

Estadística II
Curso 2010/11

Guión de la Práctica 1
Introducción a Statgraphics. Intervalos de confianza y contrastes de hipótesis para una y dos poblaciones

1. Contenidos de la práctica

- [Introducción a Statgraphics](#)
- [Intervalos de confianza y contrastes de hipótesis para una y dos poblaciones](#)

2. Introducción a Statgraphics

Statgraphics es un paquete estadístico de Software comercial basado en ventanas y orientado a la docencia, que incluye herramientas para tratar distintas técnicas estudiadas en Estadística I y II.

En los guiones de las prácticas de la asignatura se presentan indicaciones de uso de Statgraphics y ejemplos para su aplicación a estos temas. Estos guiones se han preparado para la versión de Statgraphics Plus para Windows v. 4.0, en inglés.

En ellos se mencionan únicamente aquellas opciones directamente relacionadas con los contenidos de las asignaturas. Si quisieras más información sobre Statgraphics puedes encontrarla en distintos recursos en la Web o en la propia ayuda de Statgraphics.

2.1 Contenido de la introducción

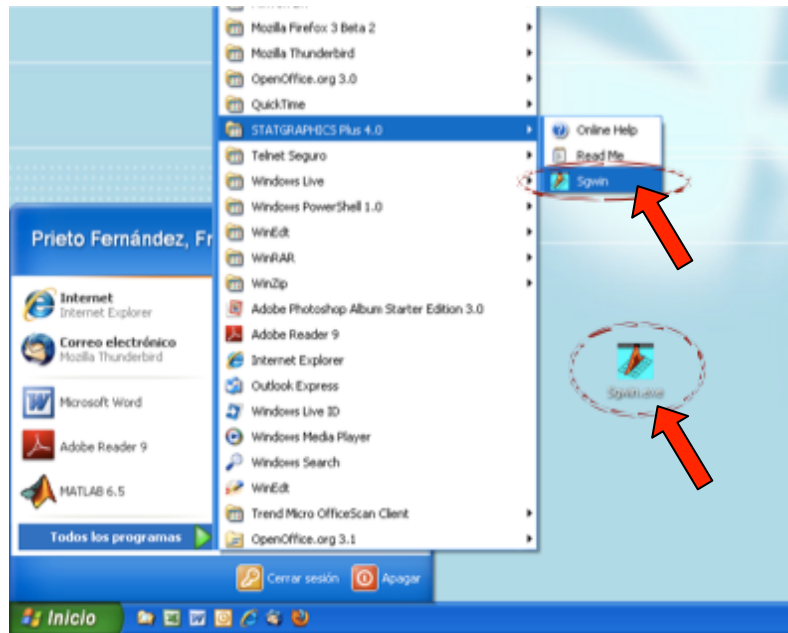
- [Cómo iniciar Statgraphics](#)
- [Cómo leer datos de un fichero en Statgraphics](#)
- [Obtención de estadísticos descriptivos básicos](#)
- [Generación de gráficos](#)
- [Cuantiles y probabilidades de distribuciones](#)

2.2 Cómo iniciar Statgraphics

Como para otros programas de Windows, puedes ir al menú de “Inicio” de Windows y seleccionar

Inicio → Todos los programas → Statgraphics Plus 4.0 → Sgwin

Alternativamente, si dispones de un icono en el Escritorio de Windows (usualmente con el nombre “Sgwin.exe”), puedes pulsar dos veces en dicho icono.



2.3 Cómo leer datos de un fichero en Statgraphics

Al iniciar Statgraphics habitualmente se inicia el “StatWizard”, un asistente para facilitar el inicio del trabajo con Statgraphics (este asistente se puede desactivar desmarcando la casilla inferior izquierda en la ventana de arranque).

Seleccionaremos en el “StatWizard”

Analyze existing data or Enter new data ... → OK

A continuación al preguntarnos donde están nuestros datos, seleccionaremos

In an existing Statgraphics Plus data file ... → OK

para cargar alguno de los conjuntos de datos ya disponibles.

Se abrirá a continuación una ventana con el título “Open Data File” y con el explorador de Windows en la carpeta “Test Data”. Para esta parte de la práctica vamos a seleccionar el fichero “Cardata.sf”, que contiene información sobre distintas características de automóviles disponibles en el mercado Norteamericano hace años. Una vez seleccionado dicho archivo y tras pulsar “Aceptar” obtendrás una ventana como la que se indica a continuación, y estarás preparada para analizar estos datos.

Alternativamente, este fichero se encuentra disponible también en la página Web de prácticas de la asignatura (Datos_1).

Antes de seguir más adelante, examina brevemente los datos y como están organizados. ¿Cuántas variables están incluidas en el conjunto de datos? ¿De cuántas observaciones se dispone para cada variable? ¿De qué tipo es cada variable?

The screenshot shows the STATGRAPHICS Plus interface with a data table titled 'Car data.sf'. The table contains 22 rows of data with the following columns: mpg, cylinders, displace, horsepower, accel, and year. The first row is highlighted.

	mpg	cylinders	displace	horsepower	accel	year
1	43,1	4	98	48	21,5	78
2	36,1	4	98	66	18,4	78
3	32,8	4	78	52	19,4	78
4	39,4	4	85	70	18,6	78
5	36,1	4	91	68	16,4	78
6	19,9	8	260	110	15,5	78
7	19,4	8	318	148	13,2	78
8	20,2	8	302	139	12,8	78
9	19,2	6	231	185	19,2	78
10	20,5	6	200	95	18,2	78
11	20,2	6	200	85	15,8	78
12	25,1	4	140	88	15,4	78
13	20,5	6	225	100	17,2	78
14	19,4	6	232	98	17,2	78
15	20,6	6	231	185	15,8	78
16	20,8	6	200	85	16,7	78
17	18,6	6	225	110	18,7	78
18	18,1	6	258	128	15,1	78
19	19,2	8	305	145	13,2	78
20	17,7	6	231	165	13,4	78
21	18,1	8	302	139	11,2	78
22	17,5	8	318	148	13,7	78

Esta lectura de datos de un fichero se puede hacer también desde el menú “File”,

File → Open → Open Data File ...

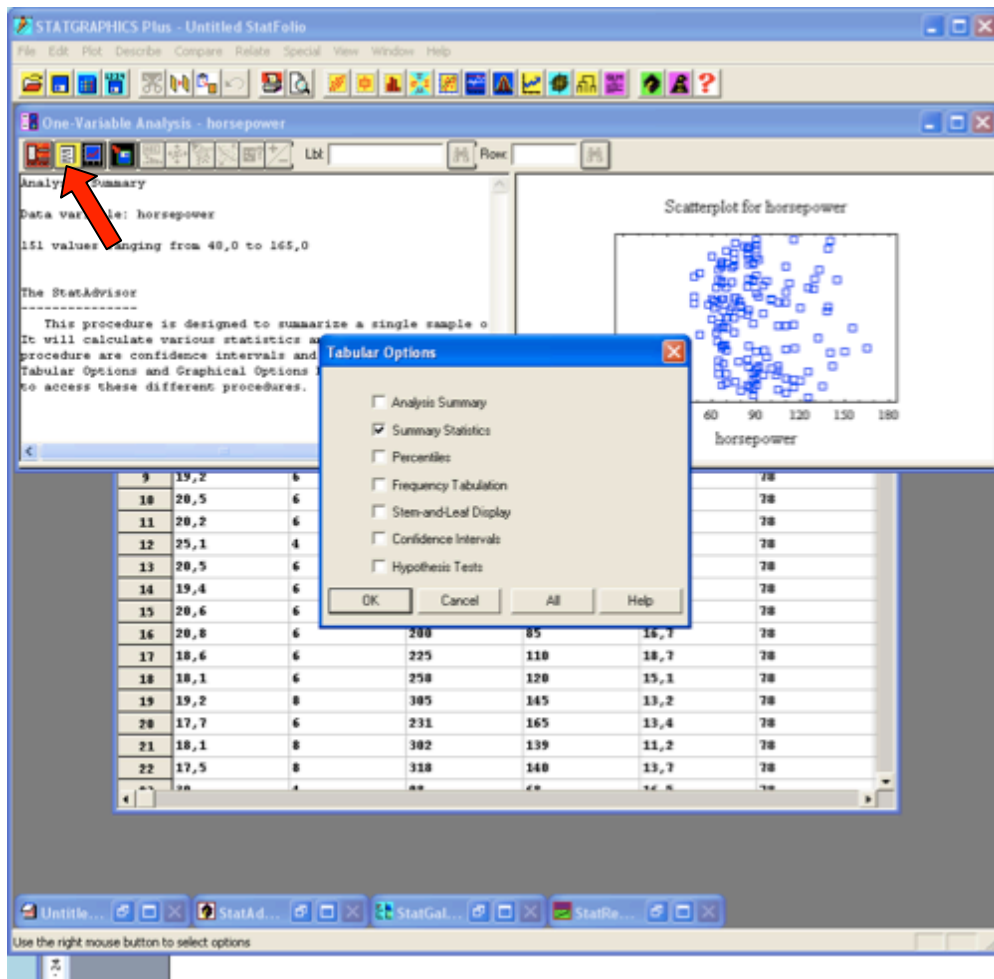
2.4 Obtener estadísticos descriptivos básicos

Para obtener un resumen de los datos seleccionamos en el menú “Describe”,

Describe → Numeric Data → One-Variable Analysis ...

Se abre una nueva ventana de diálogo en la que se nos pide que seleccionemos la variable con la que queremos trabajar. Seleccionamos la variable “Horsepower” (potencia) y aparecen dos nuevas ventanas con indicaciones del “StatAdvisor” y un gráfico de dispersión de la variable.

Para obtener los estadísticos descriptivos básicos, seleccionamos el segundo icono de la ventana desde la izquierda (con el texto flotante “Tabular options”).



En la ventana de diálogo que aparece seleccionamos “Summary Statistics”,

(I) Tabular options → Summary Statistics → OK

Se obtiene un resumen de la información en esa variable, que incluye los datos siguientes:

Count = 151	(<i>número de observaciones</i>)
Average = 89,0	(<i>promedio</i>)
Variance = 596,533	(<i>varianza</i>)
Standard deviation = 24,424	(<i>desviación típica</i>)
Minimum = 48,0	(<i>mínimo</i>)
Maximum = 165,0	(<i>máximo</i>)
Range = 117,0	(<i>recorrido</i>)
Std. skewness = 4,27089	(<i>asimetría</i>)
Std. kurtosis = 0,603367	(<i>kurtosis</i>)

Repite el análisis para un par de variables, seleccionando

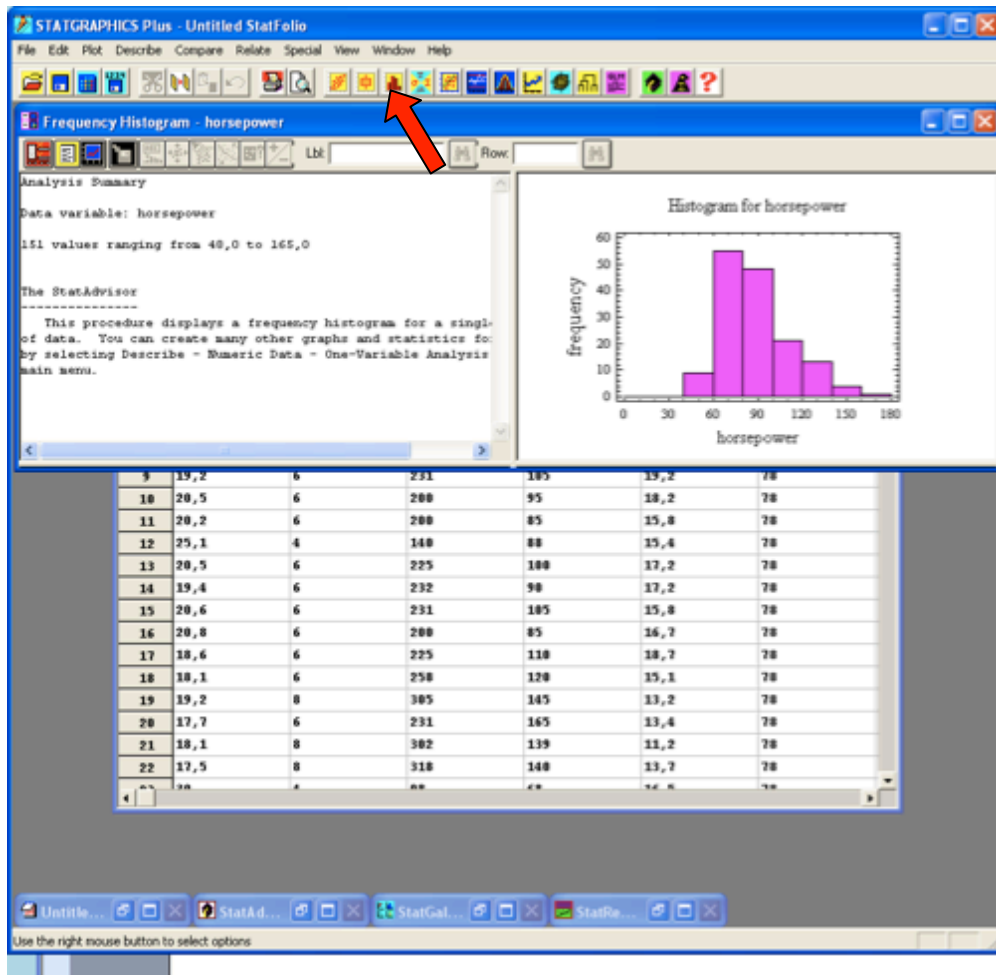
Describe → Numeric Data → Multiple-Variable Analysis ...

y a continuación las variables “Horsepower” y “Mpg” (consumo). Obtén los valores de la covarianza de ambas variables y su coeficiente de correlación.

2.5 Generación de gráficos

En la parte superior de la ventana, bajo los menús en texto, aparecen una serie de iconos. El cuarto bloque de iconos permite generar gráficos para los datos correspondientes a una o varias variables. Por ejemplo, el tercer icono de dicho grupo permite obtener los histogramas correspondientes a las diferentes variables.

En la ventana emergente seleccionamos la variable “Horsepower” y obtenemos el histograma que se indica a continuación:



Las características del gráfico, por ejemplo el número de cajas, se pueden modificar pulsando el botón derecho en la ventana del histograma y seleccionando

(BD) Pane Options → Number of Classes → 10

Genera un diagrama de caja para la variable “Mpg”.

2.6 Cuantiles y probabilidades de distribuciones

En muchos casos deseamos trabajar con distribuciones, sus cuantiles y los valores de sus funciones de distribución (probabilidades de sus colas, por ejemplo). Estos procesos se corresponden por ejemplo con la consulta de tablas de distribuciones que hemos llevado a cabo en clase.

Para obtener esta información en Statgraphics seleccionamos en el menú “Describe”

Describe → Distributions → Probability Distributions ...

En el menú emergente marcamos la distribución con la que queramos trabajar, por ejemplo la t de Student (“Student’s t”), y Statgraphics abre dos nuevas ventanas para dicha distribución (por defecto para 10 grados de libertad).

Si quisiéramos calcular la probabilidad a la derecha del valor 1,5 para una distribución t de Student con 24 grados de libertad, el procedimiento sería:

- Seleccionamos con el botón derecho en la primera ventana la opción “Analysis Options”, y escribimos 24 en la ventana de diálogo bajo “Deg. Of Freedom” (los grados de libertad).

(BD) Analysis Options → Deg. of Freedom → 24

- Pulsamos en el icono de “Tabular Options” (el segundo icono desde la izquierda) y marcamos la casilla “Cumulative Distribution”.

(I) Tabular Options → Cumulative Distribution

- Seleccionamos con el botón derecho en la primera ventana la opción “Pane Options”, y escribimos 1,5 en la ventana de diálogo bajo “Random Variable” (el valor de la variable).

(BD) Pane Options → Random Variable → 1,5

Obtenemos como resultado una salida que en parte nos indica:

```
Cumulative Distribution
-----
Distribution: Student's t

Variable      Lower Tail Area (<)
1,5           Dist. 1      Dist. 2      Dist. 3      Dist. 4      Dist. 5
              0,926672

Variable      Probability Density
1,5           Dist. 1      Dist. 2      Dist. 3      Dist. 4      Dist. 5
              0,128799

Variable      Upper Tail Area (>)
1,5           Dist. 1      Dist. 2      Dist. 3      Dist. 4      Dist. 5
              0,0733278
```

El valor de interés (la cola derecha) es 0,0733278.

Podemos también obtener el valor correspondiente a una determinada probabilidad, por ejemplo el cuantil correspondiente al 98% (el valor de la variable que deja a su derecha una probabilidad del 2%) para una t de Student con 24 grados de libertad. Para ello, tras seleccionar “Tabular Options” marcamos “Inverse CDF”, y seleccionamos con el botón derecho en la primera ventana la opción “Pane Options”. Introducimos ahora la probabilidad deseada bajo “CDF” (se han borrado los demás valores que aparecen en el cuadro de diálogo),

(I) Tabular Options → Inverse CDF
(BD) Pane Options → CDF → 0,98

Obtenemos:

Inverse CDF

Distribution: Student's t

CDF	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,98	2,17155				

El valor deseado es 2,17155.

Encuentra la probabilidad a la derecha del valor 32,37 bajo una distribución chi cuadrado (Chi-square) con 19 grados de libertad. Encuentra también los cuantiles correspondientes al 2,5% y al 97,5% para dicha distribución.

3. Intervalos de confianza y contrastes de hipótesis para una y dos poblaciones

3.1 Contenido de intervalos y contrastes

- [Intervalos de confianza para una población](#)
- [Contrastes de hipótesis para una población](#)
- [Contrastes de hipótesis para dos poblaciones](#)

3.2 Intervalos de confianza para una población

Vamos a trabajar en esta parte con uno de los ejercicios propuestos para resolver en clase en el Tema 1 de la asignatura (el ejercicio número 13). Su enunciado era:

“En una clínica se ofrecen tratamientos para la reducción de peso en pacientes. Se supone que la disminución de peso que se observa tras un tratamiento de dos meses sigue una distribución normal. De una muestra aleatoria simple de 16 pacientes se han obtenido los valores de reducción que se indican a continuación:

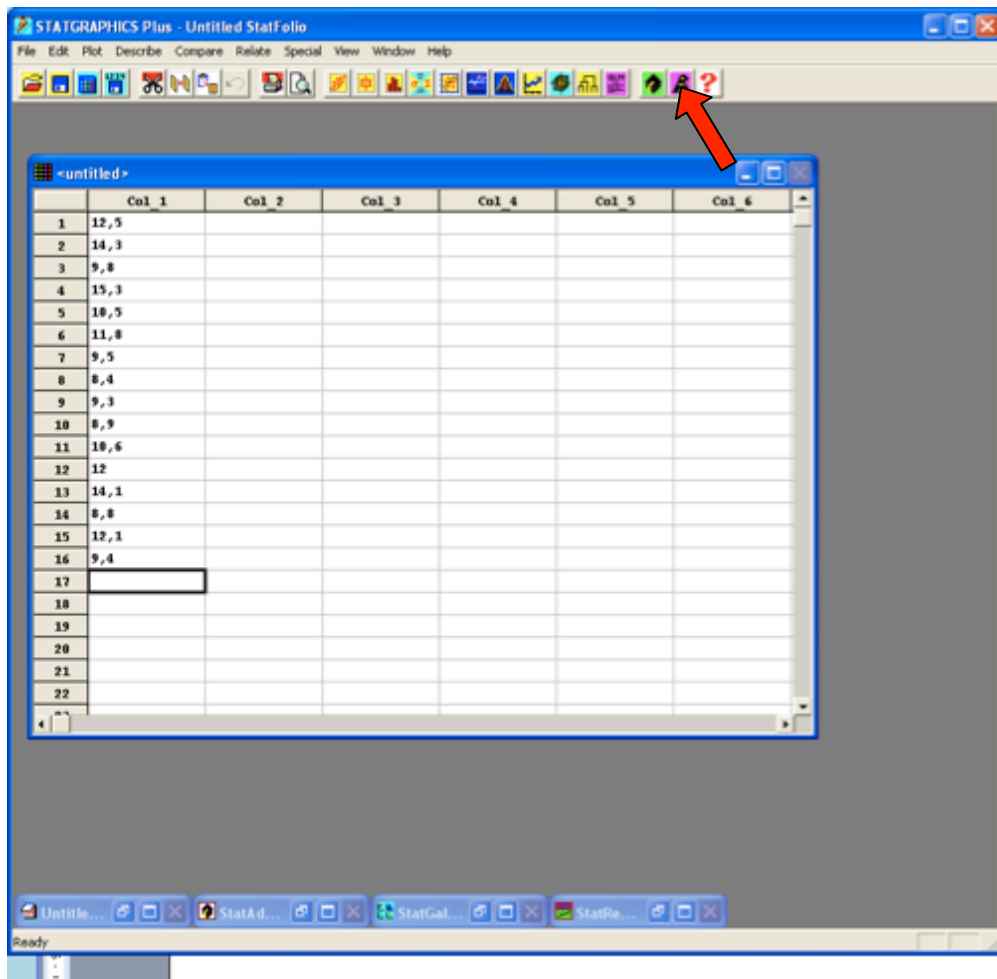
12,5 14,3 9,8 15,3 10,5 11,8 9,5 8,4
9,3 8,9 10,6 12,0 14,1 8,8 12,1 9,4

Calcula un intervalo de confianza al 95% para la reducción de peso media en todos los pacientes que siguen el tratamiento.”

Empezamos por introducir los datos del problema en Statgraphics. Para ello podemos importar los datos desde un fichero de texto o una hoja de cálculo, utilizando el StatWizard (el segundo botón desde la derecha, ver imagen) o bien introduciendo los datos directamente cerrando la ventana de datos actual seleccionando en el menú “File”,

File → Close → Close Data File

Si hacemos esto último e introducimos los datos anteriores (en una columna) en la ventana de datos vacía que aparece (“<untitled>”), tenemos



Estos datos se pueden guardar para su uso posterior seleccionando en el menú “File”,

File → Save As → Save Data File As ...

e indicando la carpeta y el nombre del archivo deseado.

El resumen de estos datos, obtenido seleccionando en el menú “Describe”,

Describe → Numeric Data → One-Variable Analysis ...
 Col_1
 (I) Tabular Options → Summary Statistics

donde tras seleccionar “One-Variable Analysis ...” hemos introducido en el cuadro de diálogo “Col_1”, y tras obtener las dos ventanas generadas por Statgraphics, hemos pulsado el icono de “Tabular options” y marcado “Summary Statistics”. El resultado es el siguiente:

```
Summary Statistics for Col_1
Count = 16
Average = 11,0813
Variance = 4,63629
Standard deviation = 2,1532
Minimum = 8,4
Maximum = 15,3
Range = 6,9
```


Std. skewness = 1,03079
Std. kurtosis = -0,599163

Para calcular el intervalo de confianza podríamos aplicar la fórmula vista en clase, esto es,

$$\bar{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$$

para lo que debemos calcular el cuantil correspondiente de la distribución t de Student, $t_{n-1,\alpha/2}$, seleccionando en el menú “Describe” como se indica en la sección “[Cuantiles y probabilidades de distribuciones](#)”,

Describe → Distributions → Probability Distributions ... → Student's t
(BD) Analysis Options → Deg. of Freedom → 15
(I) Tabular Options → Inverse CDF
(BD) Pane Options → CDF → 0,975

Obtenemos

Inverse CDF

Distribution: Student's t

CDF	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,975	2,13145				

y el intervalo será el correspondiente a

$$11,0813 \pm 2,13145 \times 2,1532/\sqrt{16} = [9,93394;12,22866]$$

Afortunadamente, existe otra manera mucho más rápida de realizar estos cálculos. Para ello, desde la ventana con los datos que hemos introducido seleccionamos en el menú “Describe”,

Describe → Numeric Data → One-Variable Analysis ...
Col_1
(I) Tabular Options → Confidence Intervals

Tras seleccionar “One-Variable Analysis ...” en el menú emergente seleccionamos “Col_1”, y una vez que aparecen las dos ventanas, pulsamos el botón “Tabular options” y marcamos “Confidence Intervals”.

Una vez completado este proceso, podemos seleccionar el nivel de confianza (cambiar el valor que aparece por defecto) pulsando el botón derecho del ratón en la ventana izquierda y seleccionando “Pane Options”. En la ventana emergente introducimos el valor deseado.

(BD) Pane Options → Confidence Level → 95

Para el 95% (el valor por defecto) obtenemos

Confidence Intervals for Col_1

95,0% confidence interval for mean: 11,0813 +/- 1,14736 [9,93389;12,2286]

95,0% confidence interval for standard deviation: [1,59058;3,3325]

Para el conjunto de datos "Cardata.sf", calcula un intervalo de confianza para la media de la variable "Mpg" al 99%.

3.3 Contrastes de hipótesis para una población

Veremos en esta sección como llevar a cabo los cálculos asociados a los contrastes de hipótesis que estudiamos en el Tema 2 mediante Statgraphics. Como ejemplo, supongamos que queremos contrastar con los datos del problema anterior si la reducción de peso obtenida con el tratamiento es significativamente superior a 10 Kg. Supongamos también que queremos hacerlo para un nivel de significación del 5%.

De nuevo podríamos llevar a cabo los cálculos manualmente, y en particular si llevamos a cabo un contraste unilateral y quisiéramos obtener el p-valor correspondiente calcularíamos el valor del estadístico como

$$(x - 10)/(s/\sqrt{n}) = (11,0813 - 10)/(2,1532/\sqrt{16}) = 2.008731$$

Podemos buscar la probabilidad para la cola derecha de este valor, para una distribución t de Student, seleccionando el menú "Describe", como se indica en la sección "[Cuantiles y probabilidades de distribuciones](#)",

Describe → Distributions → Probability Distributions ... → Student's t
(BD) Analysis Options → Deg. of Freedom → 15
(I) Tabular Options → Cumulative Distribution
(BD) Pane Options → Random Variable → 2,008731

Obtenemos

Cumulative Distribution

Distribution: Student's t

Variable	Lower Tail Area (<)				
	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,00873	0,968541				

Variable	Probability Density				
	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,00873	0,0583424				

Variable	Upper Tail Area (>)				
	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,00873	0,0314594				

El p-valor para el contraste unilateral será por tanto 0,0314594, y con ese valor y el nivel de significación indicado rechazaríamos la hipótesis nula.

De nuevo, existe una manera más sencilla de llevar a cabo este contraste. Para ello seleccionamos en el menú "Describe",

Describe → Numeric Data → One-Variable Analysis ...
Col_1
(I) Tabular Options → Hypothesis Tests

Al hacerlo obtenemos

Hypothesis Tests for Col_1

Sample mean = 11,0813
Sample median = 10,55

t-test

Null hypothesis: mean = 0,0
Alternative: not equal

Computed t statistic = 20,5856
P-Value = 2,07612E-12

Reject the null hypothesis for alpha = 0,05.

Se obtiene también otra información adicional que se ha omitido.

Obsérvese que el análisis se ha llevado a cabo para una hipótesis alternativa bilateral y para un valor del parámetro en la hipótesis nula igual a cero (las opciones por defecto). Para cambiar estas opciones a una alternativa unilateral y a un valor del parámetro igual a 10 para la hipótesis nula, debemos pulsar el botón derecho del ratón en la ventana izquierda y seleccionar “[Pane Options ...](#)”. En el diálogo introducimos “[Mean](#)” igual a 10, marcamos “[Greater than](#)” y dejamos “[Alpha](#)” en el 5%.

(BD) [Pane Options](#) → [Mean](#) → 10
[Greater Than](#)

Obtenemos ahora

Hypothesis Tests for Col_1

Sample mean = 11,0813
Sample median = 10,55

t-test

Null hypothesis: mean = 10,0
Alternative: greater than

Computed t statistic = 2,00863
P-Value = 0,0314651

Reject the null hypothesis for alpha = 0,05.

Este resultado coincide por supuesto con el análisis que habíamos realizado anteriormente.

Otra manera de llevar a cabo este contraste es seleccionar el menú “[Describe](#)”, “[Hypothesis Tests ...](#)”, marcamos “[Normal mean](#)” ya que estamos llevando a cabo un contraste sobre la media de una población normal, e introducimos en las casillas correspondientes los valores

Null Hypothesis:	10	(<i>valor de la media bajo la hipótesis nula</i>)
Sample Mean:	11,0813	(<i>media de la muestra</i>)
Sample Sigma:	2,1532	(<i>cuasidesviación típica de la muestra</i>)
Sample Size:	16	(<i>tamaño de la muestra</i>)

[Describe](#) → [Hypothesis Tests ...](#) → [Normal Mean](#)

Para el conjunto de datos "Cardata.sf", contrasta si existe evidencia suficiente para rechazar que la eficiencia de los automóviles no ha cambiado desde el año anterior, cuando el valor promedio entre todos los vehículos era de 30,3 mpg. Lleva a cabo el contraste para un nivel de significación del 1%.

3.4 Contrastes de hipótesis para dos poblaciones

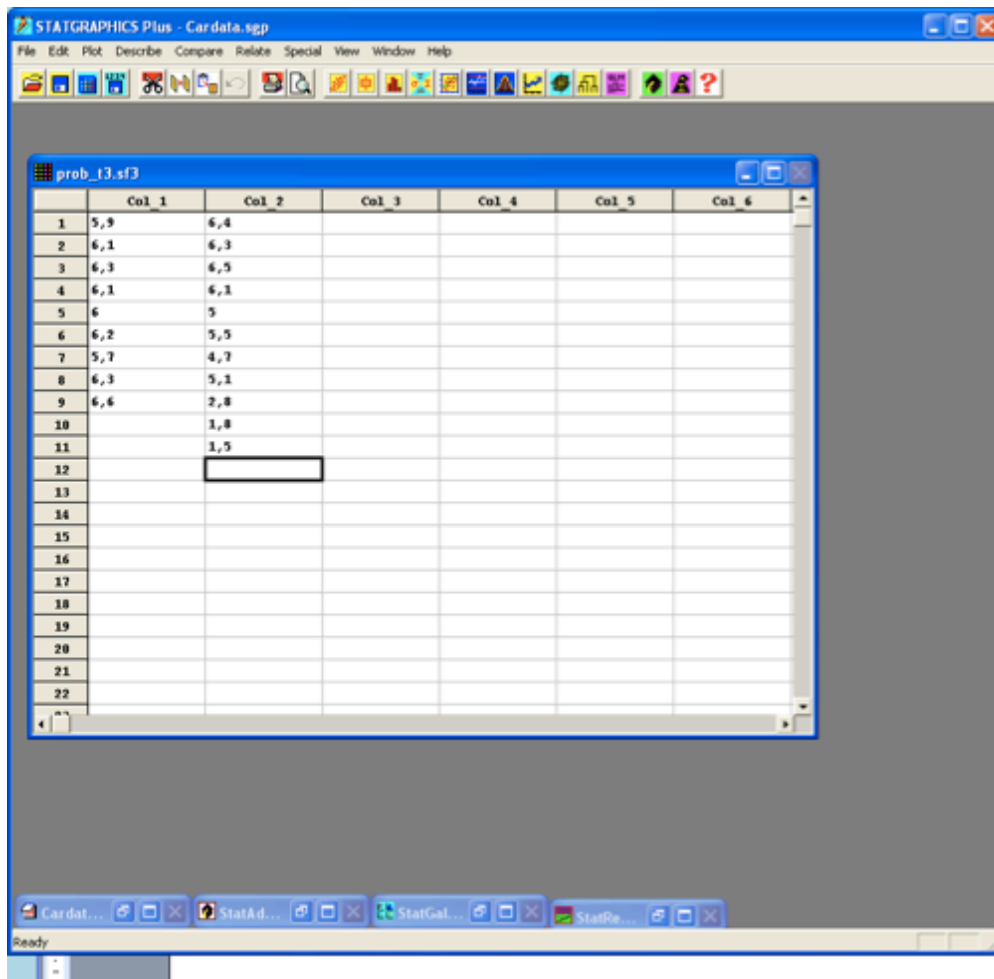
Al igual que en caso anterior empezamos tomando como referencia uno de los ejercicios de la colección correspondiente al Tema 3, en particular el ejercicio número 10. Dicho ejercicio planteaba lo siguiente:

"Estamos interesados en saber si los salarios medios en dos grandes ciudades de un país son homogéneos o si, por el contrario, se puede evidenciar que hay diferencias significativas en los salarios que perciben los trabajadores de estas 2 ciudades. Para ello se tomaron muestras aleatorias independientes de trabajadores (de actividades similares y niveles similares) en las 2 ciudades y se obtuvieron los siguientes datos:

<i>Ciudad 1</i>	5.9	6.1	6.3	6.1	6	6.2	5.7	6.3	6.6		
<i>Ciudad 2</i>	6.4	6.3	6.5	6.1	5	5.5	4.7	5.1	2.8	1.8	1.5

Plantea un contraste bilateral para ver si hay diferencias significativas en los salarios que perciben los trabajadores de estas 2 ciudades. ¿A qué conclusión llegas para un nivel de significación de 0,1? Da una cota para el p-valor."

En primer lugar, debemos introducir los datos en Statgraphics, en dos columnas consecutivas. Para ello cerramos las ventanas de datos que tengamos abiertas hasta llegar a una ventana de datos vacía ("<untitled>") e introducimos los valores indicados en las dos primeras columnas de la ventana. Obtenemos la ventana que se indica a continuación.



Si queremos estudiar este contraste (como un contraste bilateral) siguiendo el procedimiento que hemos descrito en clase, podemos calcular el valor del estadístico correspondiente a partir de los valores de cada muestra. Para calcular los estadísticos descriptivos básicos simultáneamente para ambas muestras podemos seleccionar en el menú “Compare”,

Compare → Two Samples → Two-Sample Comparison ...
 Col_1 , Col_2
 (I) Tabular Options → Summary Statistics

En el menú emergente tras seleccionar “Two-Sample Comparison ...” indicamos “Col_1” y “Col_2” como muestras 1 y 2 (“Sample 1” y “Sample 2”). Una vez que aparecen las dos ventanas pulsamos el botón “Tabular Options” y marcamos la opción “Summary Statistics”. Obtenemos

Summary Statistics	Col_1	Col_2
Count	9	11
Average	6,13333	4,7
Variance	0,0675	3,38
Standard deviation	0,259808	1,83848
Minimum	5,7	1,5
Maximum	6,6	6,5
Range	0,9	5,0
Std. skewness	0,164617	-1,22047
Std. kurtosis	0,351847	-0,449121

Con estos valores podemos calcular el estadístico del contraste, y tenemos que

$$s_p^2 = (8 \times 0,0675 + 10 \times 3,38) / 18 = 1,907778$$
$$t = (6,13333 - 4,7) / (\sqrt{1,907778 \times \sqrt{(1/9 + 1/11)}}) = 2,308792$$

Con este dato podemos calcular el p-valor del contraste para la distribución del mismo, una t de Student con 18 grados de libertad. Seleccionamos en el menú “Describe”, como se indica en la sección [“Cuantiles y probabilidades de distribuciones”](#),

Describe → Distributions → Probability Distributions ... → Student's t
(BD) Analysis Options → Deg. of Freedom → 18
(I) Tabular Options → Cumulative Distribution
(BD) Pane Options → Random Variable → 2,308792

Obtenemos finalmente

Cumulative Distribution

Distribution: Student's t

Variable	Lower Tail Area (<)	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,30879	0,983483					

Variable	Probability Density	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,30879	0,0334725					

Variable	Upper Tail Area (>)	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
2,30879	0,0165173					

Al tratarse de un contraste bilateral, su p-valor será $2 \times 0,0165173 = 0,0330346$, y no rechazamos la hipótesis nula para un nivel de significación del 1%.

Si queremos realizar este mismo contraste de una manera más directa, podemos seleccionar en el menú “Compare”,

Compare → Two Samples → Two-Sample Comparison ...
Col_1 , Col_2
(I) Tabular Options → Comparison of Means

En la ventana emergente tras seleccionar “Two-Sample Comparison ...”, indicamos “Col_1” y “Col_2” como muestras 1 y 2. En las dos ventanas que obtenemos pulsamos el botón “Tabular Options” y marcamos la opción “Comparison of Means”. Obtenemos como resultado

Comparison of Means

95,0% confidence interval for mean of Col_1: 6,13333 +/- 0,199706
[5,93363,6,33304]
95,0% confidence interval for mean of Col_2: 4,7 +/- 1,23511
[3,46489,5,93511]
95,0% confidence interval for the difference between the means

assuming equal variances: 1,43333 +/- 1,30428 [0,129049,2,73762]

t test to compare means

Null hypothesis: mean1 = mean2

Alt. hypothesis: mean1 NE mean2

assuming equal variances: t = 2,3088 P-value = 0,0330343

Estos son los mismos resultados que vimos anteriormente. Si quisiéramos cambiar el tipo de contraste a uno unilateral, o modificar el nivel de confianza para los intervalos, podemos pulsar con el botón derecho en la primera ventana, seleccionar “[Pane Options](#)” y cambiar estos datos en la ventana que aparece.

Si las muestras fuesen pareadas, bastaría con seleccionar en el menú “[Compare](#)”,

Compare → Two Samples → Paired-Sample Comparison ...

y seguir un procedimiento similar.

En la página Web de la asignatura tienes un conjunto de datos denominado Datos_2, “car_mpg.sf3”. Estos datos corresponden a consumos de automóviles fabricados en Europa, Japón y los EEUU. Lee dichos datos en Statgraphics y lleva a cabo contrastes de hipótesis para determinar si en función de estos datos puedes concluir que los consumos medios son diferentes para automóviles fabricados en las distintas zonas.