

PRÁCTICA 6. INFERENCIA SOBRE UNA POBLACIÓN NORMAL

Objetivo


El objeto de la presente sesión de práctica informática es complementar y afianzar los conceptos relativos a las técnicas de inferencia sobre una población normal vistos en clase (apartado 2, de la UD5). Asimismo se pretende que el alumno se familiarice con las opciones que el programa Statgraphics ofrece al respecto.

NOTA: Se recomienda que, durante el trabajo no presencial del alumno, los resultados obtenidos a partir del *Statgraphics* en esta práctica se calculen “a mano” y se cotejen con los mismos.

1. Comprobación de la hipótesis de normalidad

Antes de realizar cualquier estudio de inferencia sobre una población supuestamente normal debemos comprobar que NO hay indicios claros de que esta distribución NO sea la adecuada para representar nuestros datos. Para ello, entre otras herramientas, podemos generar un gráfico de **papel probabilístico normal**.

Hay dos maneras de obtener dicho gráfico en Statgraphics:

- Seleccionar la opción de menú **Plot > Exploratory Plots > Normal Probability Plot...** (Figura 1). En el cuadro de diálogo resultante, seleccionamos la variable cuyo gráfico queremos obtener y pulsamos “OK”.
- Seleccionar la opción **Describe > Numeric Data > One-variable Analysis...**. En el cuadro de diálogo, elegimos la variable cuyo papel probabilístico deseamos generar y pulsamos “OK”. En el panel de resultados, haciendo clic sobre el botón de **Graphical options** , podemos seleccionar “Normal Probability Plot” (Figura 2).

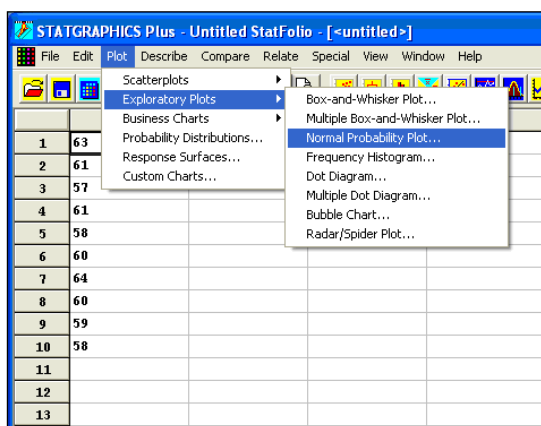


Figura 1. Opción de menú para seleccionar directamente un papel probabilístico normal.

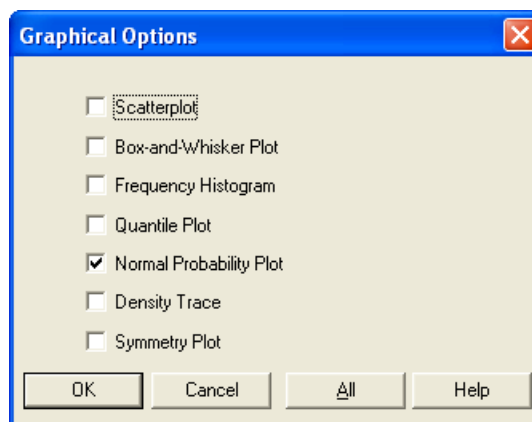


Figura 2. Selección del papel probabilístico normal, dentro de las opciones gráficas disponibles en el análisis de una variable.

En ambos casos obtenemos el gráfico del papel probabilístico normal o PPN (**Figura 3**).

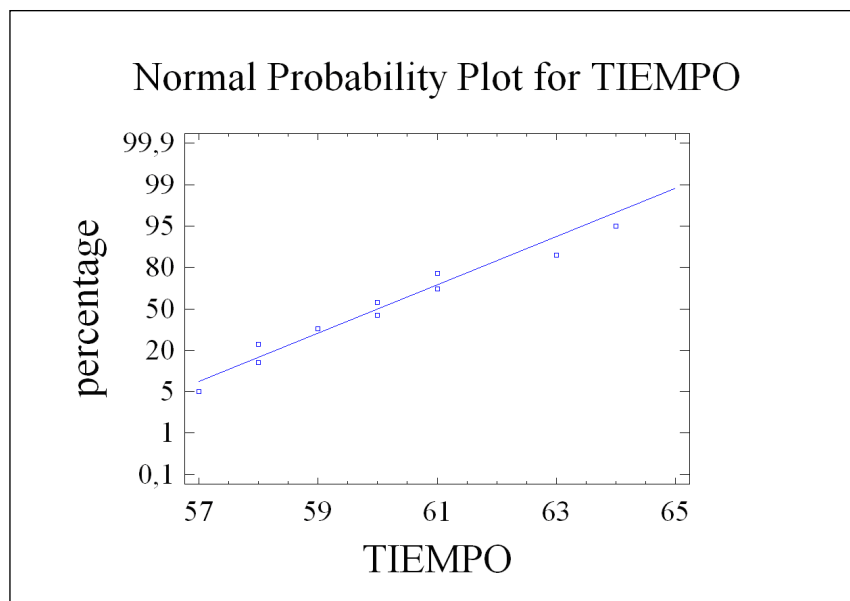


Figura 3. Ejemplo de un papel probabilístico normal obtenido con Statgraphics.

RECUERDA: La idea básica para manejar este gráfico es la siguiente: si los puntos (las observaciones) se encuentran, en general, próximos a la recta, se puede asumir (o “no se descarta”) que los datos provengan de una distribución normal.

Pregunta 1. Un programador desea estudiar las características de un nuevo sistema de depuración automática de programas. Para ello, selecciona al azar 10 programas y registra el tiempo que el nuevo sistema tarda en realizar la depuración de los mismos. Dichos tiempos, en minutos, son los siguientes: 63, 61, 57, 61, 58, 60, 64, 60, 59 y 58 (variable aleatoria TIEMPO). ¿Puede asumirse que los datos siguen una distribución normal?

RECUERDA. El PPN no deja de ser un gráfico de frecuencias relativas acumuladas. Algunas de sus aplicaciones ya se explicaron en la Unidad Didáctica 2

Pregunta 2. ¿Cuáles son los coeficientes estándar de asimetría y curtosis de los datos?, ¿Qué se puede decir sobre la distribución de la variable TIEMPO a partir de los valores obtenidos para estos parámetros?

RECUERDA. Para obtener los parámetros de posición, dispersión y forma (asimetría y curtosis) que caracterizan una muestra, debes entrar en **Describe > Numeric Data > One-variable Analysis...** y, en el panel de resultados, pulsar el botón de **Tabular options** y seleccionar **“Summary Statistics”**.

2. Contrastes de hipótesis


Una vez hemos comprobado la normalidad de las observaciones, lo siguiente que podemos hacer es plantearnos averiguar o *deducir cosas* de la variable en cuestión (en nuestro ejemplo, “tiempo de depuración”) a partir de la muestra que tenemos de ella; en otras palabras, vamos a *realizar inferencia sobre la población a partir de la muestra*.

Esto se puede hacer básicamente a través de dos vías: mediante *contrastes o tests de hipótesis*, o usando *intervalos de confianza*.

El contraste más habitual es preguntarse si la media “teórica” o poblacional de la variable (m) toma o no un determinado valor. Es lo que se conoce como un *contraste sobre la media* de la distribución de la variable que estamos considerando, y suele expresarse así:

$$\begin{cases} H_0: & m = m_0 \\ H_1: & m \neq m_0 \end{cases}$$

Es decir, tenemos un conjunto de observaciones, y queremos ver si la información que me proporciona dicha muestra *confirma o desmiente* que la media m de la variable de la cual provienen las observaciones es igual a un determinado valor m_0 , con una determinada probabilidad de equivocarnos pequeña y conocida de antemano.

Para realizar un contraste de hipótesis con Statgraphics, acudimos de nuevo a la opción de menú **Describe > Numeric Data > One-Variable Analysis...**, seleccionamos la variable que nos interesa analizar y pulsamos “OK”. Tras esto, en la ventana de resultados, pulsando el botón **Tabular options**  podemos habilitar el panel “**Hypothesis Test**” (contraste de hipótesis; **Figura 4**), cuyo aspecto inicial será, aproximadamente, el que se muestra en la **Figura 5**.

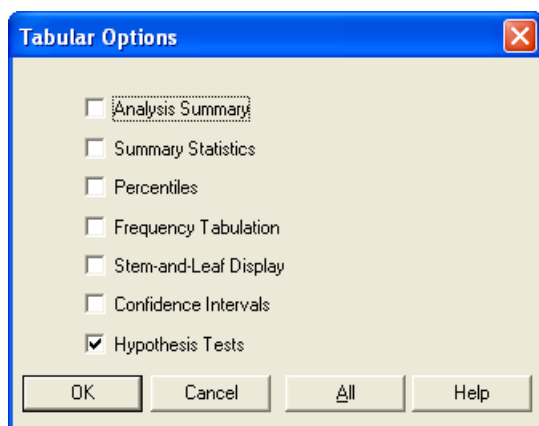


Figura 4. Selección de un contraste de hipótesis, dentro de las opciones de panel disponibles en el análisis de una variable.

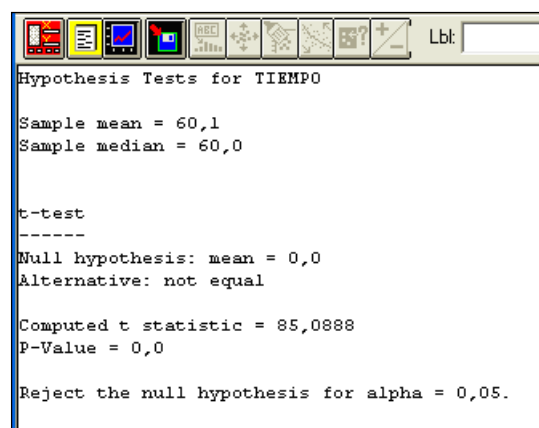


Figura 5. Detalle del aspecto inicial del panel para realizar contrastes de hipótesis sobre una población.

Para introducir los datos del contraste, hacemos clic con el botón derecho del ratón sobre el panel que acabamos de habilitar y seleccionamos “**Pane Options...**” (opciones del panel; **Figura 6**).

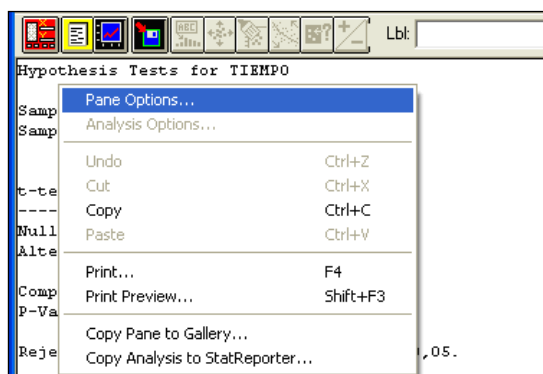


Figura 6. Selección de las opciones del panel para realizar contrastes de hipótesis sobre la media de una población.

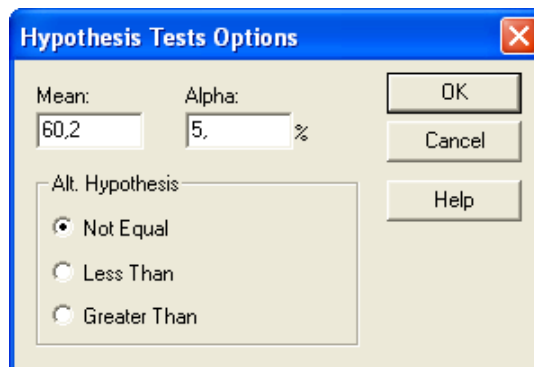


Figura 7. Definición de las opciones del contraste de hipótesis sobre la media de una población.

En el cuadro de diálogo que aparecerá (“**Hypothesis Test Options**”; **Figura 7**), debemos especificar la siguiente información:

- **Mean:** Valor teórico de la media de la población (m_0). Es el valor que queremos confirmar o desmentir a partir de la información que nos proporciona la muestra.
- **Alpha:** Representa el *riesgo de 1ª especie* o *nivel de significación*.

RECORDAR: el *riesgo de 1ª especie* representa la probabilidad que asumimos de equivocarnos, en el sentido de descartar m_0 como verdadero valor de m cuando en realidad sí lo es. Habitualmente se asignan a α valores como 10%, 5% o 1%.

- **Alt. Hypothesis:** Si, a partir de las observaciones, rechazamos la hipótesis de que m sea m_0 , entonces se considera como *hipótesis alternativa* la negación de la hipótesis inicial (o *hipótesis nula*), es decir, la opción “**Not Equal**” ($m \neq m_0$). Sin embargo, también podemos considerar la posibilidad de tomar como hipótesis alternativa sólo una de las desigualdades: “**Less Than**” ($m < m_0$) o “**Greater Than**” ($m > m_0$).

En esta práctica, como en clase, siempre consideraremos una hipótesis alternativa de tipo “**Not Equal**”¹.

Tras pulsar “**OK**”, Statgraphics resuelve el contraste, es decir, contesta a la pregunta “¿rechazamos o no rechazamos la hipótesis $m = m_0$?”.

Se ha visto en las sesiones de teoría y seminario cómo resolver de manera “manual” un contraste. Aquí, basta con saber interpretar correctamente la salida del Statgraphics.

En primer lugar, debemos fijarnos sólo en la información relativa al “**Test-t**”. En esta información se recogen los valores de m_0 y α introducidos previamente. También se muestra el valor del estadístico de contraste t o t calculada (“*Computed t statistic*”) y el p -valor del contraste (“**p-value**”).

Intuitivamente, podemos decir que el p -valor nos informa de cuán probable es que la media muestral haya sido la que ha sido, si suponemos cierta la hipótesis $m = m_0$. Por

¹ Este tipo de contrastes se denominan *contrastos bilaterales*

tanto, valores muy pequeños del *p-value* indican que sería “casi imposible” observar lo que hemos observado, si fuese cierto que la media poblacional m es m_0 .

RECUERDA. Fijado un valor de significación α , un contraste sobre la media de una población rechaza la hipótesis nula $m = m_0$ cuando el p-valor es menor que α (ver formulario y/o documentación de clase).

Pregunta 3. ¿Se puede admitir la hipótesis de que la distribución de la cual provienen las observaciones posee una media de 60,2 minutos, tomando un nivel de significación del 5%?

Pregunta 4. Tomando $\alpha = 1\%$, ¿podemos afirmar, a partir de las observaciones, que el tiempo medio que el sistema tarda en depurar cualquier programa es 62,4 minutos?

NOTA. Las dos cuestiones anteriores ilustran dos maneras distintas de preguntar por un contraste de hipótesis sobre la media de una población.

3. Intervalos de confianza


Otra manera de obtener conclusiones acerca la población a partir de la muestra es utilizar *intervalos de confianza*.

Si estamos interesados en un parámetro de la distribución que estamos estudiando (en nuestro caso, el parámetro sería m o σ), podemos decir, de manera muy intuitiva, que un intervalo de confianza para dicho parámetro no es más que *un rango de valores donde “casi seguro” se encuentra el valor del parámetro*.

Es, por tanto, una manera de dar una *estimación* del verdadero valor del parámetro.

La probabilidad de que, al construir el intervalo, *nos dejemos fuera* el verdadero valor del parámetro se denota, de nuevo, por la letra α . A la probabilidad $(1-\alpha)$ de que el parámetro a estimar se encuentre dentro del intervalo que vamos a construir se la llama *nivel de confianza*. Habitualmente, se consideran niveles de confianza del orden de 90%, 95% o 99%.

NOTA. El hecho de que exista un nivel de confianza, es decir, una probabilidad, NO significa que el parámetro a estimar sea aleatorio. Lo que es aleatorio es la muestra que utilizamos para construir el intervalo. Por eso, es posible que si repetimos el muestreo muchas veces, en un porcentaje de ellas (concretamente, $(\alpha \times 100)\%$) el verdadero valor del parámetro quede fuera del intervalo que generemos.

Para obtener un intervalo de confianza en Statgraphics, nos mantendremos en la ventana de resultados del análisis de una variable (**Describe > Numeric Data > One-Variable Analysis...**) pero ahora, en **Tabular options** , seleccionaremos **“Confidence Intervals”** (intervalos de confianza; **Figura 8**), lo cual hará que se nos muestre un panel similar al que se presenta en la **Figura 9**.

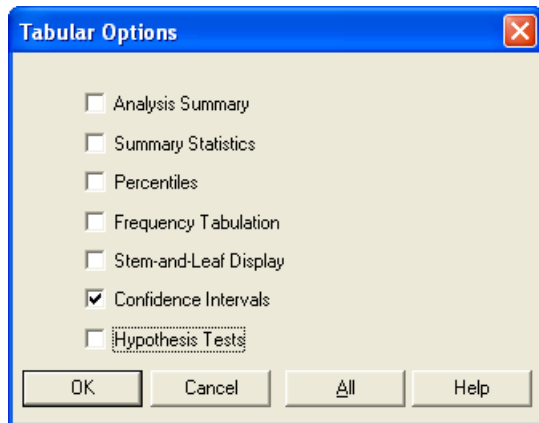


Figura 8. Selección de un intervalo de confianza, dentro de las opciones de panel disponibles en el análisis de una variable.

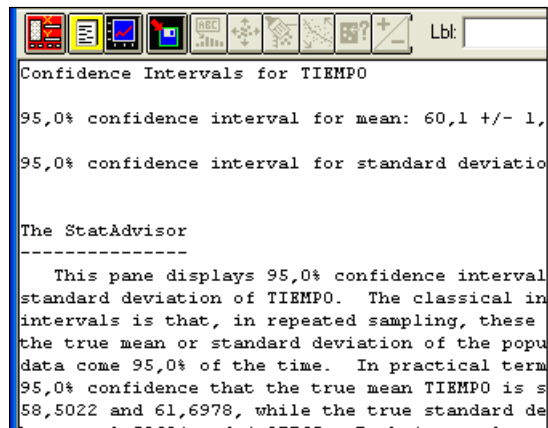


Figura 9. Detalle del aspecto inicial del panel para construir intervalos de confianza para parámetros de una población normal.

Haciendo clic con el botón derecho sobre el panel de intervalos de confianza y seleccionado “**Pane Options...**”, podemos cambiar el nivel de confianza (**Confidence Level**) de los intervalos generados.

Pregunta 5. Construye un intervalo de confianza para el tiempo medio de depuración de los programas con un riesgo de 1ª especie de 5%. ¿Es admisible la hipótesis de que la media teórica de la población (μ) de la cual provienen las observaciones sea 62 minutos?

Pregunta 6. Tomando el mismo nivel de confianza de la pregunta anterior, genera un intervalo de confianza para la desviación típica del tiempo de depuración. ¿Es admisible la hipótesis de que la desviación típica teórica de la población (σ) de la cual provienen las observaciones sea 2 minutos?

RECUERDA. Obviamente, si el verdadero valor de la desviación típica σ se encuentra en el intervalo $[a, b]$ con probabilidad 95%, entonces el verdadero valor de la varianza σ^2 se encuentra con la misma probabilidad dentro del intervalo $[a^2, b^2]$.

Pregunta 7. Tomando un nivel de significación $\alpha=0,01$, genera un intervalo de confianza para la varianza del tiempo de depuración.

Respuestas a las preguntas propuestas

Pregunta 1

A la vista del papel probabilístico normal de los datos, sí que podemos suponer que los datos provienen de una distribución normal.

Pregunta 2

Coeficiente estándar de asimetría: $0,57 \in [-2,2]$

Coeficiente estándar de curtosis: $-0,33 \in [-2,2]$

Se puede decir que TIEMPO sigue una distribución simétrica y de "apuntamiento" normal (mesocúrtica), lo que confirma la afirmación hecha en la pregunta anterior.

Pregunta 3

```
t-test
-----
Null hypothesis: mean = 60,2    (H0)
Alternative: not equal    (H1)

Computed t statistic = -0,141579    (t calculada)

P-Value = 0,890531    (p-valor)

Do not reject the null hypothesis for alpha = 0,05.    (Aceptamos H0)
```

Como nos dice el *Statgraphics*, NO podemos rechazar la hipótesis inicial ($p\text{-valor} \geq \alpha$)

La distribución de la cual provienen las observaciones de TIEMPO puede poseer una media de 60,2 minutos, tomando un nivel de significación del 5%

Pregunta 4

Como nos dice el *Statgraphics*, rechazamos la hipótesis inicial (ya que $p\text{-valor} < \alpha$)

Pregunta 5

El intervalo obtenido para la media es IC_m : [58,5022 ; 61,6978].

Como $62 \notin [58,5022 ; 61,6978] \rightarrow$ Rechazamos la H_0 .

No se puede admitir la hipótesis de que la media teórica de la población de la cual provienen las observaciones sea 62 minutos, con un nivel de confianza del 95%.

Pregunta 6

El intervalo obtenido para la desviación típica es IC_σ : [1,53634;4,07765].

Como $2 \in [1,53634;4,07765] \rightarrow$ Aceptamos la H_0 .

Sí se puede admitir la hipótesis de que la desviación típica teórica de la población de la cual provienen las observaciones sea 2 minutos, con un nivel de confianza del 95%.

Pregunta 7

El intervalo obtenido para la varianza es IC_{σ^2} : $[(1,37964)^2; (5,08724)^2] \rightarrow [1,90; 25,88]$.

Fuentes

<http://www.statgraphics.net/>



Esta obra está bajo una licencia Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-sa/2.5/es/>