

Intervalos de confianza con Statgraphics Centurion

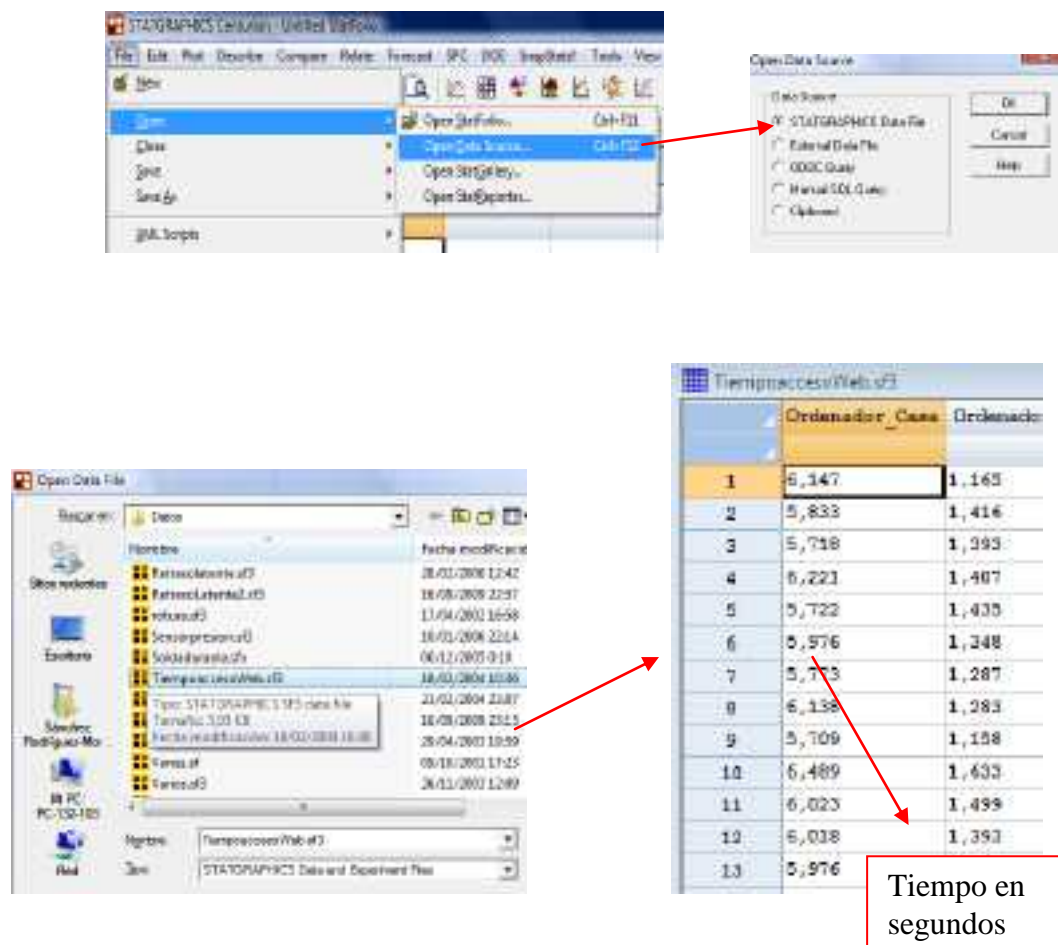
Ficheros empleados: TiempoaccesoWeb.sf3 ; TiempoBucle.sf3;

1. Ejemplo 1: Tiempo de acceso a una página Web

Se desean construir intervalos de confianza para la media μ y la desviación típica σ de la distribución del **tiempo de acceso** a la página web de la UC3M desde un domicilio particular así como desde la universidad. Los intervalos de confianza de los parámetros se construirán usando la información que proporciona una muestra de 55 datos del tiempo (en segundos) que se tarda en acceder a la página web de la UC3M. Las mediciones se hacen desde dos ordenadores: desde un domicilio particular y desde la universidad (fichero TiempoaccesoWeb.sf3)

1.1 Entrada de datos:

Lo primero que hacemos es leer ese fichero de datos.



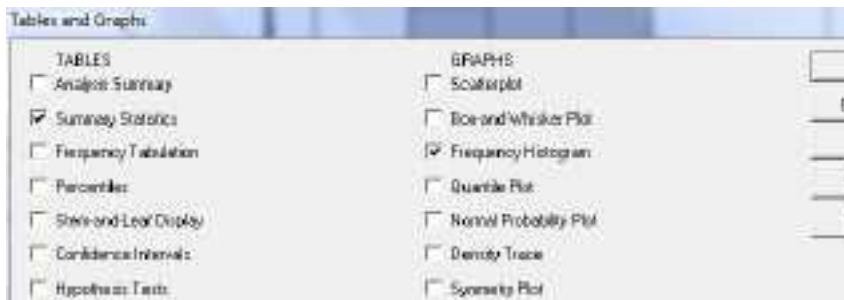
The first screenshot shows the Statgraphics Centurion interface with the 'Open' menu open. The 'Open Data Source...' option is highlighted. A red arrow points to the 'Open Data Source' dialog box in the second screenshot, which has 'STATGRAPHICS Data File' selected. The third screenshot shows the 'Open Data File' dialog box with 'TiempoaccesoWeb.sf3' selected. The fourth screenshot shows the data file loaded into a table with columns 'Ordenador_Casa', 'Ordenador', and 'Tiempo'.

	Ordenador_Casa	Ordenador	Tiempo
1	5,147	1,163	
2	5,833	1,416	
3	5,718	1,393	
4	6,221	1,407	
5	5,722	1,435	
6	5,576	1,348	
7	5,773	1,287	
8	6,138	1,283	
9	5,709	1,158	
10	6,489	1,633	
11	6,023	1,439	
12	6,038	1,393	
13	5,976		

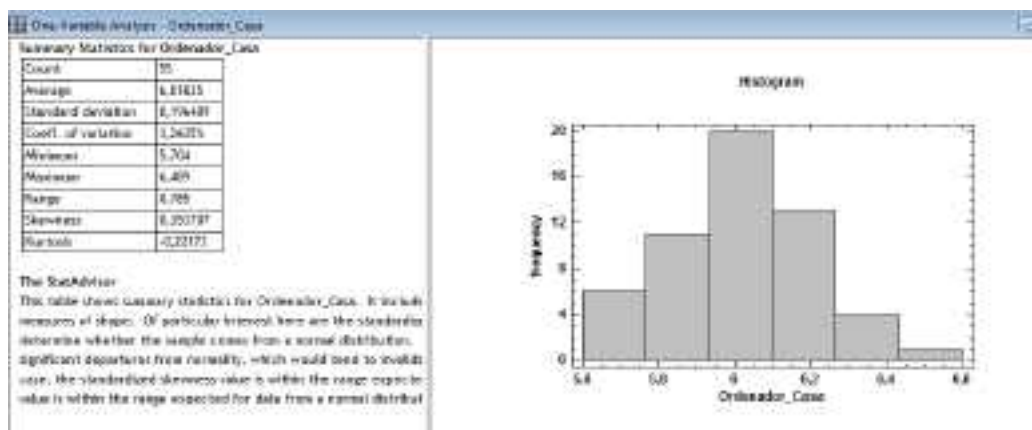
Tiempo en segundos

1.2. Análisis univariante de la variable

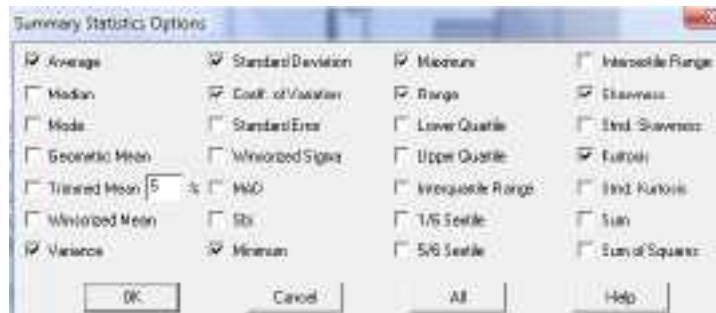
Antes de realizar cualquier análisis conviene hacer una descripción de las variables. Vamos a comenzar analizando la variable tiempo de acceso a la web desde un domicilio particular (variable Ordenador_Casa). El análisis gráfico y numérico de esta variable se hace en: Describe / Numeric Data / One-Variable Análisis, y seleccionamos la variable Ordenador_Casa. En Tables and Graphs seleccionamos el resumen estadístico y el histograma:



El resultado es el siguiente:



Los estadísticos seleccionados en el Summary Statistics se han elegido en Pane Options



Lo importante en este histograma es que la variable Ordenador_Casa tiene una distribución parecida a la normal: es bastante simétrica y con forma de campana. La hipótesis de normalidad es importante para calcular

intervalos de confianza. Por ejemplo, sólo podremos hacer intervalos de confianza para la varianza de un población basados en la distribución chi-cuadrado si la población es normal.

El resumen estadístico incluye las medidas de tendencia central, medidas de variabilidad y medidas de forma. Entre los valores obtenidos se encuentran la media y varianza muestrales que son estimaciones puntuales de la media y la varianza poblacionales. Es decir, tenemos que en esta muestra, las estimaciones ‘puntuales’ de los parámetros que nos interesan son

$$\hat{\mu} = 6.018$$

$$\hat{\sigma}^2 = 0.0386$$

Nuestro objetivo es obtener estimaciones ‘por intervalos’ de esos parámetros.

1.3. Análisis de la normalidad de la variable

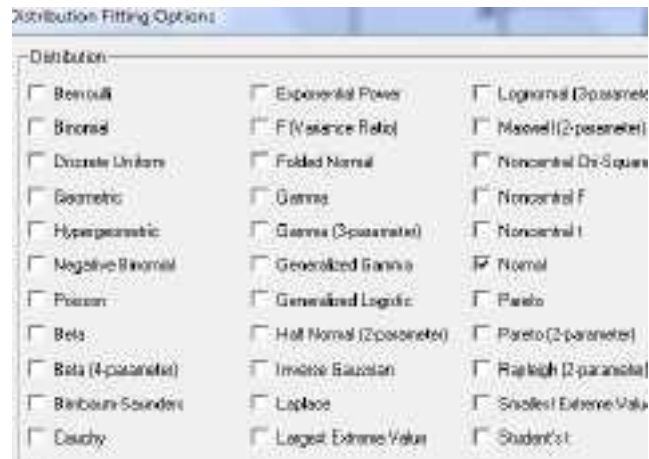
Vamos a realizar un contraste de normalidad mediante el test de la Chi-cuadrado, que nos refuerce nuestra hipótesis de que la población de la que procede nuestra muestra es normal. Seleccionamos Describe / Distribution Fitting/Fitting Uncensored Data



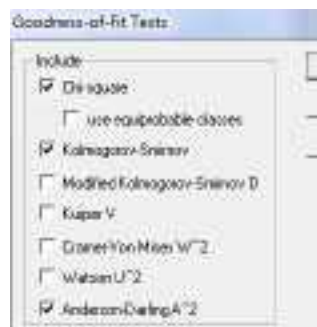
Aparece entonces la ventana



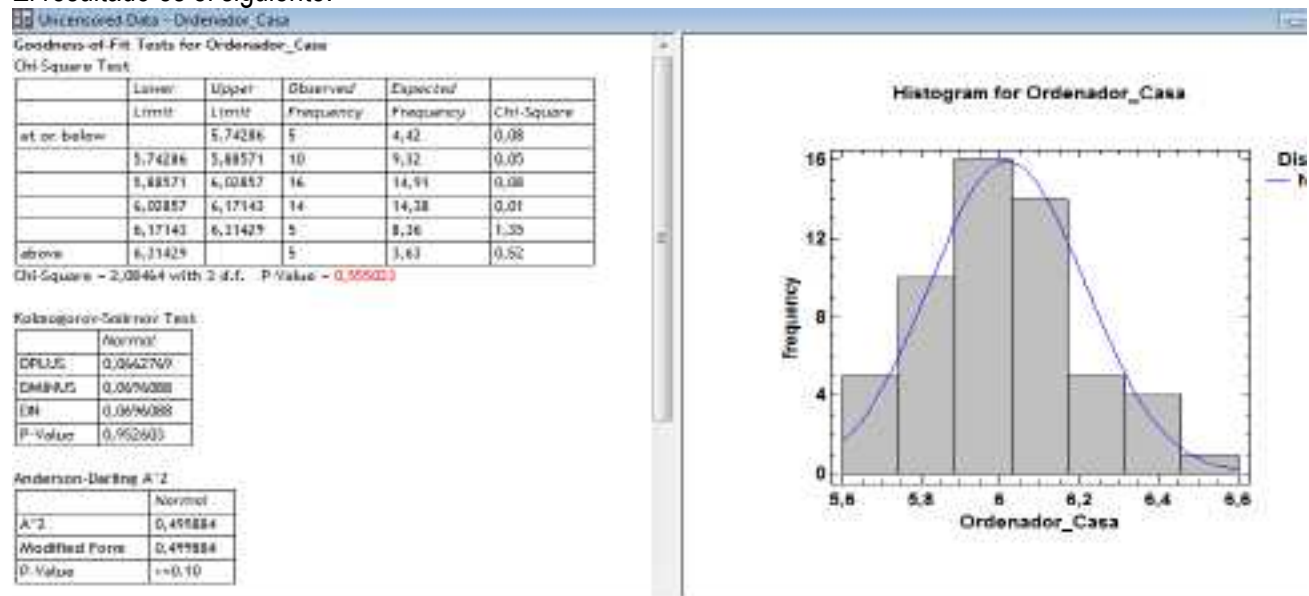
Introducimos en el campo DATA la variable Ordenador_Casa y le damos a OK. En la ventana que aparece a continuación seleccionamos la distribución normal.



Antes de hacer el test de la chi-cuadrado, lo más recomendable es visualizar el ajuste. En las opciones de Panel and Graphs seleccionamos el test de bondad de ajuste y el histograma. El histograma lo hacemos seleccionando 7 clases en la ventana de Pane Options (son 55 datos). El test de la chi-cuadrado lo hacemos sin usar clases equiprobables (ver gui n sobre ajuste de distribuciones).



El resultado es el siguiente:



Como vemos que el histograma con la curva superpuesta tiene un buen ajuste y dado que el p-valor obtenido es superior a 0.05 en los tres test de bondad de ajuste que se muestran, utilizando un nivel de significación habitual del 5% no podemos rechazar la hipótesis de que la variable proceda de una distribución normal.

Como podemos asumir normalidad en la variable Ordenador_Casa, podemos calcular los intervalos de confianza para la media y para la desviación típica. Si no pudiésemos asumir que la variable Ordenador_Casa fuese normal, no podríamos utilizar los intervalos de confianza de la desviación típica que proporciona el Statgraphics, al estar basados en la normalidad. Como la muestra es suficientemente grande ($n > 30$) todavía podríamos utilizar los intervalos de confianza de la media aunque la variable no fuese normal (¿por qué?). Por último, si la variable no fuese normal y la muestra fuese pequeña, no podríamos tampoco usar los intervalos de confianza para la media.

1.4. Intervalos de confianza

Para realizar los intervalos de confianza para μ y para σ seleccionamos: Describe / Numeric Data / One-Variable Analysis, y en Panel and Graphs seleccionamos 'Confidence Intervals'



Obtenemos la siguiente información:

Confidence Intervals for Ordenador_Casa
 95,0% confidence interval for mean: 6,01835 +/- 0,0530968 [5,96525; 6,07144]
 95,0% confidence interval for standard deviation: [0,165349; 0,241944]

Si calculamos el intervalo de confianza para μ aplicando la fórmula correspondiente obtenemos:

$$\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}} = 6,01835 \pm 1,9944 \cdot \frac{0,196409}{\sqrt{55}} = [5,96525; 6,07144]$$

que coincide con el valor del intervalo de confianza para la media proporcionado por el Statgraphics. Por tanto, el tiempo medio de acceso a la web será un valor que estará, con una confianza del 95% entre 5.96 y 6.07 segundos.

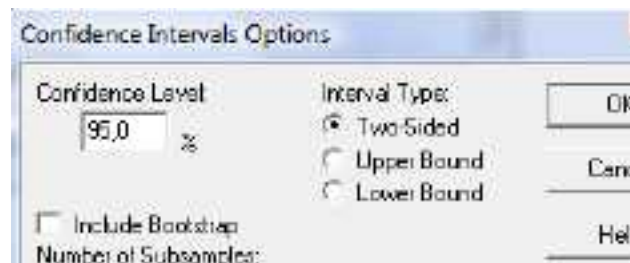
También aparece el intervalo de confianza para la desviación típica. A partir de él obtendremos, elevando al cuadrado cada término, el intervalo de confianza para la varianza [0,0273402 , 0,0585368]. Si calculamos el intervalo de confianza para σ^2 aplicando la fórmula correspondiente obtenemos

$$\frac{(n-1) \cdot \hat{s}^2}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1) \cdot \hat{s}^2}{\chi^2_{n-1, \frac{\alpha}{2}}} \Rightarrow \frac{54 \cdot 0.0385763}{\chi^2_{54, 0.975}} \leq \sigma^2 \leq \frac{54 \cdot 0.0385763}{\chi^2_{54, 0.025}}$$

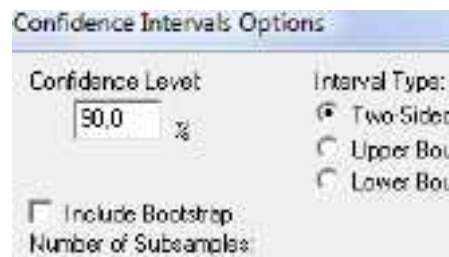
$$\Rightarrow \frac{2,0831202}{76,19555} \leq \sigma^2 \leq \frac{2,0831202}{35,586446} \Rightarrow 0,0273391 \leq \sigma^2 \leq 0,058536$$

que coincide con el intervalo de confianza proporcionado por el Statgraphics (elevándolo al cuadrado al tratarse de la desviación típica). Por tanto el tiempo de acceso a la web la podemos aproximar a una variable aleatoria normal $N(\mu, \sigma^2)$ donde, con un 95% de confianza la varianza será un valor que estará entre 0.0273 y 0.058.

Si queremos otro nivel de confianza diferente, nos situamos sobre la salida anterior, pulsamos el botón derecho del ratón y elegimos la opción PANE OPTIONS,



Vamos a calcular los intervalos de confianza del 90% para la media y la desviación típica



Se obtienen los siguientes resultados

Confidence Intervals for Ordenador_Casa

90,0% confidence interval for mean: 6,01835 +/- 0,0443223 [5,97402; 6,06267]

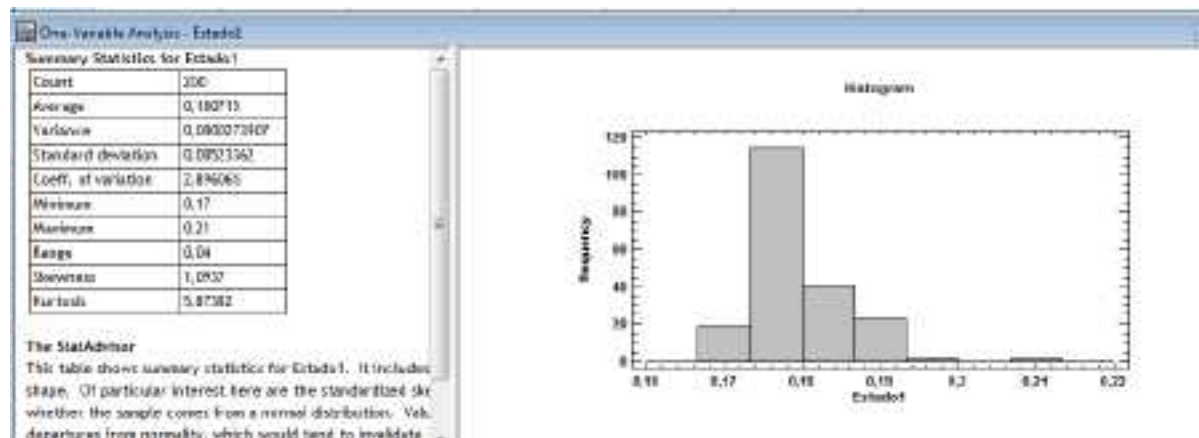
90,0% confidence interval for standard deviation: [0,169914; 0,233777]

2. Ejemplo 2: Tiempo de ejecución de un Bucle

Vamos a considerar ahora el fichero TiempoBucle que indica el tiempo en segundos de ejecución de un programa de Matlab bajo distintas circunstancias. En cada circunstancia, la ejecución se repite 200 veces. Se desean construir intervalos de confianza para la media y la varianza **poblacionales** del tiempo que tarda el programa en el Estado1

2.1 Análisis univariante de la variable

Vamos a comenzar analizando la variable tiempo de ejecución de un bucle en el Estado 1. El análisis gráfico y numérico de esta variable se hace en: Describe / Numeric Data / One-Variable Análisis. Seleccionamos las mismas opciones que en el ejemplo anterior. El histograma lo realizamos con 9 clases, pues un número superior muestra muchas irregularidades. Los resultados son los siguientes:



Podemos observar en el histograma que se trata de una distribución con una sola concentración en torno a 0,18 segundos y hay una asimetría positiva muy acusada. Además, la variable parece mucho más apuntada que una campana. No parece que esta variable sea muy normal. Entre los valores obtenidos se encuentran la media y varianza muestrales que son estimaciones puntuales de la media y la varianza poblacionales. Las estimaciones 'puntuales' son

$$\hat{\mu} = 0.1807$$

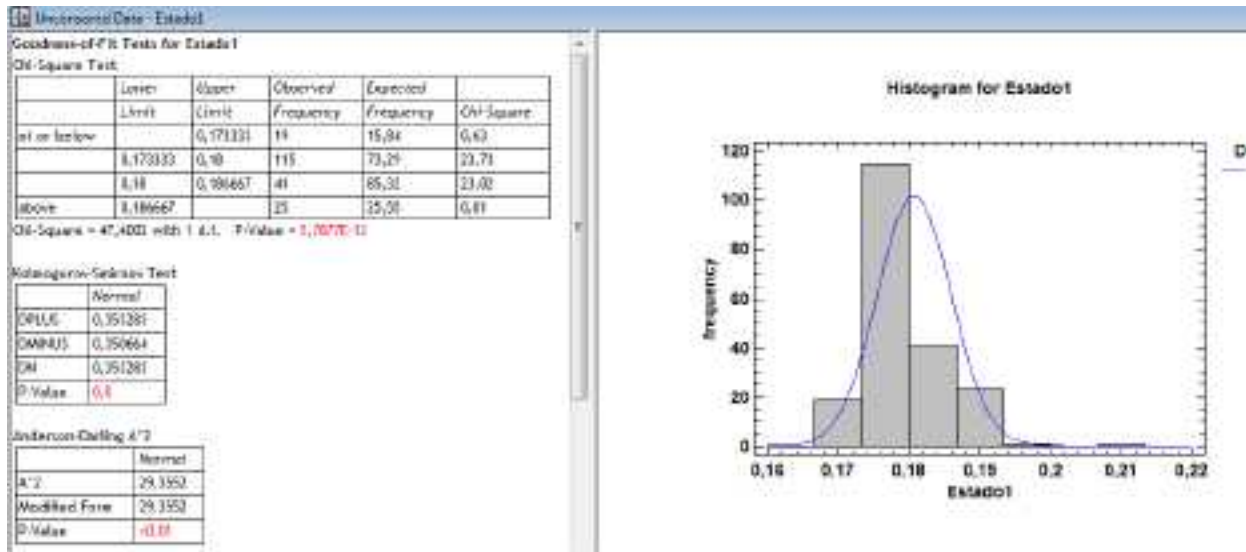
$$\hat{\sigma}^2 = 0.000274$$

Nuestro objetivo es estimar los parámetros utilizando intervalos de confianza.

2.2. Análisis de la normalidad de la variable

Para poder construir intervalos de confianza nos interesa saber si nuestra variable es o no normal (¿por qué?). Podemos observar que el histograma anterior no tiene similitud con la función de densidad de una distribución normal. Además, los coeficientes de asimetría (Skewness) y curtosis (Kurtosis) son muy elevados, indicando

que la variable tiene asimetría positiva y un apuntamiento más acusado que la normal. Nuestra variable parece que se aleja mucho de la normal. Para terminar el ejercicio de comparación de nuestros datos con la distribución normal, realizamos el test de la chi-cuadrado. Vamos a ver si los datos permiten tratar a nuestra población como una normal con el test de la Chi-cuadrado. Realizamos el test de la misma manera que en el ejemplo anterior (histograma con 9 clases y test sin clases equiprobables) y obtenemos el siguiente resultado:



Como vemos que el histograma con la curva superpuesta no tiene un buen ajuste y dado que el p-valor obtenido es inferior a 0.05, utilizando el nivel de significación habitual del 5% podemos rechazar que la variable tiempo que tarda el programa en el Estado1 proceda de una distribución normal.

Como estamos ante una muestra grande ($n=200$), tenemos que, por el teorema central del límite la media muestral, seguirá una distribución normal independientemente de cómo sea la distribución de la variable. Por lo tanto, aunque la variable Estado1 no sea normal, podemos calcular el intervalo de confianza para la media, **pero no para la desviación típica**.

2.3. Intervalos de confianza

Para obtener el intervalo de confianza para μ seleccionamos Describe / Numeric Data / One-Variable Analysis, y en Panel and Graphs marcamos la opción Confidence Intervals y obtenemos la siguiente información:

Confidence Intervals for Estado1

95,0% confidence interval for mean: 0,180715 +/- 0,000729768 [0,179985; 0,181445]

95,0% confidence interval for standard deviation: [0,00476603; 0,00580376]

Por tanto, el tiempo medio de ejecución de un programa de Matlab en el Estado 1 será un valor que estará, con una confianza del 95% entre 0.17998 y 0.18144 segundos. Al no tener una variable normal, el intervalo de la desviación típica no es fiable.