ESTADÍSTICA DESCRIPTIVA (UNIDIMENSIONAL)

Ejemplo1 (Operadores/"Generate Data"/"Modify Column")

❶ Cree un fichero llamado "Ejemplo1" que contenga la siguiente distribución de frecuencias del número de hijos de un grupo de familias:

x	n
1	6
2	3
3	2
4	3
5	4
6	1

❷ Calcule la media de esta variable, creando primero una columna llamada "x*n" que contenga el producto de las dos columnas anteriores. *Recuerde que para nombrar una columna se utiliza "Modify Column", a donde se puede acceder con el botón derecho del ratón. Utilice "SUM"*
(Haga lo mismo creando una sola columna tipo fórmula: "AVG")

❸ Cree otra columna que calcule las frecuencias absolutas acumuladas, utilizando para ello el operador "*Runtot*".

❹ Obtenga las frecuencias relativas y relativas acumuladas. Conteste:

- ¿Cuántas familias tienen 1 hijo?. ¿Y 4?
- ¿Qué proporción de familias tienen 4 hijos?
- ¿Qué proporción de familias tienen menos de 4 hijos?
- ¿Qué proporción de familias tienen 4 hijos o más?

Ejemplo2 (Operadores)

❶ Genere una columna de datos utilizando el operador "*Count(start,end,step)*" que genere datos desde el 2 hasta el 40 de dos en dos.*(Recuerde que los parámetros se separan con ;).*

❷ Calcule la media, varianza y mediana de esta variable, utilizando los operadores estadísticos disponibles.

❸Tipifique los valores de la variable x, mediante el operador "*Standarized*".

Ejemplo3 (Tabla de Frecuencias/Medidas Descriptivas/Recodificación/Gráficos)

Los datos del fichero "Mate. sf3" contienen las calificaciones de un grupo de alumnos en matemáticas:

❶ Abra el fichero de datos.

❷ Realice un resumen estadístico mediante la opción "Descripción/Datos Numéricos/ A.Unidimensional".

Con "Opciones Tabulares" puede ver algunas medidas de tendencia central, dispersión o de posición, la tabla de frecuencias, percentiles, histograma, etc. Coméntelos.(con "Opciones de Ventana" aparecen más indicadores). Analice la asimetría y curtosis de la distribución.

❸ Construya y comente el histograma de frecuencias y el gráfico de caja, usando las opciones gráficas del análisis anterior. Hay datos anómalos. Cópielos en el "Statgallery".

❹ Recodifique ("Botón Derecho/Recodificar") la variable en tres categorías del siguiente modo:

3-5	1(suspenso)
5-7	2(aprobado)
7-9	3(notable)

Nota: Para recodificar una variable hay que copiarla a otra columna, para no perder los datos originales. Llámela "calif".

❺ Construya y salve la tabla de frecuencias de la variable "Calif" mediante "Descripción /Datos cualitativos/ Tabulación". Puede salvarla con el icono "Save Result". Realice un gráfico de sectores y de barras utilizando la tabla de frecuencias anterior. Cópielos al "Statgallery".

❻ Ordene la información contenida en la variable "mate", utilizando "Edit/Sort File". Cambie el valor 8 que aparece en la columna "Mate" por un 10 y realice de nuevo el gráfico de caja, comentando resultados. ¿Cuál es la nota mínima del 20% de los alumnos que mayor nota han obtenido?

❼ Calcule las medidas descriptivas (incluyendo coeficiente de variación) de "Mate" para los alumnos suspensos (En selección: Calif="Suspenso"). Haga lo mismo para los alumnos que han obtenido un aprobado (En selección: Calif="Aprobado"). ¿Qué distribución es más homogénea?. ¿En cuál de ellas es más representativa la nota media?. En Gráficos/Gráficos Exploratorios realice un gráfico de caja múltiple para los tres grupos de alumnos.

Ejemplo4 (Tabla de Frecuencias/Medidas descriptivas/Gráficos)

En dos sectores productores de bienes de consumo, la inversión en publicidad se distribuye de la siguiente forma entre sus empresas:

Sector A	Sector B
Inversión (en millones)	**Inversión (en millones)**
70	90
70	90
80	100
80	110
80	110
90	110
90	120
90	120
100	130
100	130
110	140

❶ Cree dos columnas A y B, donde aparezcan estas inversiones. Calcule la tabla de frecuencias para cada sector y grábela (Descripción/Datos cualitativos/Tabulación).

❷ Calcule la media, desviación estándar y coeficiente de variación para cada sector, utilizando el menú "Descripción" ("Análisis multidimensional").

¿En cuál de los dos sectores la inversión en publicidad de las diferencias empresas que los componen es más homogénea?.

❸ Considerando el sector A:

i) ¿Cuál es la inversión mínima en publicidad del conjunto formado por el 15% de las empresas que más invierten en publicidad?

ii) ¿Cuál es la inversión mediana?

❹ Construya un gráfico de sectores que nos muestre la distribución de las inversiones en cada sector, utilizando la tabla de frecuencias obtenida en el primer punto.

Ejemplo 5

La tabla siguiente contiene el Peso (en Kg), Altura (en cm), Edad y Sexo de 16 individuos:

P	75	81	56	68	79	89	62	59	83	55	72	56	84	61	76	68
A	173	178	162	180	182	185	157	165	180	160	174	161	182	163	172	169
E	21	22	22	21	24	22	21	22	23	22	21	23	22	24	22	21
S	H	H	M	M	H	H	M	M	H	M	H	M	H	M	H	M

(Recodifique: H=0, M=1 en la variable Sexo y cree otra variable "Labels" con las etiquetas)

❶ Calcule las medidas descriptivas asociadas a cada variable.
(Puede utilizar si quiere Describe/Multiple analysis)
¿Qué variable es más homogénea, el Peso o la Altura?

❷ Calcule la media, mediana y desviación típica de la variable Peso según el Sexo.
¿Es más homogéneo el Peso en las mujeres o en los hombres?.
Estudie la forma de la distribución del peso en las mujeres y en los hombres, calculando el coeficiente de asimetría y aplastamiento. ¿Hay muchas diferencias entre las dos distribuciones?.

❸ Recodifique en 3 clases ("Recode"):Delgado(<60 Kg), Medio(60-75 Kg), Pesado(>75 Kg).
Obtenga la distribución de frecuencias de la variable Peso recodificada.
Realice un gráfico de sectores para reflejar esta distribución.

❹ ¿Cuál es el peso mínimo del 20% de los individuos que más pesan?

❺ Obtenga el histograma para la variable "Peso" y un gráfico de barras para la variable "Edad".

❻ Realice un diagrama de dispersión con las variables Peso y Altura y un diagrama tridimensional con Peso,Altura y Edad.

Ejemplo 6 (PROPUESTO)

Para los siguientes ejercicios usaremos el fichero **Cardata**. Para ello siga el camino: **Archivo ... Abrir ... Abrir Datos**, seleccione el fichero y pulse en **Abrir** .

1. Seleccione la primera columna libre y mediante **Modificar columna** nombre a la variable por: **log-nep**.

   Mediante la opción **Generar datos** utilice en el campo de **Expresiones** la siguiente fórmula **LOG(ABS(horsepower - displace))**. Modifique la variable, utilizando sólo dos cifras decimales.

2. Ordene de menor a mayor la variable **displace**. Mire el menor y mayor valor de esta variable. Cópiela en una columna libre de nombre **displace-c**. A continuación, seleccione **Recodificar Datos** y tome los siguientes intervalos: [70,150) ; [150,230) ; [230,310) ; [310,390). Los valores de las clases son: 1, 2, 3 y 4.

3. Para la variable **accel**, deduzca:

   - Media, Desviación típica, Varianza, Mediana, Mínimo, Máximo, Asimetría y Curtosis.
   - La Tabla de Frecuencias.
   - Los cuartiles y percentiles 35, 65 y 80.
   - El Diagrama de caja. ¿Es la mediana mayor que la media? ¿Hay valores atípicos? ¿En que fila están?
   - El Histograma con 7 clases y la curva acumulativa

   y coloque todos los resultados en el StatReporter

4. Determine el diagrama de barras y de sectores de la variable **cylinders**. Sitúe los gráficos en la StatGalery.

5. Determine un diagrama de caja múltiple para la variable **weight** utilizando como Código de Nivel a la variable **origin**.

MODELOS DE PROBABILIDAD

Ejemplo7 (Representación de modelos de distribución)

Mediante el menú *"Gráficos/Distribuciones de probabilidad"*, obtenga las representaciones gráficas de las distribuciones de probabilidad siguientes:

❶ B(10,0,1)
❷ P(1)
❸ N(50,5)
❹ N(50,20)
❺ N(40,10)
❻ t-Student con 8 g.l.
❼ Chi-Cuadrado con 8 g.l.

(Recuerde que con el botón derecho del ratón se puede acceder al menú "Opciones de análisis" donde podemos cambiar los parámetros de cada distribución).

Ejemplo8 (Cálculo de probabilidades)

Mediante el menú *"Gráficos/Distribuciones de probabilidad/Opciones Tabulares"*, realice los siguientes ejercicios:

❶ La tasa de paro de un cierto país es del 25% de la población activa. Se realiza una encuesta con diversas preguntas a 10 personas pertenecientes a este grupo de población.

a) ¿Cuál es la probabilidad de que 4 personas de las entrevistadas estén en paro?
b) ¿Cuál es la probabilidad de que al menos 3 personas estén en paro?
c) ¿Cuál es la probabilidad de que menos de 2 personas estén en paro?

❷ El número de clientes por minuto que llegan a un banco es una variable aleatoria de Poisson. Si el número promedio es de 2 por minuto:

a) ¿Cuál es la probabilidad de que exactamente una persona llegue durante un minuto seleccionado al azar?.
b) ¿Cuál es la probabilidad de que lleguen menos de 4 personas?
c) ¿Cuál es la probabilidad de que lleguen 3 o más personas?.

❸ Sea X una variable aleatoria Normal de media 10 y desviación típica 2. Calcule:

a) P[X<10]
b) P[X>8,4]
c) P[8<X<12]
d) Percentil 90 ("Inverse CDF")
e) Genere 200 números aleatorios de esta distribución y calcule su percentil 90.

❹ Sea Y una variable Chi-Cuadrado con 10 grados de libertad. Calcule:

a) P[Y<12.6]
b) P[Y>9.35]
c) Obtenga el valor de k tal que P[Y<k]=0.7

Ejercicio propuesto

1. La proporción de parados de una población es de 0.2. Se seleccionaron 60 individuos de dicha población. Obtener:

   a) Probabilidad de que 20 o más estén parados.

   b) Probabilidad de que trabajen exactamente 48.

   (Solución: $X \sim B(60, 0{,}2)$; *a*) $P[X \geq 20] = P[X > 19] = 0{,}0106683$; *b*) $P[X = 12] = 0{,}127823$).

2. Un promedio de 4 personas acuden a una oficina de información de un supermercado cada hora. Obtener la probabilidad de que:

   a) Exactamente 2 personas acudan durante una hora seleccionada al azar.

   b) Menos de 5 acudan durante una hora seleccionada al azar.

   c) Más de 7 personas acudan durante una hora seleccionada al azar.

   (Solución: $X \sim P(4)$; *a*) $P[X = 2] = 0{,}146525$; *b*) $P[X < 5] = 0{,}628837$; *c*) $P[X > 7] = 0{,}0511336$).

3. Sea X una variable aleatoria con distribución $N(4, 6)$. Calcular

   a) $P[X \leq 5]$

   b) $P[X \geq 3]$

   c) $P[2 \leq X \leq 6]$

   d) Percentiles: P_{12}, P_{25}, P_{52}, P_{85} y P_{90}

   (Solución: $X \sim N(4, 6)$; *a*) $P[X \leq 5] = 0{,}566186$; *b*) $P[X \geq 3] = 0{,}566186$; *c*) $P[2 \leq X \leq 6] = 0{,}261122$; *d*) $P_{12} = -3{,}04994$, $P_{25} = -0{,}046951$, $P_{52} = 4{,}30093$, $P_{85} = 10{,}2186$ y $P_{90} = 11{,}6893$).

4. Representar gráficamente las funciones de densidad de las siguientes distribuciones:

   a) $N(0, 2)$

   b) t_4

   c) χ^2_4

   d) $F_{8,12}$

5. Generar una muestra aleatoria de tamaño 200 de una distribución $N(5, 4)$ y representarla mediante un gráfico de normalidad. Hacer un análisis descriptivo de la muestra generada.

ESTIMACIÓN.TEST DE HIPÓTESIS

Ejemplo9 (Intervalos de confianza y Test de hipótesis)

Se ha extraído una muestra de 20 paquetes cuyos pesos en gramos son los siguientes:

520	436	521	520	503	478	524	428
538	463	515	519	447	525	550	491
506	494	457	548				

Mediante el menú "*Descripción/Datos Numéricos/Análisis Unidimensional*", obtenga la media y desviación típica muestral, y realice lo siguiente*:*

❶ Estime el peso medio de la población a un nivel de confianza:
   a) del 80%
   b) del 95%

❷ El proveedor asegura que el peso medio de los paquetes del envío es superior a los 520 gramos. Contraste esta hipótesis a un nivel de significación del 10%.

Mediante el menú "*Descripción/Contraste de Hipótesis*":

❸ Contraste la hipótesis nula de que la desviación típica es igual a 22 al 5%.

❹ Si suponemos que la verdadera media poblacional es 500 gr. y la desviación típica vale 30 gr., determine el tamaño muestral necesario para que el error absoluto de la media de la muestra no difiera más de 10 gr.

❺ Estudie la normalidad de la muestra mediante la representación gráfica del histograma y gráfico de normalidad.

Ejemplo10 (Intervalos de confianza y Test de hipótesis)

Se encuestó a 300 personas seleccionadas al azar de una población, para conocer la proporción de votantes favorables a un candidato. De los 300 encuestados sólo 100 se mostraron favorables al candidato.

❶ Obtenga un intervalo de confianza para la verdadera proporción de la población a un nivel de confianza del 95%.

❷ Contraste la hipótesis de que más del 25% de la población de muestre favorable al candidato.

Ejemplo11

❶ El salario anual (en millones de pesetas) de un grupo de trabajadores sigue una distribución Normal de media 2,5 y desviación típica 0,5. Determine:

a) Probabilidad de que el salario anual de un individuo elegido al azar sea superior a 3 millones de pesetas.
b) ¿Cuál es el menor salario que cobra el 45% de los trabajadores mejor pagados?
c) Represente gráficamente la distribución de probabilidad para esta variable.
d) Genere 40 valores de esta distribución y grábelos como Norm_40. Comente los estadísticos de resumen de esta variable y realice el apartado b.

❷ Los niveles de audiencia (en miles de personas) de un programa de televisión, medidos en 10 emisiones elegidas aleatoriamente, han sido los siguientes:

682 553 555 666 657 649 522 568 700 552

(Suponga que los niveles de audiencia siguen una distribución normal)

a) Obtenga un intervalo de confianza para la audiencia media de toda la población a un nivel del 80%.

b) Obtenga un intervalo de confianza para la desviación típica poblacional a un nivel del 90%.

c) Un directivo de la cadena que emite este programa afirma que la audiencia media del programa es de 600.000 espectadores. Contraste esta hipótesis para un nivel de significación del 5%.

d) La compañía productora del programa afirma que este acapara una audiencia fiel y que su desviación típica sería menor de 15.000 espectadores. ¿Puede probar esta hipótesis con los datos disponibles y a un nivel de significación del 5%?.

e) Estudie la normalidad de esta muestra mediante el histograma y gráfico de normalidad.

Nota: *En el caso de la realización de contrastes de hipótesis para dos poblaciones seleccione "Comparación/dos muestras/Contrastes de Hipótesis"*

Ejercicios propuestos

1. El tiempo de paro, en meses, de 20 individuos en una oficina de empleo es el siguiente:

20	33	24	31	23	11	43	22	14	25
32	21	43	22	15	30	21	16	34	13

   a) Obtener estimaciones puntuales para la media y la varianza.
   b) Obtener un intervalo de confianza para la media a un nivel del 90 %.
   c) Obtener un intervalo de confianza para la varianza a un nivel del 94 %.
   d) Ralizar el siguiente contraste a un nivel de significación del 4 %.

$$\begin{cases} H_0 : \mu \leq 30 \\ H_1 : \mu > 30 \end{cases}$$

   e) Ralizar el siguiente contraste a un nivel de significación del 8 %.

$$\begin{cases} H_0 : \sigma^2 = 5 \\ H_1 : \sigma^2 \neq 5 \end{cases}$$

   (Solución: a) 24,65 ; 85,3974 ; b) [21,077 , 28,223] ; c) $[7,10324^2, 13,2799^2]$; d) P-valor=0,991, se acepta H_0 para $\alpha = 0,04$; e) P-valor=0,0, se rechaza H_0 para $\alpha = 0,08$).

2. En una ciudad se desea estudiar la proporción de individuos que tienen alguna mascota, para ello se selecciona una muestra de 500 individuos, de los cuales resulta que 300 tienen alguna mascota. Obtener un intervalo de confianza del 95 % para la proporción de individuos de la ciudad que tienen alguna mascota. Realizar el siguiente test a un nivel de significación del 10 %:

$$\begin{cases} H_0 : p \leq 0,5 \\ H_1 : p > 0,5 \end{cases}$$

   (Solución: Intervalo: [0,566334 , 0,6347] ; P-valor=0,00000477248, se rechaza H_0 para $\alpha = 0,1$).

3. Se ha extraído una muestra de 20 paquetes cuyos pesos en gramos son los siguientes:

520	436	521	520	503	478	524	428	538	463
515	519	447	525	550	491	506	494	457	548

   Se pide:

   a) Estimación puntual y por intervalos del peso medio de la población al nivel de confianza del 80 %.
   b) El proveedor asegura que el peso medio de los paquetes del envío es superior o igual a los 520 gramos. Contrastar esta hipótesis a un nivel de significación del 10 %.
   c) Contrastar la hipótesis nula de que la desviación típica es igual a 22 al nivel de significación del 5 %.
   d) Si suponemos que la verdadera media poblacional es 500 gramos y la desviación típica vale 30 gramos. Determine el tamaño muestral necesario para que el error absoluto de la media de la muestra no difiera en más de 10 gramos.

(Solución: a) 499,15 ; Intervalo: [488,335 , 509,965] ; b) P-valor=0,00958238, se rechaza H_0 para $\alpha = 0,1$; c) P-valor=0,000128011, se rechaza H_0 para $\alpha = 0,05$; d) 38 observaciones).

4. De una población se extrae una muestra de 300 personas de las que fuman 83.

   Se desea conocer la proporción de fumadores en la población. Obtener:

   a) El intervalo de confianza para la proporción de fumadores poblacional a un nivel de confianza del 90 %.

   b) El contraste de hipótesis de que menos de la cuarta parte de la población es fumadora al 5 % de significación.

   c) Si la verdadera proporción de fumadores fuera del 28 %, ¿qué tamaño muestral mínimo es necesario para que la proporción muestral no difiera en mas/menos el 2 % del valor poblacional a un nivel de significación del 5 %.

   (Solución: a) Intervalo: [0,234305 , 0,322355] ; b) P-valor=0,871463, no se rechaza H_0 para $\alpha = 0,05$; c) 2047 observaciones).

5. Contraste de hipótesis para la media de una población Normal.

   Datos:

   $H_0 : \mu \geq 85$, $H_1 : \mu < 85$ (contraste unilateral) ; Tamaño muestral=10 ; Media muestral=82,3 ; Desviación típica muestral=7,57261 ; Nivel de significación=5 %.

   (Solución: P-valor=0,144345>0,05. No se rechaza la hipótesis nula para $\alpha = 0,05$).

6. Contraste de hipótesis para la proporción de una población Binomial.

   Datos:

   $H_0 : p = 0,7$, $H_1 : p \neq 0,7$ (contraste bilateral) ; Tamaño muestral=500 ; Proporción muestral=0,68 ; Nivel de significación=1 %.

   (Solución: P-valor=0,35387>0,01. No se rechaza la hipótesis nula para $\alpha = 0,01$).

7. Contraste de hipótesis para la desviación típica de una población Normal.

   Datos:

   $H_0 : \sigma \leq 9,2$, $H_1 : \sigma > 9,2$ (contraste unilateral) ; Tamaño muestral=10 ; Media muestral=183,5 ; Desviación típica muestral=12,0485 ; Nivel de significación=8 %.

   (Solución: P-valor=0,0796369<0,08. Se rechaza la hipótesis nula para $\alpha = 0,08$).

8. Contraste de hipótesis para la diferencia de medias de poblaciones Normales. (muestras independientes)

   Datos:

   $H_0 : \mu_1 - \mu_2 \leq 0$, $H_1 : \mu_1 - \mu_2 > 0$ (contraste unilateral) ; Tamaños muestrales=9 y 9 ; Medias muestrales=329 y 283 ; Desviaciones típicas muestrales=45 y 43 ; Nivel de significación=5 %.

   (Solución: Primero se efectúa el contraste $H_0 : \sigma_1/\sigma_2 = 1$, $H_1 : \sigma_1/\sigma_2 \neq 1$. Del que se obtiene un P-valor=0,900824>0,05, por lo que se asume que las varianzas son iguales. A continuación, se efectúa el contraste de diferencia de medias pedido y se obtiene un P-valor=0,0207207<0,05, por lo que se rechaza la hipótesis nula (la medias son distintas) para $\alpha = 0,05$).

9. Contraste de hipótesis para la diferencia de medias de poblaciones Normales. (muestras apareadas)

Datos:

Var1	98,2	107,9	95,3	102,1	99,9	98,7	100,5
Var2	98,1	103,0	96,5	98,3	89,8	90,6	94,1

Se genera la variable D = Var1 − Var2. Que son 7 valores comprendidos entre -1,2 y 10,1.

$H_0: D = 0$; $H_1: D \neq 0$. Media muestral para D=4,6.

(Solución: P-valor=0,0247712<0,05. Se rechaza la hipótesis nula para $\alpha = 0,05$).

10. Contraste de hipótesis para la diferencia de proporciones de poblaciones Binomiales.

Datos:

$H_0: p_1 - p_2 \leq 0,04$, $H_1: p_1 - p_2 > 0,04$ (contraste unilateral) ; Tamaños muestrales=75 y 75 ; Proporciones muestrales=0,63 y 0,57 ; Nivel de significación=5 %.

(Solución: P-valor=0,40111>0,05. No se rechaza la hipótesis nula para $\alpha = 0,05$).

REGRESIÓN SIMPLE/MÚLTIPLE

Ejemplo 12

Galton estudió en 1977 la relación entre el diámetro de los guisantes y el diámetro medio de sus descendientes con los resultados siguientes:

Diámetro padres	21	20	19	18	17	16	15
Diámetro medio descendientes	17.26	17.07	16.37	16.40	16.13	16.17	15.98

Los datos están en pulgadas x100 (1 pulgada=2,54 cm.).

❶ Obtenga la representación gráfica de la nube de puntos, utilizando el menú "Plot/ Scatterplots". ¿Cree que hay alguna relación entre las dos variables?.

❷ Calcule las varianzas marginales y covarianza mediante la opción "Describe/Numeric Data/ Multiple-Variable Análisis". Seleccione también la matriz de correlaciones y coméntela.

❸ Calcule la expresión matemática de la recta de regresión simple ("Relate/Simple Regression"). ¿Cuánto vale el coeficiente de correlación de Pearson?.¿Y R^2?. ¿Qué conclusiones pueden extraerse a la vista de estos valores?. Intreprete los coeficientes de regresión estimados.
Recuerde: Si $Y=\alpha+\beta X$, entonces α="intercept" y β="slope".

❹ Realice el gráfico del modelo ajustado (mediante las opciones gráficas del menú anterior) y salve los valores predichos (con el botón "salvar resultados").

❺ Prevea el diámetro medio de los descendientes de guisantes con diámetro 25 pulgadas (con la opción "Tabular Opción/Forecast").

❻ Para comprobar si el modelo es adecuado es necesario analizar los residuos (comprobar su aleatoriedad y normalidad, con media 0). Realice el gráfico de los residuos y después grábelos mediante "Save Result". ¿Parecen ser aleatorios?. Obtenga también el gráfico probabilístico normal para comprobar su normalidad (One-Variable Analysis).

❼ Genere de forma manual la columna de valores predichos (Generate Data) y llámela "mis_predichos". Obtenga también otra columna llamada "mis_residuos" donde genere los residuos de este modelo, utilizando la columna de la variable dependiente y la de valores predichos. Compruebe que los residuos que ha obtenido son iguales a los que proporciona el programa.

❽ ¿Cuánto valen la suma de cuadrados explicada, residual y total? Realice los contrastes de hipótesis para comprobar la significación de los parámetros del modelo.

Ejemplo 13

Los datos siguientes muestran el número de contratos realizados y el número de accidentes en jornada de trabajo con baja, en las 8 provincias andaluzas durante el año 1996.

PROVINCIA	CONTRATOS	ACCIDENTES
Almería	120596	71
Cádiz	292702	136
Córdoba	312886	87
Granada	138434	72
Huelva	188244	69
Jaén	247632	75
Málaga	267035	136
Sevilla	505679	215

❶ Obtenga la representación gráfica mediante el diagrama de dispersión. ¿Existe relación lineal entre las dos variables?.

❷ Realice un ajuste de regresión lineal simple, que nos permita estimar el número de accidentes de trabajo en función del número de contratos realizados. Estime los parámetros del modelo e interprete sus valores.

❸ Contraste la significación de los parámetros estimados y del modelo en su conjunto.

❹ Interprete la tabla del análisis de la varianza.

❺ ¿Cuál es el coeficiente de determinación?. ¿Qué indica?.

❻ Obtenga el gráfico del modelo ajustado, el de valores observados/predichos y el de los residuos. ¿Existen comportamientos anómalos? ¿Existen valores anómalos?, ¿Existen puntos influyentes?, ¿Se observa aleatoriedad en los residuos? .

❼ Grabe en una variable los valores predichos y los residuos. Prediga el número de accidentes que ocurrirán en Granada en el año 2004 si se prevé que el número de contratos ascenderá a 200.000 en ese año.

Ejemplo 14

Una empresa de colonias ha decidido lanzar al mercado una nueva colonia. Se sospecha que existe relación entre los gastos destinados a la promoción (millones) y la cantidad de unidades vendidas (miles). Se desea verificar si las ventas aumentan de forma lineal a los gastos en promoción.¿Qué recomendaría la empresa para mejorar el modelo empírico?.

sugerencia:

❶ Obtenga el modelo lineal y compruebe si modelo es significativo.
❷ Represente gráficamente el modelo ajustado. ¿Cree que otro modelo ajustaría mejor?
❸ Compare el ajuste obtenido con otros modelos de regresión utilizando "Tabular Opcion /Comparison of Alternative Models" (compare resultados).

<u>Nota:</u> Podrá obtener la expresión ajustada para cualquier modelo seleccionando con el botón derecho, las opciones de análisis del menú "Análisis summary"

Publ	25	20	24	12	27	24	22	19	22	18	26	25	20	22
Vent	6.15	4.50	5.27	2.54	6.23	5.97	5.55	4.32	5.14	4.18	5.78	5.65	4.96	5.06

Considere ahora el modelo de regresión múltiple (cuadrática):

$$Ventas=a+ b1*Public+b2*(Public)^2$$

❹ Genere la variable $(Public)^2$ y llámela "Publ2".

❺ Ajuste el modelo propuesto anteriormente y compare el resultado con el obtenido en el apartado ❸. ¿Qué conclusión podemos extraer?.

❻ Realice un análisis completo de la bondad del ajuste anterior.

❼ Establezca si existen puntos raros o influyentes ("Tabular options").

❽ Si una empresa dedicara 30 millones a promocionar sus productos, ¿qué ventas tendría que esperar?.
Nota: Realice este apartado tomando distintos modelos y compare.

Ejemplo 15

En el sector textil de un país, la evolución de la producción,(Q), del capital invertido en planta y equipo (K), y del empleo (L) ha sido la siguiente:

AÑO	PRODUCCIÓN Q	CAPITAL K	EMPLEO L
1975	80	240	60
1976	95	270	70
1977	100	300	80
1978	120	360	90
1979	150	410	100
1980	165	480	120
1981	190	600	130
1982	230	700	140
1983	250	750	160
1984	260	900	180

Dada la función de Cobb-Douglas:

$$Q = \beta L^{\beta 1} K^{\beta 2} e^{\varepsilon}$$

❶ Estime los parámetros β, β1 y β2.
Recuerde que el modelo anterior se puede linealizar tomando logaritmos neperianos:

$$Ln(Q) = Ln(\beta) + Ln(L^{\beta 1}) + Ln(K^{\beta 2}) + Ln(e^{\varepsilon}) \Rightarrow$$
$$Ln(Q) = Ln(\beta) + \beta 1 Ln(L) + \beta 2 Ln(K) + \varepsilon$$

Para introducir las variables del modelo linealizado en el menú “Multiple Regresion”, utilice el botón “transform” y tome logaritmos.

❷ Contraste la significatividad de β1 y β2, de forma individual y conjunta. ¿Ha observado algo inusual?.

❸ Interprete la tabla del análisis de la varianza.

❹ ¿Cuál es la bondad del ajuste?.

❺ Obtenga el gráfico de los valores observados/predichos y el de los residuos. ¿Existen comportamientos anómalos? ¿Existen valores anómalos?, ¿Existen puntos influyentes?, ¿Se observa aleatoriedad en los residuos? .

❻ Prediga la producción del sector textil para el año 2002 si se sabe que el capital invertido será de 1200 y el empleo de 200.

Ejemplo 16

Los datos siguientes reflejan los salarios de un grupo de 14 trabajadores según el tipo de empresa a que pertenecen (estatal, no estatal y profesional libre) y según su sexo (hombre, mujer):

Salario	Tipo empresa	Sexo
190	Estatal	Hombre
200	Estatal	Mujer
250	Estatal	Hombre
90	No Estatal	Mujer
100	No Estatal	Hombre
80	No Estatal	Mujer
110	No Estatal	Hombre
105	No Estatal	Mujer
120	No Estatal	Hombre
200	Libre	Hombre
250	Libre	Hombre
230	Libre	Mujer
260	Libre	Mujer
255	Libre	Mujer

❶ Recodifique las variables Tipo y Sexo.

(Cree dos variables ficticias para Tipo y una para Sexo)

❷ Realice un ajuste múltiple, que nos permita explicar el salario de un trabajador en función del tipo de empresa a la que pertenece y su sexo.

❸ Interprete los coeficientes estimados.

❹ ¿Cree que el modelo obtenido es aceptable?. ¿Porqué?.

❺ ¿Cuál será el salario de una mujer profesional libre?.

Ejemplo 17 (PROPUESTO)

La tabla siguiente contiene el Peso (en Kg), Altura (en cm), Edad y Sexo de 16 individuos:

P	75	81	56	68	79	89	62	59	83	55	72	56	84	61	76	68
A	173	178	162	180	182	185	157	165	180	160	174	161	182	163	172	169
E	21	22	22	21	24	22	21	22	23	22	21	23	22	24	22	21
S	H	H	M	M	H	H	M	M	H	M	H	M	H	M	H	M

(Recodifique la variable Sexo: H=0, M=1)

Realice todos los pasos que crea necesarios para ajustar y verificar un modelo que prediga el Peso de un individuo en función de su Altura, Edad y Sexo.

Ejemplo 18 (PROPUESTO)

La tabla siguiente proporciona la latitud en grados (L), la altura en metros (A) y la temperatura media anual (T) de 6 capitales marítimas españolas.

CAPITAL	L	A	T
Gijón	43.4	22	13.9
Vigo	43.2	45	14.9
Barcelona	41.3	95	16.4
Valencia	39.5	24	17.2
Almería	36.8	7	18
Cádiz	36.5	30	18

(Datos del I.N.E.)

❶ Construya e interprete un modelo de regresión múltiple para explicar la temperatura en función de estas dos variables.

❷ Contraste la hipótesis de que los coeficientes estimados son nulos (individualmente). Seguidamente contraste esta hipótesis de forma conjunta.

❸ Calcule el coeficiente de determinación y la varianza residual.

❹ Prevea la temperatura media de Tortosa sabiendo que la latitud es 40.5 y la altitud 50 m.

❺ Considere otros posibles modelos y diga con cuál de ellos se quedaría.

ANÁLISIS DE DOS VARIABLES: TABLAS DE CONTINGENCIA

Ejemplo 19

Un equipo de trabajadores sociales realizó un estudio sobre 75 pacientes con problemas debidos al consumo de drogas, y trataron de investigar la relación entre el grado de adicción que presentaban y la clase social de sus familias. En principio, este equipo suponía que el nivel social familiar de un adicto a las drogas podía influir en su grado de adicción. A la vista de los resultados, ¿cree que su suposición era cierta?.

Clase Social	**Grado de adicción**		
	Bajo	**Alto**	**Muy Alto**
Baja	3	9	16
Media	7	8	14
Alta	10	5	3

(Cree una variable "Etiqueta" con las etiquetas de las modalidades de la clase social).
Utilice el test Chi-Cuadrado. Hágalo primero de forma manual y luego compare su resultado con el que proporciona Statgraphics.

Ejemplo 20

De 1000 conductores de turismo, 734 son hombres y 266 mujeres; 100 hombres han sufrido accidentes, mientras que las mujeres accidentadas son 14. Determine si existe asociación entre el sexo del conductor y el hecho de tener o no accidentes.

Ejemplo 21

Doce individuos se clasificaron según el sexo (varón, mujer) y su deseo de ver o no una final de campeonato de fútbol que será televisada. (Utilice "Crosstabulation").

Sexo	V	H	V	V	V	H	H	H	V	V	V	H
Fútbol	SI	NO	SI	NO	SI	NO	NO	SI	SI	SI	SI	NO

❶ Obtenga la tabla de frecuencias cruzadas.
❷ ¿Son independientes las variables Sexo e interés por el fútbol?.

Ejemplo 22 (Propuesto)

En un estudio sobre los antecedentes personales y sociales de los líderes del movimiento nazi, se investigaron los historiales de 15 hombres que constituyeron el gabinete alemán a finales de 1934. Estos hombres fueron clasificados en dos grupos: nazis (N) y no nazis (NN). Para probar la hipótesis de que los líderes nazis habían hecho del trabajo partidista su carrera, mientras que los no nazis habían surgido de otras ocupaciones más estables y convencionales, se clasificaron de acuerdo con el primer trabajo de su carrera: ocupación estable (E) o relacionado con el partido (P). Los resultados fueron:

Ocupación	**Ideología**	
	N	**NN**
E	1	6
P	8	0

Confirme o no esta hipótesis.

Ejemplo 23 (Propuesto)

En un centro se dispone de tres libros de texto para impartir Estadística (L1, L2, L3). Se supone que la calidad de los textos influye en las calificaciones de los alumnos. Para comprobar si esta suposición es real, se realiza un experimento consistente en que tres grupos de alumnos estudien con cada uno de los libros. Los resultados fueron:

LIBROS	**CALIFICACIONES**			
	Susp	**Aprob**	**Notab**	**Sobr**
L1	10	32	12	40
L2	14	12	10	30
L3	8	10	6	16

¿Es lógico admitir esta suposición?.