (transformación de Box-Cox, HEV)

- Homocedasticidad

Homocedasticidad

- Homocedasticidad= Homogeneidad de varianzas = **HOV**
- La varianza es constante (no varía) en los diferentes niveles del factor.
- La falta de homocedasticidad se denomina heterocedasticidad =**HEV**.

Homocedasticidad. (2)

- Si el diseño es balanceado ($n_i = n_j$, i, j = 1,..,I), la heterocedasticidad no afecta tanto a la calidad de los contrastes, a no ser que la varianza de la respuesta para algún grupo particular sea considerablemente mayor que para otros.
  - Balanceadas y HEV, OK si

$$\frac{\hat{S}^2_{Max}}{\hat{S}^2_{Min}} < 3$$

Si no hay balanceo esta regla funciona con 2.

Homocedasticidad.(3)

- Si los tamaños muéstrales son muy distintos, se verifica que:
  - Si los grupos con tamaños muéstrales pequeños tienen mayor varianza la probabilidad de cometer un error de **tipo I** en las pruebas de hipótesis será menor de lo que se obtiene y los niveles de confianza de los intervalos serán inferiores a lo que se cree.→ Conservador
  - Si los tratamientos con tamaños muestrales grandes tienen mayor varianza, entonces se tendrá el efecto contrario y las pruebas serán mas liberales.

Ojo!: Es peor que los grandes la tengan grande

Homocedasticidad. (4)

- Contrastes para contrastar la Homogeneidad de varianzas
  - El contraste de **Levene** es el más utilizado para contrastar HOV.
    - Se puede hacer sobre la media, mediana o sobre otras medidas de tendencia central.
  - F-Test. (Excel)

Homocedasticidad. (5)

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
Numeracy	Basándose en la media	12,350	1	98	,001
	Basándose en la mediana.	13,025	1	98	,000
	Basándose en la mediana y con gl corregido	13,025	1	80,252	,001
	Basándose en la media recortada	12,527	1	98	,001

P<0.05, Significativo, rechazamos H_0, ed
No alcanzamos el supuesto de HOV, HEV

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
Percentage of lectures attended	Basándose en la media	,545	1	98	,462
	Basándose en la mediana.	,598	1	98	,441
	Basándose en la mediana y con gl corregido	,598	1	97,401	,441
	Basándose en la media recortada	,554	1	98	,458

P>0.05, No Significativo, alcanzamos el supuesto de HOV

Homocedasticidad. (6)

- En Excel con el F-Test.

Prueba F para varianzas de dos muestras

Prueba F para varianzas de dos muestras

	Variable 1	*Variable 2*
Media	7,71875	7,073170732
Varianza	5,316468254	7,369512195
Observaciones	64	41
Grados de libertad	63	40
F	0,721413862	
P(F<=f) una cola	0,121149706	
Valor crítico para F (una co	**0,630961949**	

P>0.5, NO sig

Var(mayor)/var(menor)
1,386166877 No hay 'grandes' diferencias entr elas varianzas

Aceptamos la hip. de igualdad de varianzas

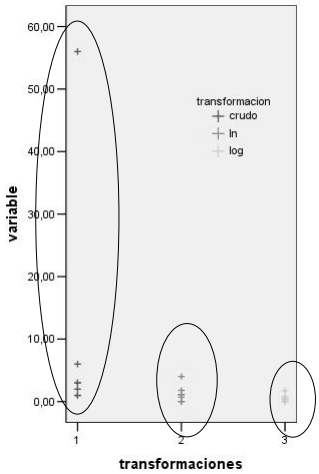
Transformación de variables

- Posibles soluciones a la falta de HOV:
  - Transformaciones Y=Ln(X), Y=1/X, $Y=X^{1/2}$, $Y=1/X^{1/2,}$ (Box-Cox).
  - La heterocedasticidad suele ir unida a la falta de normalidad.

Transformación de variables (2)

SD=20.23

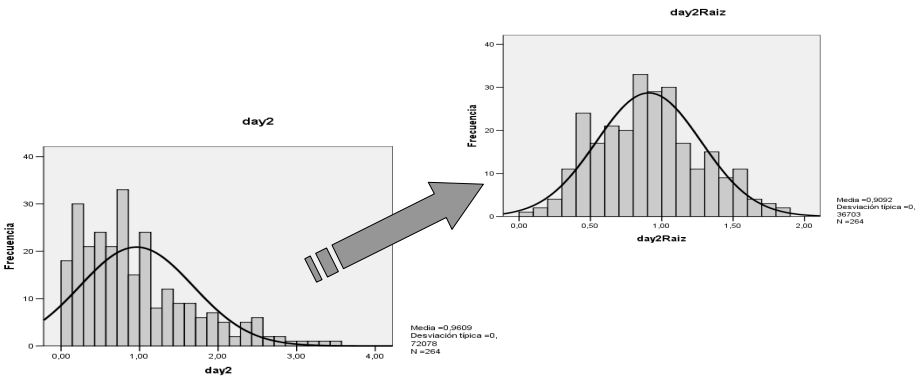
SD=1.38



SD=0.60

Transformación de variables (3)

- Otro ejemplo de transformación.



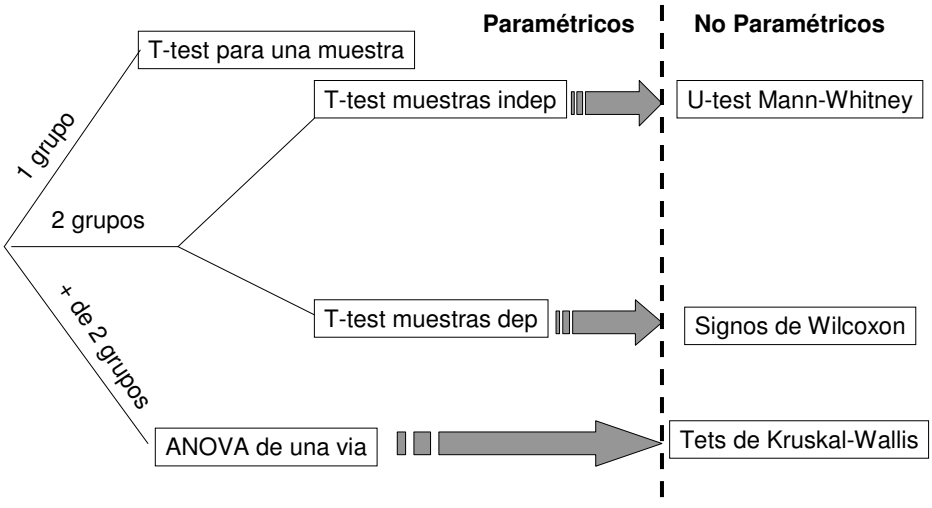
Hipótesis iniciales.

- La violación de los supuestos paramétricos no invalida el análisis, ya que estas pruebas suelen ser lo suficientemente robustas para no verse seriamente afectadas por ligeras violaciones de los supuestos paramétricos.
- Existe clara evidencia empírica, de que las pruebas con una sola variable dependiente (por ejemplo ANOVA) son altamente robustas bajo la violación de normalidad y homocedasticidad.
- La principal, excepción, la tenemos cuando las muestras son muy pequeñas (menores de 10) y desiguales.

Hipótesis iniciales. (2)

- La situación se complica cuando pasamos a los métodos multivariantes.
  - La tendencia actual está en considerar que en muestras grandes (n> 30) los análisis multivariantes son lo suficientemente robustos como para ser insensibles a ligeras desviaciones de los supuestos paramétricos, principalmente el normalidad multivariante y de la homocedasticidad.

¿Que test elegir?.



¿Que test elegir?. (2)

- Datos categóricos → Chi Cuadrado

  McNemar (Antes/después)

TEMA 5

Correlación Lineal

Índice

- Correlación bivariada
- Correlaciones parciales

Objetivo

- Dos grandes objetivos del Análisis de datos
  - Comparar grupos
  - Estudiar relaciones entre grupos
    - *Queremos ser capaces de cuantificar estas relaciones : Grado de relación existente entre 2 variables (índices estadísticos).*

Correlación lineal simple

- El concepto de *relación o correlación* se refiere al grado de variación conjunta entre dos o más variables.
  - Nos vamos a centrar en:
    - *relaciones lineales*
    - *entre dos variables: simple o bivariada.*

(más de 2 variables => concepto de 'regresión')

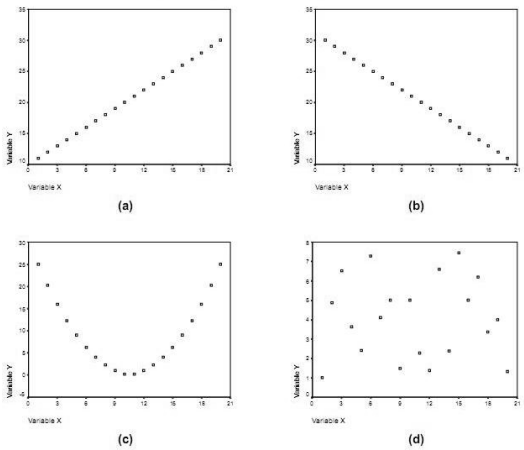
Correlaciones positivas y negativas

- Correlación/relación lineal positiva entre dos variables X e Y, indica que las dos variables varían de forma parecida: *los sujetos que puntúan alto en X puntúan alto en Y.*
- Correlación/relación lineal negativa: varían justamente al revés: *los sujetos que puntúan alto en X puntúan bajo en Y.*

Diagrama de dispersión

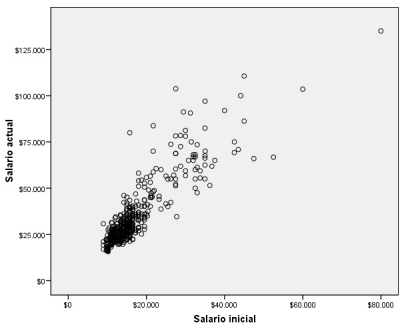
- Primera impresión: *diagrama de dispersión*
  - Directo
  - Intuitivo
    - Un *diagrama de dispersión (Scatter plot)* es un gráfico en que una de las variables se coloca en el eje de abscisas (X) y la otra en el de ordenadas (Y) y los pares (xi, yi) se represntan como una **nube de puntos**.

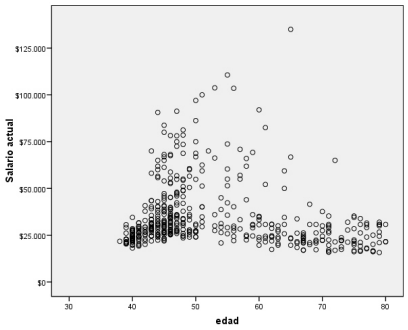
Diagrama de dispersión (2)



Necesidad de cuantificar

- No siempre tenemos relaciones tan 'claras'





Coeficientes de correlación

Coeficientes de correlación: índices numéricos que sirven para cuantificar el grado de relación lineal entre dos variables. También sirven para valorar el grado de ajuste de una nube de puntos a una línea recta.

- Pearson
- Tau-b de Kendall
- Sperman

r de Pearson

- Coef de correlación de Pearson (1896)
  - Es quizás el mejor
  - Variables de tipo intervalo o razón.
  - Es el más utilizado
  - Se representa por 'r'
  - Toma valores entre -1 y 1
    - Si r=1 indica relación lineal perfecta positiva
    - Si r=-1 indica relación lineal perfecta negativa
    - r=0 indica relación lineal nula

$$r_{xy} = \frac{\sum x_i y_i}{n S_x S_y}$$

Nota: un coef de correlación alto no implica Causalidad.
(dos variables pueden estar altamente correlacionadas sin que ninguna sea causa de la otra)

Tau-b de Kendall

- Apropiado para estudiar relaciones entre variables ordinales.
  - Toma valores entre -1 y 1
    - Si b=1 indica relación lineal perfecta positiva
    - Si b=-1 indica relación lineal perfecta negativa
    - b=0 indica relación lineal nula
- Se basa en el concepto de inversión, no-inversión y empate.

$$\tau_b = (n_P - n_Q) \Big/ \sqrt{(n_P + n_Q + n_{E(X)})\,(n_P + n_Q + n_{E(Y)})}$$

- Cuando las variables no alcanzan el nivel de medida de Intervalo y no podemos suponer que la distribución poblacional conjunta de las variables sea normal.

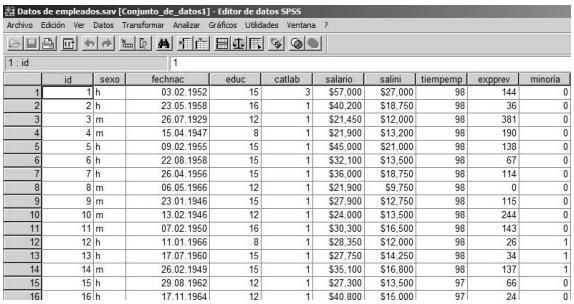
rho de Sperman

- El coeficiente de correlación de Sperman (1904) es el mismo que el de Pearson pero después de transformar las puntuaciones originales en rangos.
- Como en el caso de la tab-b de Kendall. Sperman puede utilizarse como una alternativa a Perason cuando las variables son ordinales y/o se incumple el supuesto de Normalidad.
  - Toma valores entre -1 y 1
    - Si b=1 indica relación lineal perfecta positiva
    - Si b=-1 indica relación lineal perfecta negativa
    - b=0 indica relación lineal nula

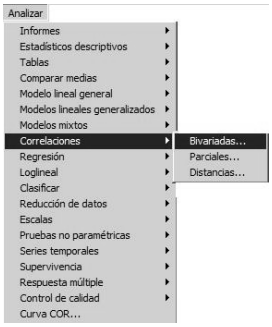
- Fuerza de la relación (Según Best and Khan, 1989).
  - Insignificante (r=0.00 a 0.2)
  - Baja (r=0.20 a 4.0)
  - Moderada (r=0.40 a 0.6)
  - Sustancial (r=0.60 a 0.8)
  - Alta o muy alta (r=0.80 a 1.00)

Ejemplo con SPSS

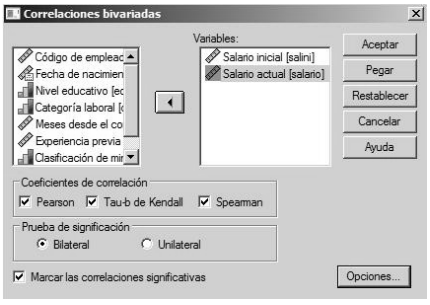
- Fichero: C:\Program Files\SPSS15\Datos de empleados.sav

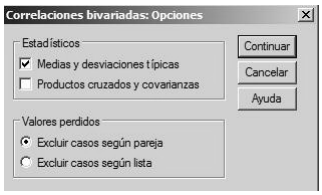


	id	sexo	fechnac	educ	catlab	salario	salini	tiempemp	exprev	minoría
1	1	h	03.02.1952	15	3	$57,000	$27,000	98	144	0
2	2	h	23.05.1958	16	1	$40,200	$18,750	98	36	0
3	3	m	26.07.1929	12	1	$21,450	$12,000	98	381	0
4	4	m	15.04.1947	8	1	$21,900	$13,200	98	190	0
5	5	h	09.02.1955	15	1	$45,000	$21,000	98	138	0
6	6	h	22.08.1958	15	1	$32,100	$13,500	98	67	0
7	7	h	26.04.1956	15	1	$36,000	$18,750	98	114	0
8	8	m	06.05.1966	12	1	$21,900	$9,750	98	0	0
9	9	m	23.01.1946	15	1	$27,900	$12,750	98	115	0
10	10	m	13.02.1946	12	1	$24,000	$13,500	98	244	0
11	11	m	07.02.1950	16	1	$30,300	$16,500	98	143	0
12	12	h	11.01.1966	8	1	$28,350	$12,000	98	26	1
13	13	h	17.07.1960	15	1	$27,750	$14,250	98	34	1
14	14	m	26.02.1949	15	1	$35,100	$16,800	98	137	1
15	15	h	29.08.1962	12	1	$27,300	$13,500	97	66	0
16	16	h	17.11.1964	12	1	$40.800	$15.000	97	24	0



Ejemplo con SPSS (2)





Ejemplo con SPSS (3)

Estadísticos descriptivos

	Media	Desviación típica	N
Salario inicial	$17,016.09	$7,870.638	474
Salario actual	$34,419.57	$17,075.661	474

Correlaciones

		Salario inicial	Salario actual
Salario inicial	Correlación de Pearson	1	**,880****
	Sig. (bilateral)		,000
	N	474	474
Salario actual	Correlación de Pearson	,880**	1
	Sig. (bilateral)	,000	
	N	474	474

**. La correlación es significativa al nivel 0,01 (bilateral).

Correlaciones no paramétricas

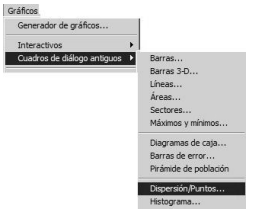
Correlaciones

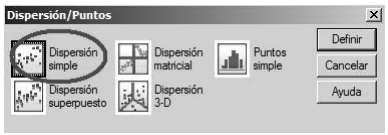
			Salario inicial	Salario actual
Tau_b de Kendall	Salario inicial	Coeficiente de correlación	1,000	**,656****
		Sig. (bilateral)	.	,000
		N	474	474
	Salario actual	Coeficiente de correlación	,656**	1,000
		Sig. (bilateral)	,000	.
		N	474	474
Rho de Spearman	Salario inicial	Coeficiente de correlación	1,000	**,826****
		Sig. (bilateral)	.	,000
		N	474	474
	Salario actual	Coeficiente de correlación	,826**	1,000
		Sig. (bilateral)	,000	.
		N	474	474

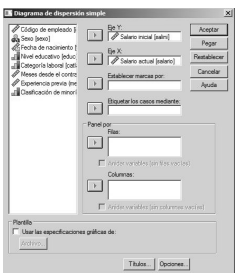
**. La correlación es significativa al nivel 0,01 (bilateral).

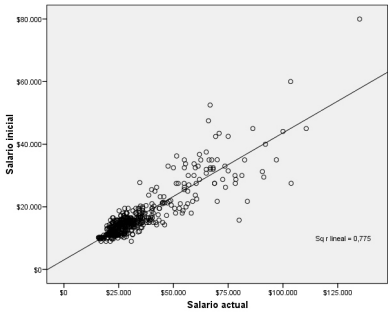
Ejercicio: añadir al análisis la variable '*Meses desde el contrato*'. ¿Qué sucede?

Ejemplo con SPSS (4)



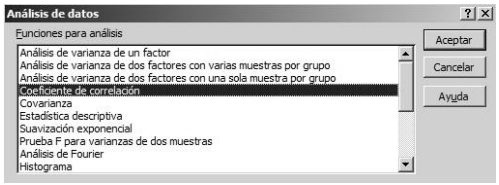


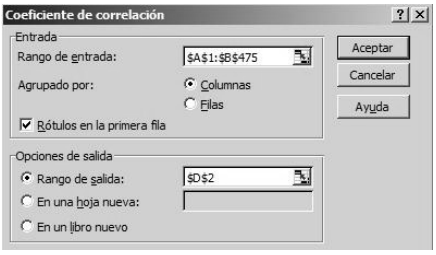




Ejemplo con MS-excel

Menú: Herramientas /Análisis de datos





	Salario actual	*Salario Inicial*
Salario actual	1	
Salario Inicial	0,880117466	1

¿Qué echamos de menos en comparación con la salida de SPSS?

Correlaciones parciales

- Permiten estudiar la relación lineal entre dos variables controlando el posible efecto de una o más variables extrañas.

Por ejemplo:

- Variables a estudiar: 'Inteligencia' y 'rendimiento escolar'
- Terceras variables: 'nº de horas de estudio', 'nivel educativo de los padres'.

orden

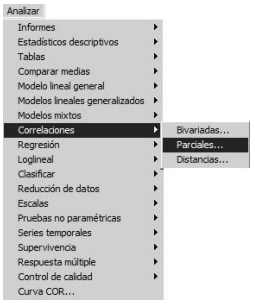
- La ecuación para obtener el coeficiente de correlación parcial depende del número de variables que estemos controlando (primer orden -> una variable, segundo orden -> dos variables, etc..)

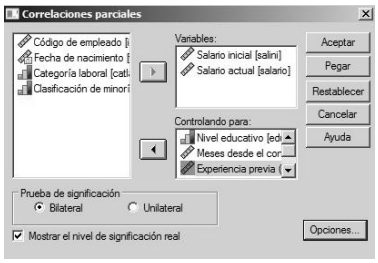
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad \textbf{(correlación parcial de primer orden)}$$

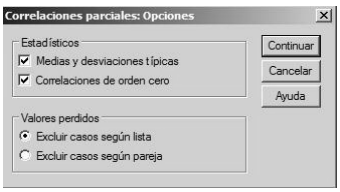
$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \qquad \textbf{(correlación parcial de segundo orden)}$$

- Los coeficientes de mayor orden se obtienen siguiendo la misma lógica.
- Cuando no controlamos ninguna variable, hablamos de *correlación de orden cero*, y es el coeficiente 'r' de correlación de Perason.

Ejemplo con SPSS







Corr. parciales

Estadísticos descriptivos

	Media	Desviación típica	N
Salario inicial	17016,09	7870,638	474
Salario actual	34419,57	17075,661	474
Nivel educativo	13,49	2,885	474
Meses desde el contrato	81,11	10,061	474
Experiencia previa (meses)	95,86	104,586	474

Ejemplo con SPSS (2)

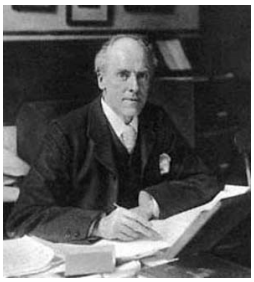
Correlaciones

Variables de control			Salario inicial	Salario actual	Nivel educativo	Meses desde el contrato	Experiencia previa (meses)
-ninguno-[a]	Salario inicial	Correlación	1,000	**,880**	**,633**	-,020	,045
		Significación (bilateral)	.	,000	,000	,668	,327
		gl	0	472	472	472	472
	Salario actual	Correlación	,880	1,000	**,661**	,084	-,097
		Significación (bilateral)	,000	.	,000	,067	,034
		gl	472	0	472	472	472
	Nivel educativo	Correlación	,633	,661	1,000	,047	**-,252**
		Significación (bilateral)	,000	,000	.	,303	,000
		gl	472	472	0	472	472
	Meses desde el contrato	Correlación	-,020	,084	,047	1,000	,003
		Significación (bilateral)	,668	,067	,303	.	,948
		gl	472	472	472	0	472
	Experiencia previa (meses)	Correlación	,045	-,097	-,252	,003	1,000
		Significación (bilateral)	,327	,034	,000	,948	.
		gl	472	472	472	472	0
Nivel educativo & Meses desde el contrato & Experiencia previa (meses)	Salario inicial	Correlación	1,000	**,812**			
		Significación (bilateral)	.	,000			
		gl	0	469			
	Salario actual	Correlación	,812	1,000			
		Significación (bilateral)	,000	.			
		gl	469	0			

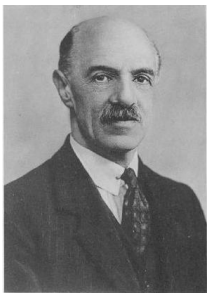
a. Las casillas contienen correlaciones de orden cero (de Pearson).

Controlando

Historia



Karl Pearson (Londres 1857- 1936) fue un prominente científico, matemático, historiador y pensador británico, que estableció la disciplina de la estadística matemática. Desarrolló una intensa investigación sobre la aplicación de los métodos estadísticos en la biología y fue el fundador de la bioestadística. Cofundó el '*Men and Women's Club*' cuya finalidad era permitir libre discusión entre hombres y mujeres.



Charles Edward Spearman (Londres,1863-1945). Psicólogo inglés. Realizó importantes aportes a la psicología y a la estadística, desarrollando el Análisis Factorial. Gracias a él propuso la existencia de un factor general de inteligencia (Factor G), que subyace a las habilidades para la ejecución de las tareas intelectuales.

TEMA 6

Regresión Lineal

Índice

- Introducción
  - Recta de regresión
  - La mejor recta de regresión
  - Bondad de ajuste
- Análisis de regresión lineal simple
- Análisis de regresión lineal múltiple
- Supuestos del modelo de regresión lineal

Introducción

- Técnica estadística par estudiar la relación entre variables.
- Se adapta a multitud de situaciones (inv social, inv de mercados, en fisica,…)
- Regresión lineal simple: relación entre dos variables.
- Regresión lineal múltiple: más de dos variables.

Introducción

- Se utiliza para explorar y cuantificar la relación entre una variable llamada dependiente o criterio (Y) y una o más variables llamadas independientes o predictoras (X_1, X_2,…X_k).
- Tb para construir una ecuación lineal con fines predictivos.

Introducción

- Existen una serie de procedimientos de diagnostico que nos informaran sobre la estabilidad e idoneidad del análisis (análisis de residuos, puntos influyentes,..)

Supuestos del modelo

IMPORTANTE

Recta de regresión

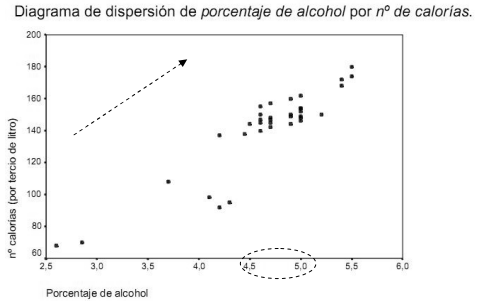
- Supongamos que queremos estudiar la relación entre el grado de alcohol de 35 marcas de cerveza y su contenido calórico.

–Partimos de un diagrama de dispersión

– ¿Como podríamos describir estos datos?

• Hablar del sentido de la relación (correcto pero poco específico).

• Listar todos los datos (preciso pero poco informativo).



Recta de regresión

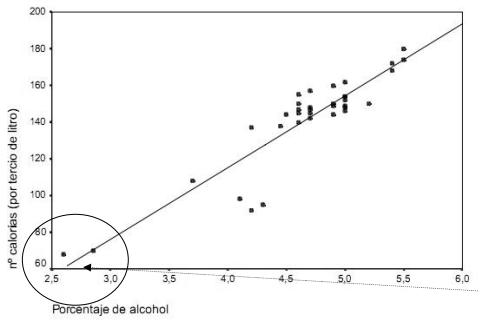
- Podríamos describir la pauta observada mediante una función matemática simple: una línea recta.

$$Y=B_0+B_1X$$

- Coeficiente B_1: cambio medio en el número de calorías (Y) por cada unidad de cambio que se produce en el porcentaje de alcohol (X)
- Coeficiente B_0: punto de corte de la recta con el eje vertical. (nº medio de calorías para una cerveza de 0º de alcohol)

Recta de regresión

Diagrama de dispersión y recta de regresión (*% de alcohol* por *nº de calorías*).



$$Y_i = -33{,}77 + 37{,}65\ X_i$$

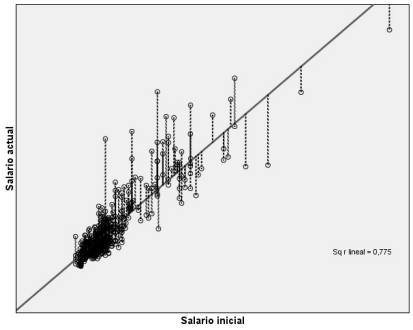
nº de calorías = − 33,77 + 37,65 (% de alcohol)

Y(0)=-33.77 calorías

• El origen de la recta aporta información sobre lo que podría ocurrir si extrapolamos hacia abajo la pauta observada.
•Al hacer esto estaríamos haciendo pronósticos más allá de lo que abarcan nuestros datos ← extremadamente arrisgado en el contexto del análisis de regresión

La mejor recta de regresión

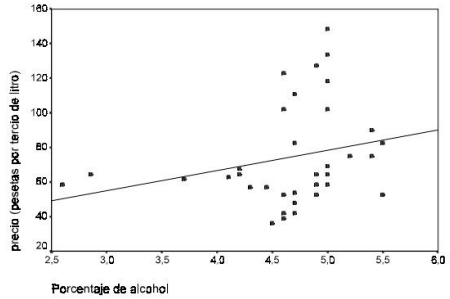
- Existen varios criterios para calcular la mejor recta de regresión.
- Tradicionalmente la recta más aceptada ha sido *la recta que hace mínima la suma de los cuadrados de las distancias verticales entre cada punto y la recta* (Ajuste por mínimos cuadrados)
  - Solo existe una y solo una recta que minimice los residuos al cuadrado.



Bondad de ajuste

- Falta un indicador del grado en que la recta se ajusta a la nube de puntos!!!
  - la mejor recta podría no ser buena

Diagrama de dispersión, recta de regresión y ajuste (*% de alcohol* por *precio*).



$$\text{precio} = 20{,}16 + 11{,}61\ (\%\ \text{de alcohol})$$

$$R^2 = 0{,}06$$

Bondad de ajuste

- Hay muchas maneras de cuantificar el ajuste, la medida que más aceptación tiene es el **Coeficiente de determinación** R^2.

$$R^2 = \frac{s_{XY}^2}{s_X^2 \, s_Y^2}$$

- R^2: el cuadrado del coeficiente de correlación múltiple.
- Toma valores entre -1 y 1.
- Representa el grado de ganancia que podemos obtener al predecir una variable basándonos en le conocimiento que tenemos de otra u otras variables.

Bondad de ajuste

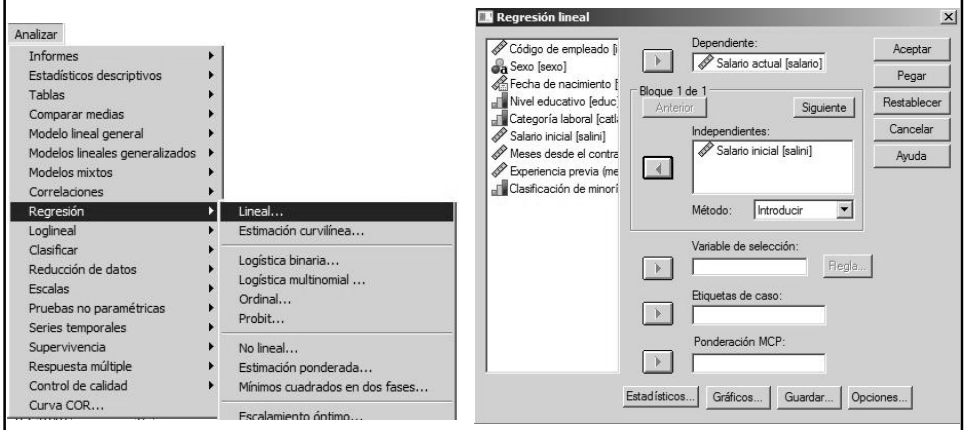
- En el ejemplo de las calorías y el grado de alcohol R^2=0.83

$$\text{nº de calorías} = -33{,}77 + 37{,}65\ (\%\ \text{de alcohol})$$

- Podemos mejorar nuestro pronostico si en lugar de usar como pronostico el número medio de calorías (modelo trivial) lo basamos en el % de alcohol.
- En el ejemplo de calorías y precio R^2=0.06.
- Parece evidente que las calorías están más relacionadas con el % de alcohol que con el precio.

Análisis de regresión simple con SPSS

- Fichero: C:\Program Files\SPSS15\Datos de empleados.sav



Análisis de regresión simple con SPSS (2)

- Bondad de ajuste. La primera tabla se refiere al coeficiente de correlación múltiple (que al tener solo dos variables es el coeficiente de correlación de Pearsón). Su cuadrado es el **coeficiente de determinación**.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,880[a]	,775	,774	$8,115.356

a. Variables predictoras: (Constante), Salario inicial

Nos indica que el 75% de la variación de *salario* está explicada por *salini*

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

$$R^2_{\text{corregida}} = R^2 - [p(1-R^2)/(n-p-1)]$$

Es una corrección a la baja de R^2 que se basa en el número de casos y de variables

Análisis de regresión simple con SPSS (3)

- El error típico de la estimación es la desviación típica de los residuos*.
- Representa una medida de la parte de la variabilidad de la variable dependiente que no es explicada por la regresión (cuanto mejor es el ajuste menor el SSE).

$$\text{Error típico de estimación} = S_e = \sqrt{\sum (Y_i - \hat{Y}_i)^2 / (n-2)}$$

• Debería ser marcadamente inferior a la desviación típica de la variable dependiente.

*Se= raíz cuadrad de la media cuadrática residual de la tabla ANOVA

Análisis de regresión simple con SPSS (3)

- La tabla ANOVA nos informa sobre si existe o no relación significativa entre las variables.
- Contrasta si *el valor poblacional de R es cero. En el modelo de regresión simple equivale a contrastar si la pendiente de la recta de reg vale cero.*

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,07E+011	1	1,1E+011	1622,118	,000[a]
	Residual	3,11E+010	472	65858997		
	Total	1,38E+011	473			

a. Variables predictoras: (Constante), Salario inicial

b. Variable dependiente: Salario actual

En nuestro caso sig<0.05, en consecuencia ambas variables están relacionadas linealmente.

Análisis de regresión simple con SPSS (4)

- Ecuación de regresión.

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Error típ.	Coeficientes estandarizados Beta	t	Sig.
1	(Constante)	1928,206	888,680		2,170	,031
	Salario inicial	1,909	,047	,880	40,276	,000

a. Variable dependiente: Salario actual

$$B_0 = \bar{Y} - B_1\bar{X}$$

$$B_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

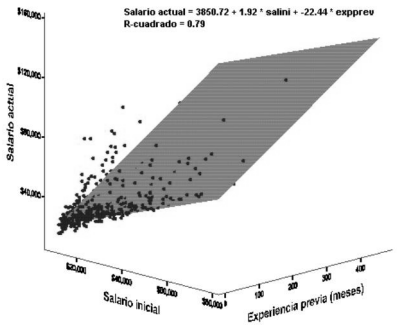
Pronóstico en salario = 1928.206 + 1.909 salini

Análisis de regresión simple con SPSS (5)

- Coeficientes de regresión estandarizados o coeficientes *Beta.*
- Se obtienen tras convertir las puntuaciones directas en típicas.
- En la regresión múltiple nos permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación.
- En regresión simple B_1= R de Pearson.

Regresión lineal múltiple

- La regresión lineal múltiple no define una recta en el plano, sino un hiperplano en un espacio multidimensional.



Con una variable dependiente y tres independientes necesitamos tres ejes para el correspondiente diagrama de dispersión. Y así sucesivamente.

Regresión lineal múltiple

- Con más de una variable independiente el diagrama de dispersión no resulta tan útil, ni intuitivo.
- Es más fácil partir de la ecuación del modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- Combinación lineal
- Coeficientes B_k que indican el peso relativo de esa variable en la ecuación.
- Epsilon: componente aleatorio, residuos.

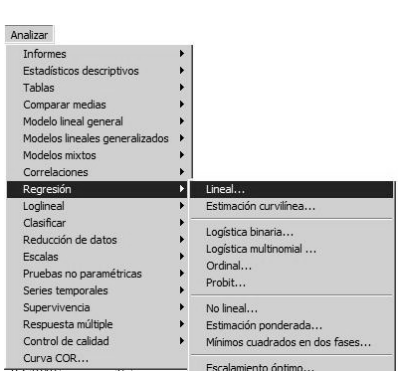
Regresión lineal múltiple

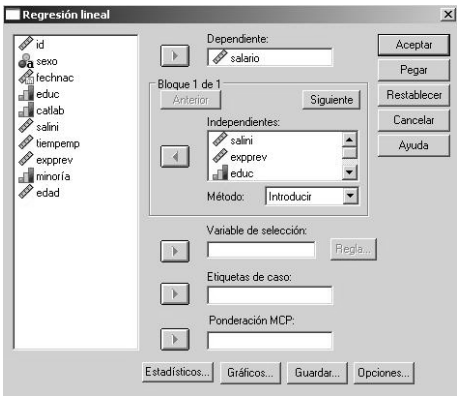
- Al igual que en la regresión lineal simple se ajusta la recta por el **método de los mínimos cuadrados**. Es decir, haciendo que las diferencias entre los valores observados y los pronosticado sean mínimas.

- Este modelo se basa en una serie de supuestos (linealidad, independencia, normalidad, homocedasticidad y no co-linealidad) que veremos después.

Regresión lineal múltiple con SPSS

- Fichero: C:\Program Files\SPSS15\Datos de empleados.sav





Regresión lineal múltiple con SPSS (2)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,895[a]	,802	,800	$7,631.917

a. Variables predictoras: (Constante), Nivel educativo, Experiencia previa (meses), Salario inicial

Las tres variables independientes explican un 80% de la varianza de la var dependiente

El Error típico de la estimación es mejor que el del análisis simple (8.115,35) ➔ mejoramos el ajuste

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1,11E+011	3	3,7E+010	632,607	,000[a]
	Residual	2,74E+010	470	58246157		
	Total	1,38E+011	473			

a. Variables predictoras: (Constante), Nivel educativo, Experiencia previa (meses), Salario inicial

b. Variable dependiente: Salario actual

sig<0,05, la ecuación de reg ofrece un buen ajuste a la nube de puntos

Regresión lineal múltiple con SPSS (3)

- Esta tabla contiene toda la información necesaria para construir la ecuación de regresión mínimo-cuadrática

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error típ.	Coeficientes estandarizados Beta	t	Sig.
1	(Constante)	-3661,517	1935,490		-1,892	,059
	Salario inicial	1,749	,060	,806	29,198	,000
	Experiencia previa (meses)	-16,730	3,605	-,102	-4,641	,000
	Nivel educativo	735,956	168,689	,124	4,363	,000

a. Variable dependiente: Salario actual

$$Pronóstico\ en\ salario\ =$$
$$= -3.661{,}517 + 1{,}749\ salini - 16{,}730\ expprev + 735{,}956\ educ$$

¡Nota!: Estos coeficientes no son independientes entre si (*coef de reg parcial*). El valor estimado para cada coef se ajusta teniendo en cuenta la presencia del resto de variables independientes.

Regresión lineal múltiple con SPSS (4)

- Los coeficientes beta están basados en puntuaciones típicas y por tanto son comparables entre si.
- Indican la cantidad de cambio que se producirá en la variable dependiente (Y) por cada cambio de una unidad en la correspondiente variable independiente (manteniendo constante el resto de variables independientes).
- Son una 'pista' muy útil sobre la importancia relativa de cada variable independiente en la ecuación.
- En nuestro ejemplo la variable con mayor peso relativo es *salini*, después *educ* y por último *expprev*.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-3661,517	1935,490		-1,892	,059
	Salario inicial	1,749	,060	,806	29,198	,000
	Experiencia previa (meses)	-16,730	3,605	-,102	-4,641	,000
	Nivel educativo	735,956	168,689	,124	4,363	,000

a. Variable dependiente: Salario actual

Regresión lineal múltiple con SPSS (5)

- Pruebas de significación:
  - Sirven para contrastar la hip nula de que un coeficiente de regresión vale cero en la población.
  - Si es no sig (>0,05) nos indica ausencia de relación lineal.
  - Los cof significativos son relevantes en el modelo. Los otros deberíamos quitarlos del modelo y recalcular los coeficientes (que no son independientes)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-3661,517	1935,490		-1,892	,059
	Salario inicial	1,749	,060	,806	29,198	,000
	Experiencia previa (meses)	-16,730	3,605	-,102	-4,641	,000
	Nivel educativo	735,956	168,689	,124	4,363	,000

a. Variable dependiente: Salario actual

Ejercicio: repetir el ajuste sin termino constante. Observar R^2 y SSE del nuevo modelo y compararlo

Supuestos del modelo de regresión lineal

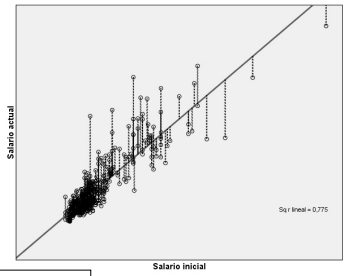
- **Linealidad** (error de especificación).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

  - Ejemplos de error de especificación:
    - Omisión de VIs importantes.
    - Inclusión de VIs irrelevantes.
    - Relación no lineal entre VD e VIs
    - Parámetros no constantes.
    - No aditividad (sensibilidad a los niveles de alguna otra VI).

Supuestos del modelo de regresión lineal

- **Independencia**.
  - Los residuos han de ser independientes entre si.
  - Es decir, son una V.A.
  - Error: residuos autocorrelados.



Los residuos son las diferencias entre los valores observados y los pronosticados.

Supuestos del modelo de regresión lineal

- **Homocedasticidad**.
  - *Para cada valor de la variable independiente la varianza de los residuos es constante.*
- **Normalidad**
  - *Para cada valor de la variable independiente los residuos se distribuyen normalmente con media cero.*

Supuestos del modelo de regresión lineal

- **No-Colinealidad**
  - *No existe relación lineal exacta entre ninguna de las variable independientes.*
  - Cuando se incumple decimos que hay colinealidad o multicolinealidad.

Supuestos del modelo de regresión lineal

- Linealidad
  - Diagrama de dispersión.
  - Gráficos parciales
- La colinealidad no se da en reg simple. (solo hay una VI)
- Independencia
- Homocedasticidad
- Normalidad

Comportamiento de los **residuos**

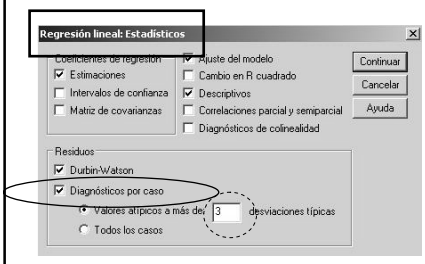
Análisis de los Residuos

Análisis de los residuos

- Los residuos nos informan sobre el grado de exactitud de los pronósticos.
- Cuando menor es el error típico de la estimación (Se o SSE),mejor son los pronósticos, o dicho de otra forma '*mejor se ajusta la recta a la nube de puntos*'.
- Residuos grandes nos pueden ayudar a localizar valores atípicos. ➔ perfeccionar la ecuación
- Los residuos nos proporciona información crucial sobre varios supuestos del modelo de regresión lineal: independencia, homocedasticidad, normalidad y linealidad.

Diagnostico por casos

- Diagnósticos por casos: Listado de los más grandes en valor absoluto.



• Pedimos valores atípicos a más de [3] Desviaciones típicas
•Los residuos tipificados tienen media 0 y Desv típica 1.
•Debemos estudiar los casos con residuos grandes.
•La media de los residuos vale 0.

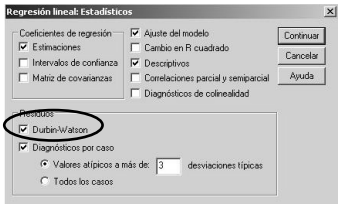
Diagnósticos por caso[a]

Número de caso	Residuo tip.	salario Salario actual	Valor pronosticado	Residuo bruto
18	6,381	$103,750	$55,048.80	$48,701.198
32	3,095	$110,625	$87,004.54	$23,620.458
103	3,485	$97,000	$70,405.22	$26,594.783
106	3,897	$91,250	$61,505.37	$29,744.628
205	-3,781	$66,750	$95,602.99	-$28,852.993
218	5,981	$80,000	$34,350.68	$45,649.323
274	4,953	$83,750	$45,946.77	$37,803.233
449	3,167	$70,000	$45,829.66	$24,170.345
454	3,401	$90,625	$64,666.70	$25,958.303

a. Variable dependiente: salario Salario actual

Independencia (Durbin-Waston)

- El estadístico de Durbin-Waston (1951) proporciona información sobre el grado de independencia de los residuos
  - DW oscila entre 0 y 4
  - Si DW=2 los residuos son independientes
  - DW>2 autocorrelación positiva.
  - Podemos asumir independencia cuando **1.5<DW<2.5**



Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,895[a]	,802	,800	$7,631.917	1,832

a. Variables predictoras: (Constante), educ Nivel educativo, expprev Experiencia previa (meses), salini Salario inicial

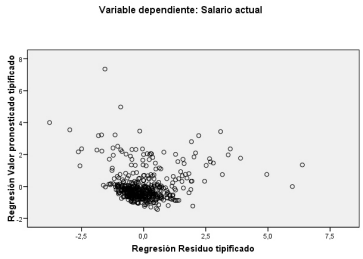
b. Variable dependiente: salario Salario actual

Homocedasticidad y Normalidad

- SPSS no proporciona algunos gráficos útiles a este propósito.

Homocedasticidad

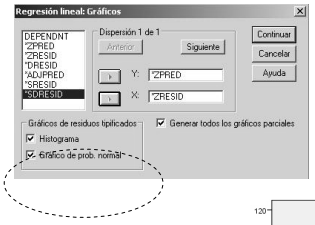
- El grafico ZPRED*ZRESID nos informa sobre la homocedasticidad ➔ No debe mostrar ninguna pauta
  - ZPRED: pronostico tipificado
  - ZRESID: residuos tipificados

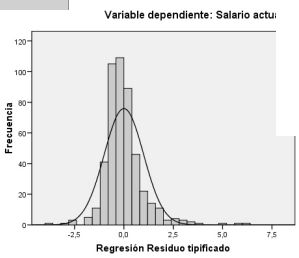


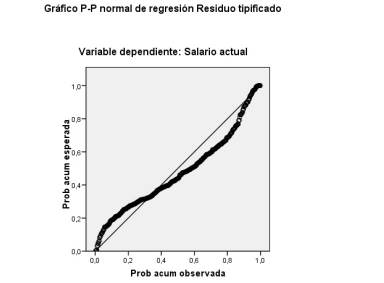
Cuando una diagrama de dispersión delata presencia de varianza heterogéneas puede utilizarse una transformación de la VD para resolver o paliar el problema (p.e. logarítmica, raíz cuadrada). Ojo con los posibles problemas de interpretación del modelo posteriores.

Normalidad

- Gráficos de residuos tipificados.







La dist de los residuos No parece seguir un modelo de prob normal

Normalidad (2)

- Otra aproximación al problema de la normalidad la podríamos obtener con la opción **guardar** de spss

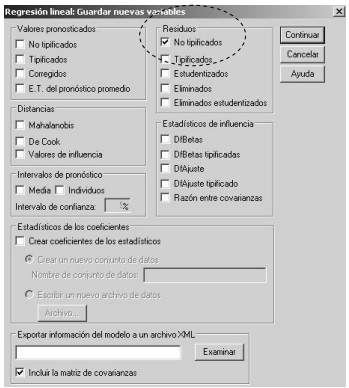
- Posteriormente le aplicaríamos un test de normalidad a la variable RES_1 (Unstandardized Residual)

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
RES_1 Unstandardized Residual	,122	474	,000	,867	474	,000

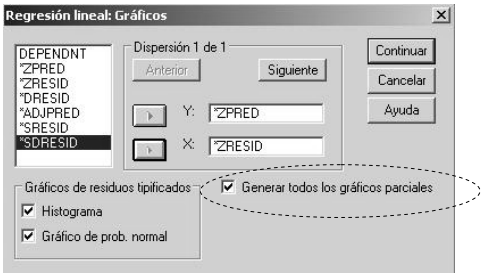
a. Corrección de la significación de Lilliefors

Confirma lo que ya sabíamos por los gráficos anteriores



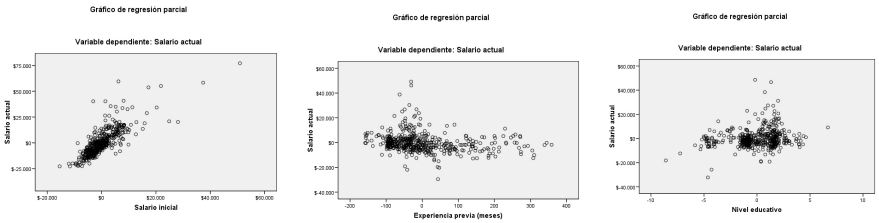
Linealidad

- Los diagramas de regresión parcial nos permiten hacernos una idea sobre la forma que adopta una relación. (están basado en los residuos y no en puntuaciones directas)
- Muestran relación neta entre las variables representadas (controlan el resto de variables).



Linealidad (2)

- Los gráficos de regresión parcial deben de mostrar relaciones lineales
- Permiten formarse una idea del tamaño y el signo de los coeficientes de regresión parcial.
- Ojo con los valores extremos, pude ser necesario investigarlos.

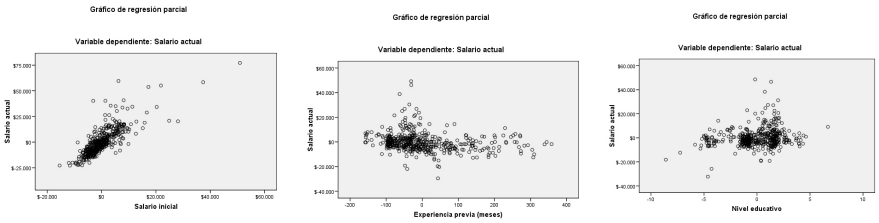


Linealidad (3)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-3661,517	1935,490		-1,892	,059
	Salario inicial	1,749	,060	,806	29,198	,000
	Experiencia previa (meses)	-16,730	3,605	-,102	-4,641	,000
	Nivel educativo	735,956	168,689	,124	4,363	,000

a. Variable dependiente: Salario actual



Colinealidad

- Excede el nivel

Validez del modelo de regresión

- Puede ser validado con nuevos casos.
  - Obtener los pronósticos para los nuevos casos
  - Calcular el coeficiente de correlación entre los observados y la nuevos casos.
  - Este coeficiente de correlación (Pearson) deberá ser igual (o muy parecido) al coeficiente de correlación múltiple de nuestro análisis de regresión (R).
  - Si no tenemos nuevos datos podemos optar por reservar datos para la validación (muestras grandes)
  - Un modelo fiable nos debería de llevar a obtener una correlación similar entre los valores observado y pronosticados en ambas mitades.
- ¿Tamaño de la muestra?
  - varias teorías: (k=nº de VIs)
    - n>=50+8k
    - n>10K o 15K

¿Por que la llamamos regresión?

- Etimología
  - El término regresión se utilizó por primera vez en el estudio de variables antropométricas: al comparar la estatura de padres e hijos, resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio.

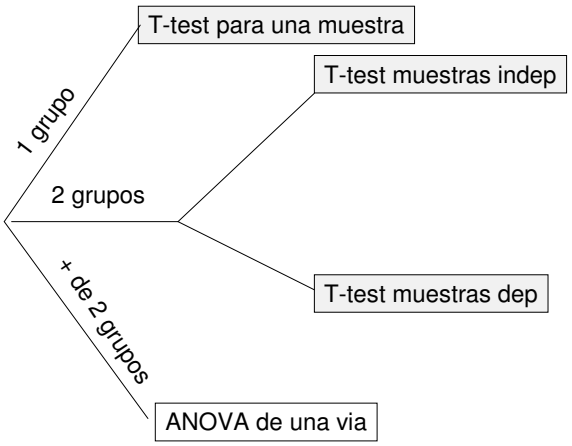
Curiosidad y consideraciones

- En medicina, las primeras evidencias relacionando la mortalidad con el fumar tabaco vinieron de estudios que utilizaban la regresión lineal.
- Los investigadores incluyen una gran cantidad de variables en su análisis de regresión en un esfuerzo por eliminar factores que pudieran producir correlaciones espurias. En el caso del tabaquismo, los investigadores incluyeron el estado socio-económico para asegurarse que los efectos de mortalidad por tabaquismo no sean un efecto de su educación o posición económica.
- No obstante, es imposible incluir todas las variables posibles en un estudio de regresión
- En el ejemplo del tabaquismo, un hipotético gen podría aumentar la mortalidad y aumentar la propensión a adquirir enfermedades relacionadas con el consumo de tabaco. Por esta razón, en la actualidad las pruebas controladas aleatorias son consideradas mucho más confiables que los análisis de regresión.

TEMA 7

t-test. Comparación de medias.

Tema 7
t-test. Comparación de medias.



t-test

- Es un procedimiento para comparar medias muestrales.
- Dadas dos muestras. ¿Hay suficiente evidencia para inferir que las medias poblacionales difieren?.

t-test para dos muestras.

- Compara las medias de dos muestras independientes o relacionadas.
- Por ejemplo:
  - La diferencia en la puntuación en un examen entre hombres y mujeres.
  - La cantidad de glóbulos rojos antes y después de administrar un tratamiento.

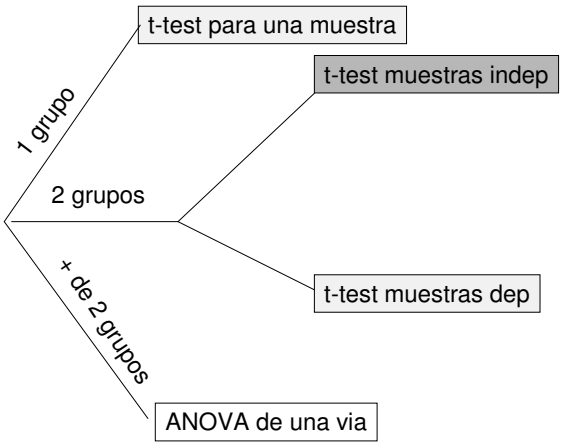
Hipótesis nula: La media de las dos poblaciones (normales) son iguales.

Exigencias del t-test

- Normalidad.
  - Test de Kolmogorov-Smirnov o Shapiro-Wilk
- Homocedasticidad. Cuando se viola esta hipótesis, existe una variante del test que se llama test de Welch o t-test heterocedastico.
  - Test de Levene, F-test.
- Independencia de las observaciones.

Depende del diseño experimental!

t-test muestras independientes.



t-test muestras independientes.

t-test muestras independientes. (2)

t-test muestras independientes. (3)

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
quiz1	Se han asumido varianzas iguales	2,180	,143	1,305	103	,195	,646	,495	-,335	1,627
	No se han asumido varianzas iguales			1,259	75,304	,212	,646	,513	-,376	1,667

P>0.05 ➔ HOV

Welch

En caso de que p<0.05 ➔ HEV
(este no es el caso ahora)

En ambos casos P>0.05, No existe dif significativas

t-test muestras independientes. (4)

- Con MS-Excel
  - 1º Estudiamos la homocedasticidad

    Prueba F para varianzas de dos muestras
  - 2º Aplicamos el test
    - HOV
    - HEV

    Prueba t para dos muestras suponiendo varianzas iguales
    Prueba t para dos muestras suponiendo varianzas desiguales

t-test muestras independientes. (5)

Prueba F para varianzas de dos muestras

	Variable 1	*Variable 2*
Media	7,71875	7,073170732
Varianza	5,316468254	7,369512195
Observaciones	64	41
Grados de libertad	63	40
F	0,721413862	
P(F<=f) una cola	0,121149706	
Valor crítico para F (una cc	**0,630961949**	

Aceptamos la hip. de igualdad de varianzas

Var(mayor)/var(menor)

1,386166877 No hay 'grandes' diferencias entr elas varianzas

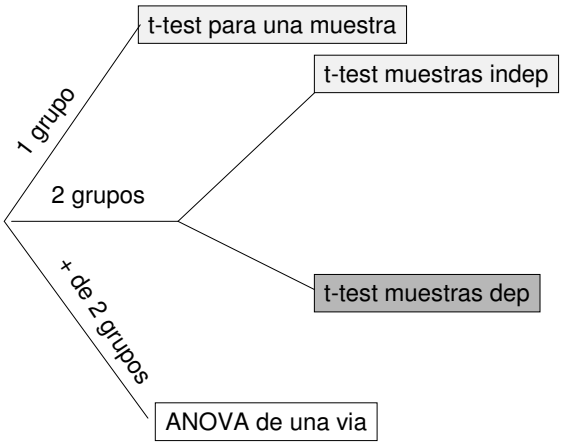
t-test muestras independientes. (6)

Prueba t para dos muestras suponiendo varianzas iguales

	Variable 1	*Variable 2*
Media	7,71875	7,073170732
Varianza	5,316468254	7,369512195
Observaciones	64	41
Varianza agrupada	6,113766872	
Diferencia hipotética de las medias	0	
Grados de libertad	103	
Estadístico t	1,305216045	
P(T<=t) una cola	**0,097363759**	
Valor crítico de t (una cola)	1,659782356	
P(T<=t) dos colas	0,194727519	
Valor crítico de t (dos colas)	**1,983262337**	

→ 0.09>0.05 ,,
No Dif Sig

Tema 6
t-test. Comparación de medias.



Tipos de variación

- Variación sistemática:
  - Producida por el experimento.
  - Por ejemplo: Aplico un tratamiento a un grupo y nada a otro grupo (control).
- Variación no-sistemática:
  - Debida a factores aleatorios que puede que existan entre las distintas condiciones experimentales.
  - Por ejemplo: personas que sin saber por que tienen diferentes tolerancias a ciertas drogas.

Ventajas del diseño dependiente

- En un diseño dependiente (*de medidas repetidas*).
  – Se controla mejor la variación no sistemática.
  – Es más fácil descubrir efectos de nuestra manipulación experimental.
  - En diseños de medidas repetidas el ruido se mantiene en el mínimo.
  - Los diseños de medidas repetidas tienen más potencia estadística, detectan diferencias que existen realmente más fácilmente que los diseños independientes.

t-test muestras dependientes.

t-test muestras dependientes.

Prueba de muestras relacionadas

		Diferencias relacionadas							
					95% Intervalo de confianza para la diferencia				
		Media	Desviación típ.	Error típ. de la media	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	quiz1 - quiz2	-,514	1,835	,179	-,869	-,159	-2,872	104	,005

Existe una diferencia significativa entre quiz1 y quiz2
Lo mismo nos dice el intervalo de confianza para la diferencia

t-test muestras dependientes.

- SPSS nos da algo más de información.

Correlaciones de muestras relacionadas

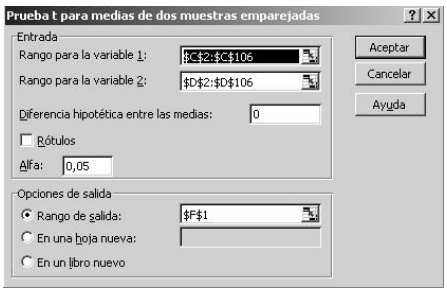
		N	Correlación	Sig.
Par 1	quiz1 y quiz2	105	,673	,000

- Existe una correlación significativa entre quiz1 y quiz2.

t-test muestras dependientes.

- Nota: en el T-test para muestras relacionadas no hablamos de homocedasticidad !!.

- Con Excel:
  - Herramientas/Análisis de datos…



Prueba t para medias de dos muestras emparejadas

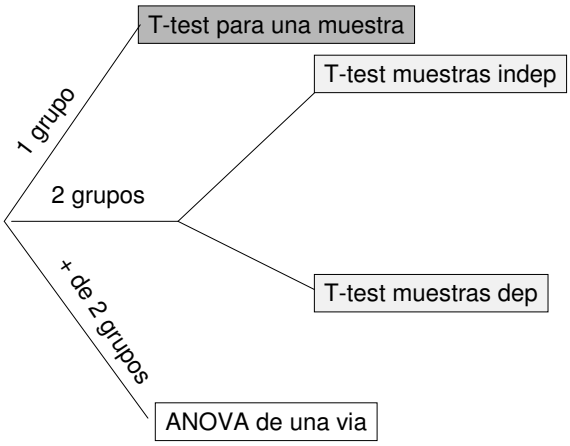
t-test muestras dependientes.

	Variable 1	*Variable 2*
Media	7,466666667	7,980952381
Varianza	6,155128205	2,634249084
Observaciones	105	105
Coeficiente de correlación de Pearson	0,673234117	
Diferencia hipotética de las medias	0	
Grados de libertad	104	
Estadístico t	-2,871706119	
P(T<=t) una cola	0,002474156	
Valor crítico de t (una cola)	1,659636837	
P(T<=t) dos colas	0,004948312	
Valor crítico de t (dos colas)	1,983034963	

P<0.05, así que rechazamos la hipótesis de igualdad de medias.

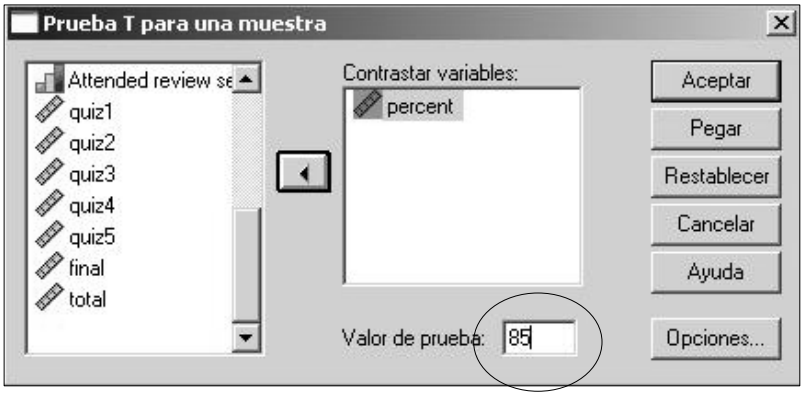
➔ Existe una diferencia significativa

t-test para una muestra.



t-test para una muestra.(2)

- Con SPSS



t-test para una muestra.(3)

Estadísticos para una muestra

	N	Media	Desviación típ.	Error típ. de la media
percent	105	80,34	12,135	1,184

Prueba para una muestra

	Valor de prueba = 85					
					95% Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
percent	-3,932	104	,000	-4,657	-7,01	-2,31

Son significativamente diferentes

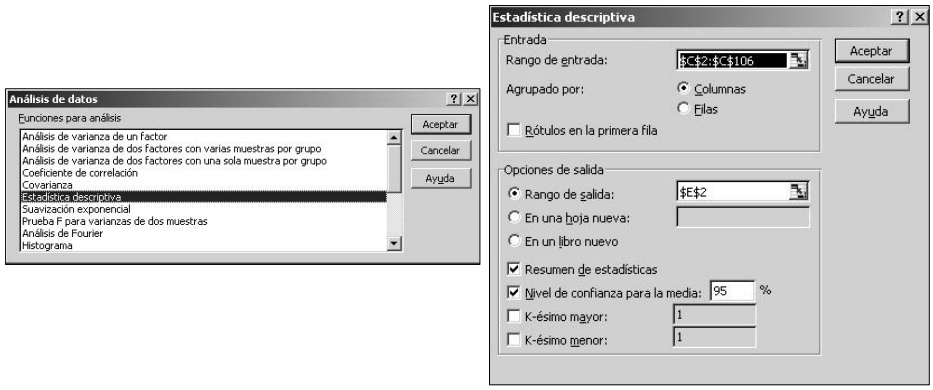
t-test para una muestra.(4)

- Con Excel basta con usar la opción de 'Estadística descriptiva' y calcular un intervalo de confianza para X-μ.
- Primero calculamos una nueva variable (columna) : percent-85

C2 =B2-85

A	B	C
id	percent	X-85
6	64	-21
10	77	-8
10	78	-7
7	82	-3
7	86	1
10	98	13
10	90	5
10	96	11
10	98	13
10	99	14
7	78	-7
8	94	9
8	89	4
3	67	-18
5	63	-22
5	75	-10
5	74	-11

t-test para una muestra.(5)



t-test para una muestra.(6)

Columna1	
Media	-4,657142857
Error típico	1,184285736
Mediana	-3
Moda	-7
Desviación estándar	12,13531763
Varianza de la muestra	147,2659341
Curtosis	0,952085001
Coeficiente de asimetría	-0,834024463
Rango	58
Mínimo	-44
Máximo	14
Suma	-489
Cuenta	105
Nivel de confianza(95,0%)	2,348480021
-7,005622878	-2,308662836

(Media- Nivel de conf, Media+ Nivel de conf)

(-7.0, -2.3)
No Contiene al '0'
➔ La dif es significativa

Tamaño del efecto

- Incluso cuando un *valor t* resulte significativo, no significa que el efecto sea importante en términos prácticos.

Convertirnos un *t valor* en un tamaño de efecto, ***r***, así:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

- *r* =0.10, efectos pequeños
- *r* =0.30, efectos medianos
- *r* =0.50, efectos grandes

Análisis estadístico. Paso a paso

- Tabular datos en el paquete estadístico.
- Explorar los datos. (hipótesis iniciales)
  - Estadísticos descriptivos
  - Conclusiones sobre las distribuciones de muestras.
    - Normalidad
    - Homocedasticidad
    - Balanceo
- Establecer hipótesis.
- Realizar tests (según hipótesis iniciales) . (t-test,..)
- Interpretar resultados.

Reglillas particulares

- K-S (n>50), S-W (n<50)
  - p>0.05 → no sig, 'Existe Normalidad'.
  - p<0.05 → Sig, 'No Existe Normalidad'.
- Levene
  - p>0.05 → no Sig, Existe HOV.
  - p<0.05 → Sig, No Existe HOV, HEV.
- t-test y otros (ANOVA,..)
  - p<0.05 → Sig, Existen Diferencias sig
  - p>0.05 → no Sig, No Existen Diferencias sig

recordar

- H_o="*No existen diferencias significativas*" ~ Hipótesis de igualdad.
- Contraste:
  - Si p<0.05, → Resultado significativo, Rechazamos la Hipótesis Nula , existen diferencias significativas.
  - Si p>0.05, → No rechazamos la Hipótesis Nula, se acepta la igualdad. (las diferencias observadas se deben al azar)

Como reportamos un *t*-test

- Medias y desviaciones típicas de cada grupo.
- Valor de *t* y grados de libertad.
- Significativo ($p<0.05$) o no significativo ($p>0.05$).
- Si es significativo, tamaño del efecto, *r*.

Como reportamos un *t*-test (2)

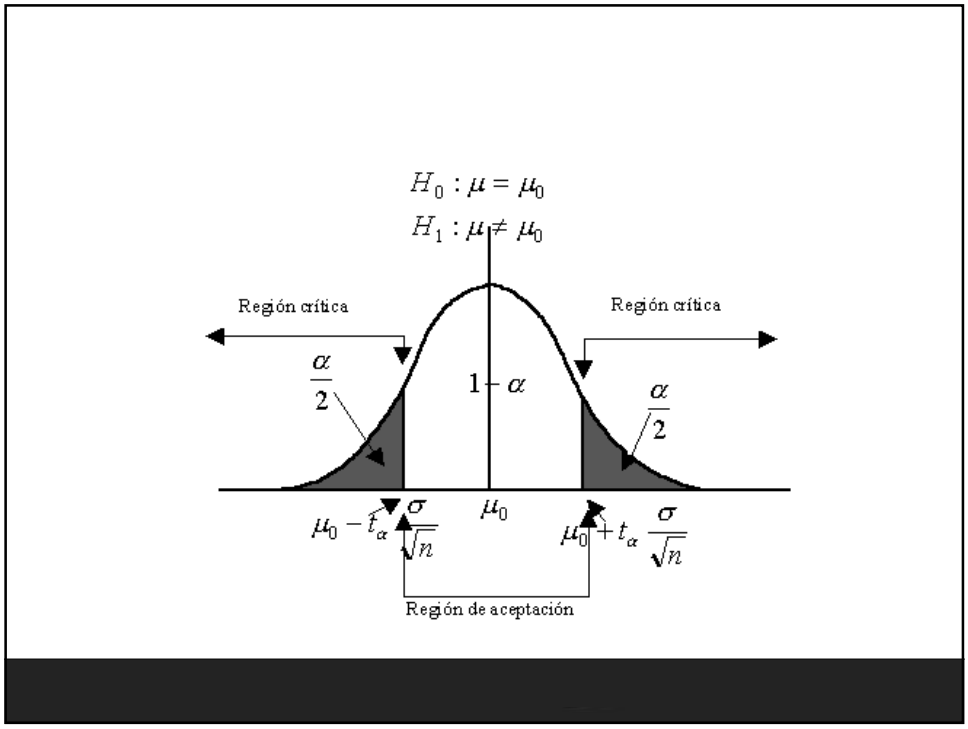
Estadísticos de grupo

	trat	N	Media	Desviación típ.	Error típ. de la media
var_ef	A	22	10,42500	3,662014	,780744
	M	26	12,88346	4,312720	,845794

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
var_ef	Se han asumido varianzas iguales	1,141	,291	-2,107	46	,041	-2,458462	1,167050	4,807611	-,109312
	No se han asumid varianzas iguales			-2,136	45,998	,038	-2,458462	1,151055	4,775419	-,141504

➢Se han encontrado diferencias significativas entre el tratamiento A (M_A=10.42, DT_A=3.66) y el M (M_M=12.88, DT_M=4.31, $t(46)$=-2.107, $p<0.05$ y r =0.29)


$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$
Región crítica
Región crítica
$\frac{\alpha}{2}$
$1-\alpha$
$\frac{\alpha}{2}$
$\mu_0 - t_\alpha \frac{\sigma}{\sqrt{n}}$
μ_0
$\mu_0 + t_\alpha \frac{\sigma}{\sqrt{n}}$
Región de aceptación

Actividades Tema 01.

Entorno de trabajo SPSS

Ejercicio 1. Crear un fichero de datos llamado *misdatos.sav.*

1. Crear las siguientes variables con los siguientes atributos:

- **genero**: numérico, anchura 1, decimales 0, etiqueta "*genero*", valores : 1 'masculino', 2 'femenino' y medida: '*nominal*'.
- **edad**: numérico, anchura 3, decimales 0, etiqueta '*Edad*', valores '999= 'NS/NC', en perdidos poner el '999', medida: '*ordinal*'.
- **coche**: numérico, anchura 1, decimales 0, etiqueta "*¿tiene coche propio?*", valores : 1 'si', 0 'no', y medida: '*nominal*'.
- **altura**: numérico, anchura 6, decimales 3, etiqueta '*altura en cm*', valores '999= 'NS/NC', en perdidos poner el '999' y en medida: '*escala*'.
- **peso**: numérico, anchura 6, decimales 3, etiqueta '*peso en kg*', valores '999= 'NS/NC', medida: '*escala*'.
- **color**: cadena, anchura 20, etiqueta '*Color favorito*', valores '999= 'NS/NC', medida: '*nominal*'.
- **nhijos**: numérico, anchura 2, decimales 0, etiqueta '*Número de hijos*', valores '99= 'NS/NC', medida: '*ordinal*'.
- **ojos**: numérico, anchura 1, decimales 0, etiqueta "*¿color de ojos?*", valores : 1 'castaño', 2 'negro', 3 'miel o avellana', 4 'verdes', 5 'azules', 6 'grises', 7 'de varios colores', 8 'otros'.

  Introduce tus datos personales para esas variables.

2. Introduzca los siguientes datos (variables: g*enero; edad; coche; altura; peso; color; nhijos; ojos*)

masculino	*25*	*no*	*1,75*	*73,00*	*verde*	*0*	*castaño*
masculino	*34*	*si*	*1,76*	*79,00*	*azul*	*2*	*verdes*
masculino	*89*	*si*	*1,65*	*89,37*	*violeta*	*4*	*verdes*
femenino	*55*	*no*	*1,50*	*57,69*	*naranja*	*2*	*negro*
femenino	*90*	*si*	*1,52*	*89,20*	*azul cielo*	*2*	*negro*

3. Importar los datos del fichero xls "*tema01-datos-para-importar-ejercicio-1.xls*".

4. Crear una nueva variable con la utilidad **'calcular'** que sea el índice de masa corporal. La variable ha de llamarse 'imc', numérica con dos decimales.

Tener en cuenta que: IMC=peso/altura^2

5. Crear una variable nueva llamada '***grupopeso***" con la utilidad "**recodificar en distintas variables**" de modo que si imc=<18.5 se le asocie el valor 1, si 18.5<imc<25 se le asocie el valor 2, y si imc>=25 se le asocie el valor 3.

Crear etiquetas para esta variable 'grupopeso' de modo que 1= infrapeso, 2= normal, 3=sobrepeso.

6. Cree una variable llamada '***caseid***', numérica , anchura 3, decimales 0, que sea un contador de casos (pistas, calcular, casenum...).

Ejercicio 2. Con el fichero '**Datos de empleados.sav**'

1. Calcular una nueva variarle, 'sfineuro' que sea el salario final en euros. Considerando que $1=0,64€.
2. Solo para los 'Administrativo' calcular la media de su salario (variable 'catlab').
3. Recodificar la variable sexo en una variable numérica donde 0=m y 1=h.
4. Calcular la media de sueldo para los 'Administrativos' y personal de seguridad que además sea hombre.

Ejercicio 3. Con el fichero '**Smonking.sav**'

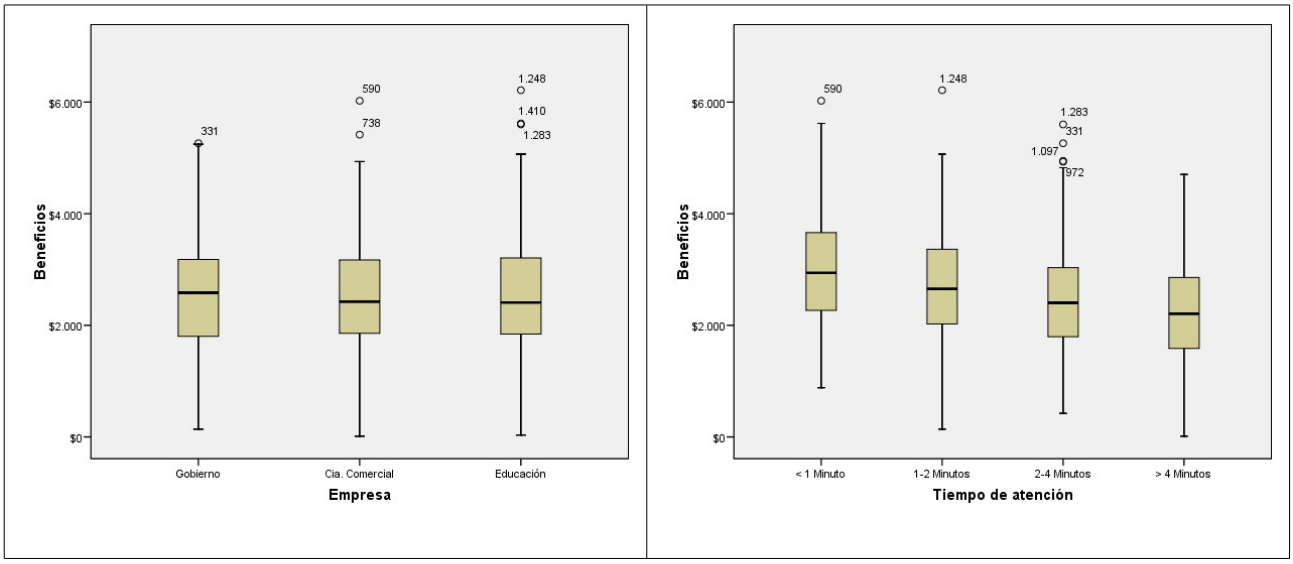
1. Construir un tabla de contingencia con las variables personal y tabaco, ponderando por la variable ***frec***.
2. Construir otra tabla de contingencia sin considerar los casos de la variable '***tabaco***' que tienen que ver con el consumo de 'Alcohol'.
3. La misma tabla de contingencia pero uniendo las categorías de empleados: directivos juntos, empleados y secretarios juntos.

Actividades Tema 2 y 3.

Estadística Descriptiva con SPSS y MS-EXCEL

Ejercicio 1. Con el fichero '**sales.sav**'

1. Para la variable '*ingresos*' calcular los siguientes estadísticos descriptivos: media, mediana, SD, Rango, Asimetría y curtosis.
2. Calcular los mismos estadísticos separadamente para los dos tipos de cliente.
    1. ¿Que tipo de cliente tiene una dispersión de los datos menor?
    2. ¿Que tipo de cliente tiene una distribución menos asimétrica?
3. Ídem según la variable 'atención'.
4. Crear gráficos de cajas y bigotes parara la variable 'salario' según la variable 'cliente' y según la variable 'atención'.



Ejercicio 2. Con el fichero **tema02-ej01.xls**

(que te puedes descargar de la web **http://www.um.es/ae-spss/curso**)

1. Par al variable 'total' calcular los estadísticos descriptivos: Media, Mediana, Moda, Desviación estándar, Varianza, Curtosis, Coeficiente de asimetría, Rango, Mínimo, Máximo.

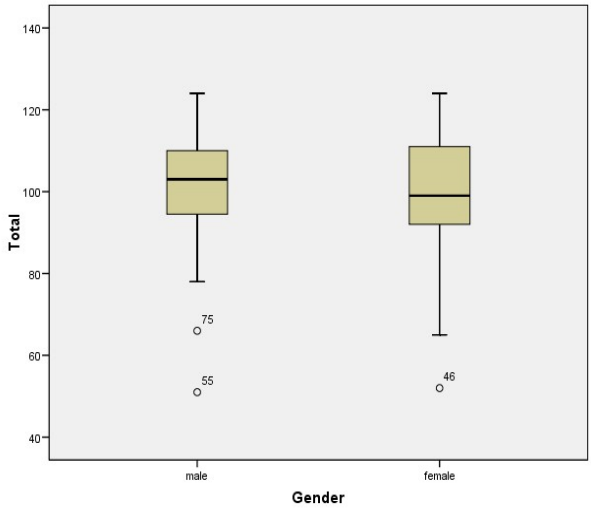
Ejercicio 3. Con el fichero **tema02-ej01.sav**

1. Crear diagramas de cajas para la variable 'total' según la variable 'gender'.
    1. ¿Que diferencias observas entre las dos distribuciones?
    2. Calcula estadísticos descriptivos y respalda tus observaciones.

Informe

Total

Gender	Media	Desv. típ.	Mediana	Rango	Asimetría	Curtosis	N
male	102,03	13,896	103,00	73	-,925	1,958	64
female	98,29	17,196	99,00	72	-,653	,078	41
Total	100,57	15,299	103,00	73	-,837	,943	105



¿Que ocurre si eliminamos los atípicos?

Ejercicio 4. Con el fichero '**tema02-ej04-pilarm.sav**',

(que te puedes descargar de la web **http://www.um.es/ae-spss/curso**)

1. Hacer un estudio descriptivo completo de la variable 'ratio'.
2. Sesgar la muestra a solo los casos que verifican la condición de que LP>5 y hacer el mismo estudio. ¿Que diferencias encuentras?
3. Crear los gráficos de cajas según 'gender'. ¿aparentemente hay mucha diferencia entre los dos grupos?.
4. Localizar y sustituir '*missings*' por la mediana y luego por la media. ¿Qué ocurre con esos estimadores en cada caso?

Actividades Tema 04. *Normalidad y Homocedasticidad.*

I) Con el fichero "**tema04-ej04-Abejas.sav**"

1. Comprobar si se cumple el supuesto de Normalidad según 'tratamiento' para las variables 'var_ef' y 'var_e'.
2. Comprobar si se cumple el supuesto de Homogeneidad de varianzas según 'tratamiento' para las variables 'var_ef' y 'var_e'.
3. * Sugiere algún cambio de variable que haga que los datos sean homocedasticos según 'tratamiento'.
4. Representa gráficamente usando un box plot las dos variables según tratamiento y también la variable transformada resultante del apartado 3 (para lo que debes de crearla primero usando el menú Transformar/calcular variable aplicar un logaritmo en base 10 para transformar).

II) Con el fichero "**tema04_ej05_gradesGPA_.sav**"

1. Estudiar si tenemos normalidad para la variable 'gpa'.
2. Si no tenemos normalidad para la variable estudiar si si la tenemos según alguno de los factores 'genero' , 'ethnicity'.
3. Según el factor 'ethnicity' estudiar si tenemos homogeneidad de varianzas.

III) Con el fichero '**tema04-ej06-edades.xls**'.

1. Comprobar si ambos sexos tienen varianzas estadísticamente homogéneas.

Actividades Tema 05.

Correlación

Ejercicio 1.

Con el fichero correlacion-ejercicio01.sav.

Los datos respuestas de 15 participantes en una encuesta.

- Calcular el coeficiente de correlación de Pearson.
- ¿es el coeficiente significativo?
- ¿es una correlación positiva o negativa?
- ¿Cual es el coeficiente de determinación?. ¿Que te dice el coeficiente?
- ¿Que forma tiene el diagrama de dispersión?
- ¿Observas valores atípicos?

Ejercicio 2.

Con el fichero Correlacion-parcial-ejercicio02.sav

- Verificar que las variables "v96, Life Satisfaction", "v116, Job Satisfaction" y "v363, Income scale" son de intervalo u ordinales. (Análisis/estadística descriptiva/frecuencias)
- Calcula la correlación de orden cero entre las variables "v96, Life Satisfaction" y "v116, Job Satisfaction".
- Calcula la correlación de orden cero entre las variables "v96, Life Satisfaction" y "v116, Job Satisfaction" controlando las posibles relaciones con la variable "v363, Income scale".

Ejercicio 3.

Con el fichero Correlacion-encuesta-ejercicio03.sav

1. Explorar gráficamente la relación entre Peso en Kg. y altura en cm. Interpreta la gráfica obtenida.
2. Indicar gráficamente y mediante un índice la relación entre Metros cuadrados del aula y Nº de escalones.
3. Controla el efecto de la variable Peso en Kg. en la relación entre Edad (años) y Número de Escalones.
4. ¿Podemos pensar que la relación entre Peso en Kg. y Altura en cm. se ve afectada por la variable Edad (años)? Razona tu respuesta.

Actividades Tema 06.

Regresión Lineal

Ejercicio 1.

Con el fichero Regresion-Record1-ejercicio01.sav

1. Ajusta el modelo de regresión simple tomando como variable dependiente 'sales' y como independiente 'adverts'.
2. ¿Crees que es un buen modelo para explicar las ventas?
3. ¿Cuanta varianza de 'sales' no explica el modelo?
4. ¿es este un modelo mejor que considerar la media de 'sales' como modelo predictivo?
5. Si el modelo es útil usalo para predecir cuales serán las ventas si el presupuesto es de 100.000€.
6. Crea un diagrama de dispersión de sales*adverts e inserta la recta de regresión.

Ejercicio 2.

Un estudiante de modas esta interesado en los factores que pueden predecir los salarios de las modelos de una famosa agencia de modelos. Así que a cada modelo le pregunto: Salario que obtenían por día trabajado (**salary**), edad (**age**), cuantos años llevaba trabajando como modelo(**years**) y pidió a un panel de expertos que puntuaran su atractivo de 0 a 100% siendo 100% 'perfectamente atractiva/o'. Los datos están en el fichero regresion-Supermodelos-ejercicio02.sav.

1. Interpreta el modelo. ¿cuanta varianza de 'salary' explica?
2. ¿Que puedes decir sobre los residuos del modelo?
3. ¿Quitarías alguna variable predictora del modelo?
4. ¿Que ocurre con 'years' y 'age'? ¿hay alguna relación entre ellas?, ¿afecta esta relación a la calidad del modelo?
5. ¿Es un buen modelo?

Ejercicio 3.

Deseamos estimar el efecto que las variaciones en el nivel de ocupación de las empresas tienen sobre la cuenta de resultados. Para eso se han recogido las tasas de variación de esas variables en diversos sectores y los valores son:

Resultados	**Plantillas**
2,1	6,1
2,2	11,0
7,8	1,4
15,3	7,2
17,2	4,2
20,3	4,2
25,0	-1,9
26,3	-2,4

Se pide:

1. Estimar un modelo de regresión simple.
2. Interpretar el coeficiente de la variable independiente y contrastar la significación de los parámetros del modelo.
3. Obtener una representación gráfica de los residuos. Valorar la bondad del ajuste (Durbin-Waston, normalidad de residuos).
4. ¿Es un buen modelo?.

Ejercicio 4.

Los datos de la siguiente tabla representan las estaturas (X, cm) y los pesos (Y, kg) de una muestra de 12 hombres adultos.

X	152	155	152	155	157	152	157	165	162	178	183	178
Y	50	61.5	54.5	57.5	63.5	59	61	72	66	72	84	82

Ajuste un modelo de regresión y analiza los residuos.

Actividades Tema 07. *T-Student*

I) Con el fichero "**tema07-ej04_ansiedad-araña01.sav"**

Una experiencia de investigación ha consistido en medir la ansiedad en dos grupos de individuos. Al primer grupo se le mostró una foro de una araña peluda y al segundo grupo se les mostró la araña peluda real. Con el ansionómetro se midieron las ansiedades individuales después de tan interesante experiencia.

1. Se quiere saber si podemos concluir que la araña real produce más ansiedad que la foto.

II) Con el fichero "**tema07-ej05_ansiedad-araña02.sav**".

Se ha realizado una experimento parecido al del apartado I. Solo que esta vez se les enseño primero una foto de una araña horripilante y meses después (a los mismos señores) se les mostró la araña horripilante real. En ambos casos hemos medido la ansiedad que producía la experiencia.

1. Se quiere saber que produce más ansiedad en los sujetos, si ver la foto o ver la araña real.
2. Como curiosidad comprobar que son los mismos datos que el ejercicio I, solo que consideramos el modelo de medidas repetidas.

III) Con el fichero "**tema07-ej06-libros-felicidad.sav**".

Se ha medido la felicidad que produce en 500 lectores leer un libro de Estadística (concretamente "Discovering how to design and report experiments") y la felicidad que produce en estos mismos sujetos leer un libro banal que proporciona la revista Cosmopolitan.

1. Queremos saber que libro produce más felicidad.

IV) Con el fichero "**tema07-ej07-felicidad-VitC.sav**"

1. *A un grupo de pacientes con una enfermedad X se les ha dividido en dos grupos, a un grupo se le ha administrado un tratamiento consistente en vitamina C y al otro un tratamiento placebo. Si la variable 'indice' mide el incremento de felicidad, ¿que tratamiento resulta más efectivo?.*

Notas: