

PRÀCTICA 6. INFERÈNCIA SOBRE UNA POBLACIÓ NORMAL

Objectiu


L'objecte de la present sessió de pràctica informàtica és complementar i afermar els conceptes relatius a les tècniques d'inferència sobre una població normal vistos en classe (apartat 2, de la UD5). Així mateix es pretén que l'alumne es familiaritze amb les opcions que el programa Statgraphics ofereix sobre aquest tema.

NOTA: Es recomana que, durant el treball no presencial de l'alumne, els resultats obtinguts a partir de l'Statgraphics en aquesta pràctica es calculen amb les fórmules.

1. Comprovació de la hipòtesi de normalitat

Abans de realitzar qualsevol estudi d'inferència sobre una població suposadament normal hem de comprovar que NO hi ha indicis clars que aquesta distribució NO siga l'adequada per a representar les nostres dades. Per a açò, entre altres eines, podem generar un gràfic de paper probabilístic normal. Hi ha dues maneres d'obtenir aquest gràfic en Statgraphics:

Selecioneu l'opció de menú **Plot > Exploratory Plots > Normal Probability Plot...** (Figura 1). En el quadre de diàleg resultant, seleccionem la variable el gràfic de la qual volem obtenir i premem “OK”.

• Seleccioneu l'opció **Describe > Numeric Data > One-Variable Analysis...** En el quadre de diàleg, triem la variable de la qual desitgem generar el paper probabilístic i premem “OK”. En el panell de resultats, fent clic sobre el botó de **Graphical Options** , podem seleccionar “Normal Probability Plot” (Figura 2).

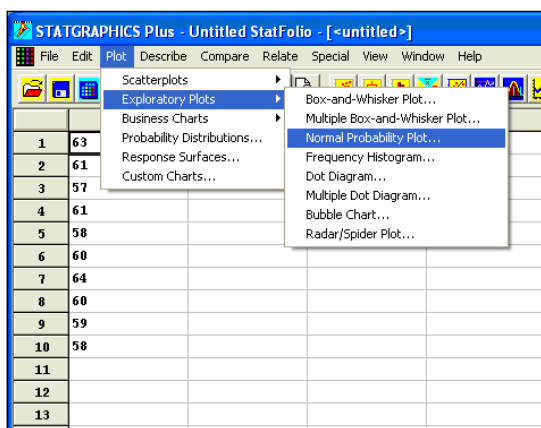


Figura 1. Opció de menú per a seleccionar directament un paper probabilístic normal.

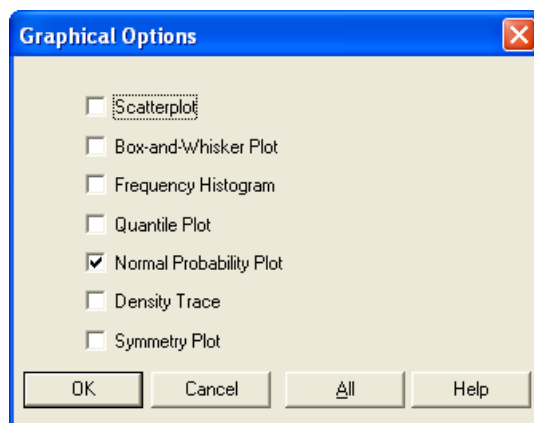


Figura 2. Selecció del paper probabilístic normal, dins de les opcions gràfiques disponibles en l'anàlisi d'una variable.

En tots dos casos obtenim el gràfic del paper probabilístic normal o PPN (Figura 3).

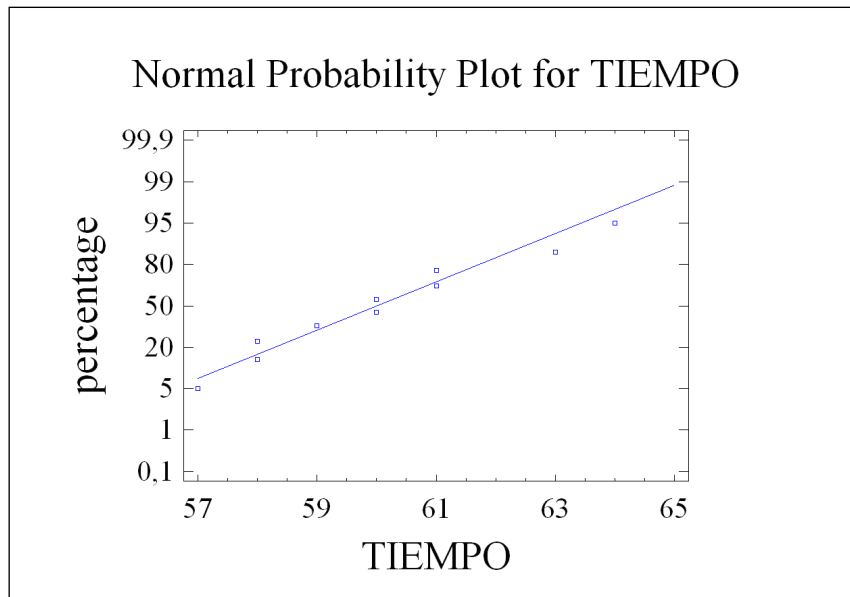


Figura 3. Exemple d'un paper probabilístic normal obtingut amb Statgraphics.

RECORDEU: La idea bàsica per a manejar aquest gràfic és la següent: si els punts (les observacions) es troben, en general, pròxims a la recta, es pot assumir (o “no es descarta”) que les dades són d'una distribució normal.

Pregunta 1. Un programador vol estudiar les característiques d'un nou sistema de depuració automàtica de programes. Per a fer-ho, selecciona a l'atzar 10 programes i registra el temps que el nou sistema tarda en realitzar la depuració dels mateixos. Aquests temps, en minuts, són els següents: 63, 61, 57, 61, 58, 60, 64, 60, 59 i 58 (variable aleatòria TEMPS). Es pot assumir que les dades segueixen una distribució Normal?

RECORDEU. El PPN no deixa de ser un gràfic de freqüències relatives acumulades. Algunes de les seues aplicacions ja es van explicar en la Unitat Didàctica 2

Pregunta 2. Quins són els coeficients estàndard d'asimetria i curtosi de les dades?, Què es pot dir sobre la distribució de la variable TEMPS a partir dels valors obtinguts per a aquests paràmetres?

RECORDEU. Per a obtenir els paràmetres de posició, dispersió i forma (asimetria i curtosi) que caracteritzen una mostra, heu d'entrar en **Describe >Numeric Data >One-Variable Analysis...** i, en el panell de resultats, feu clic la icona de **Tabular Options** i seleccioneu “**Summary Statistics**”.

2. Contrastos d'hipòtesi

Una vegada hem comprovat la normalitat de les observacions, el següent que podem fer és plantejar-nos esbrinar o deduir coses de la variable en qüestió (en el nostre exemple, “temps de depuració”) a partir de la mostra que tenim; en altres paraules, anem a realitzar inferència sobre la població a partir de la mostra.

Açò es pot fer bàsicament a través de dues vies: mitjançant contrastos o tests d'hipòtesi, o usant intervals de confiança.

El contrast més habitual és preguntar-se si la mitjana “teòrica” o poblacional de la variable (μ) presa o no un determinat valor. És el que es coneix com un contrast sobre la mitjana de la distribució de la variable que estem considerant, i sol expressar-se així:

És a dir, tenim un conjunt d'observacions, i volem veure si la informació que em proporciona aquesta mostra

confirma o desmenteix que la mitjana m de la variable de la qual provenen les observacions és igual a un determinat valor m_0 , amb una determinada probabilitat d'equivocar-nos xicoteta i coneguda per endavant.

Per a realitzar un contrast d'hipòtesi amb Statgraphics, acudim de nou a l'opció de menú **Describe > Numeric Data > One-Variable Analysis...**, es selecciona la variable que ens interessa analitzar i premem “ **OK**”. Després d'açò, en la finestra de resultats, prement el botó **Tabular Options** podem habilitar el panell “**Hypothesis Test**” (contrast d'hipòtesi; Figura 4), l'aspecte inicial del qual serà, aproximadament, el que es mostra en la Figura 5.

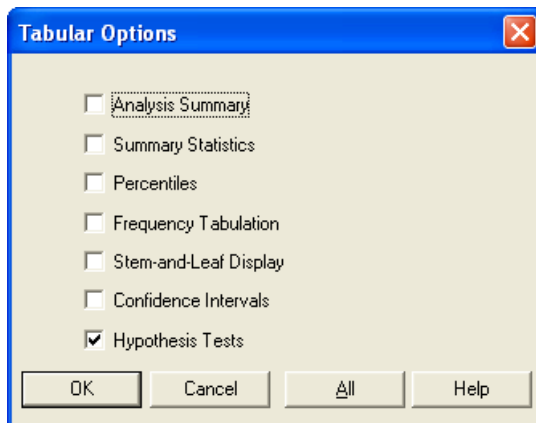


Figura 4. Selecció d'un contrast d'hipòtesi, dins de les opcions de panell disponibles en l'anàlisi d'una variable.

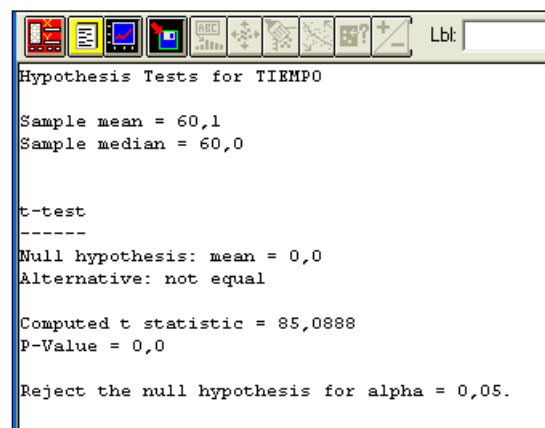


Figura 5. Detall de l'aspecte inicial del panell per a realitzar contrastos d'hipòtesis sobre una població.

Per a introduir les dades del contrast, fem clic amb el botó dret del ratolí sobre el panell que acabem d'habilitar i seleccionem “ **Pane Options...**” (opcions del panell; Figura 6).

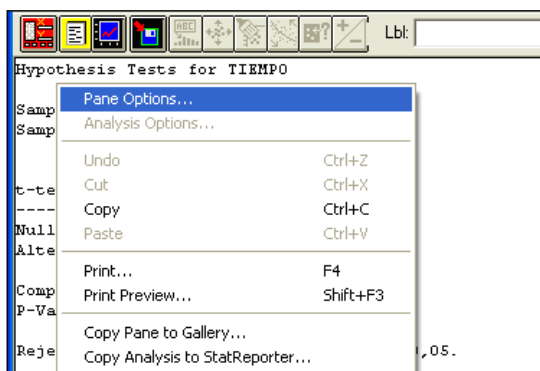


Figura 6. Selecció de les opcions del panell per a realitzar contrastos d'hipòtesis sobre la mitjana d'una població.

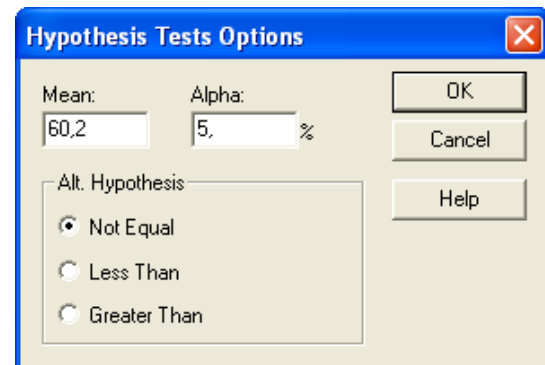


Figura 7. Definició de les opcions del contrast d'hipòtesi sobre la mitjana d'una població.

En el quadre de diàleg que apareixerà (“ **Hypothesis Test Options**”; Figura 7), hem d'especificar la següent informació:

- Mean: Valor teòric de la mitjana de la població (m_0). És el valor que volem confirmar o desmentir a partir de la informació que ens proporciona la mostra.

- Alpha: Representa el risc de 1ª espècie.

RECORDEU: el risc de 1ª espècie representa la probabilitat que assumim d'equivocar-nos, en el sentit de descartar m_0 com a vertader valor de m quan en realitat sí ho és. Habitualment s'assignen a α valors com 10%, 5% o 1%.

- Alt. Hypothesis: Si, a partir de les observacions, rebutgem la hipòtesi que m siga m_0 , llavors es considera com a hipòtesi alternativa la negació de la hipòtesi inicial (o hipòtesi nul·la), és a dir, l'opció “ Not Equal” ($m \neq m_0$). No obstant açò, també podem considerar la possibilitat de prendre com a hipòtesi alternativa només una de les desigualtats: “Less Than” ($m < m_0$) o “Greater Than” ($m > m_0$). En aquesta pràctica, com en classe, sempre considerarem una hipòtesi alternativa de tipus “ Not Equal”.

Després de prémer “ **OK**”, Statgraphics resol el contrast, és a dir, contesta a la pregunta “rebutgem o no rebutgem la hipòtesi $m = m_0$?”

S'ha vist en les sessions de teoria i seminari com resoldre de manera “manual” amb les fórmules un contrast. Ací, n'hi ha prou amb saber interpretar correctament l'eixida del Statgraphics.

En primer lloc, hem de fixar-nos només en la informació relativa a el “Test-t”. En aquesta informació s'arrepleguen els valors de m_0 i α introduïts prèviament. També es mostra el valor de l'estadístic de contrast t o t calculada (“**Computed t statistic**”) i el p-valor del contrast (“**p- value**”).

Intuïtivament, podem dir que el p-valor ens informa de quant probable és que la mitjana mostral haja sigut la que ha sigut, si suposem certa la hipòtesi $m = m_0$. Per tant, valors molt xicotets del p- value indiquen que seria “quasi impossible” observar el que hem observat, si fóra cert que la mitjana poblacional m és m_0 .

RECORDEU. Fixat un valor de significació α , un contrast sobre la mitjana d'una població rebutja la hipòtesi nul·la $m = m_0$ quan el p-valor és menor que α (veure formulari i/o documentació de classe).

Pregunta 3. Es pot admetre la hipòtesi que la distribució de la qual provenen les observacions posseeix una mitjana de 60,2 minuts, prenent un nivell de significació del 5%?

Pregunta 4. Prenent $\alpha = 1\%$, podem afirmar, a partir de les observacions, que el temps mitjà que el sistema tarda en depurar qualsevol programa és 62,4 minuts?

NOTA. Les dues qüestions anteriors il·lustren dues maneres diferents de preguntar per un contrast d'hipòtesi sobre la mitjana d'una població.

3. Intervals de confiança

Una altra manera d'obtenir conclusions de la població a partir de la mostra és utilitzar intervals de confiança.

Si estem interessats en un paràmetre de la distribució que estem estudiant (en el nostre cas, el paràmetre seria m o σ), podem dir, de manera molt intuïtiva, que un interval de confiança per a aquest paràmetre no és més que un rang de valors on “quasi segur” es troba el valor del paràmetre.

És, per tant, una manera de donar una estimació del vertader valor del paràmetre.

La probabilitat que, en construir l'interval, ens deixem fóra el vertader valor del paràmetre es denota, de nou, per la lletra α . A la probabilitat $(1-\alpha)$ que el paràmetre a estimar es trobe dins de l'interval que anem a construir li la crida nivell de confiança. Habitualment, es consideren nivells de confiança de l'ordre de 90%, 95% o 99%.

NOTA. El fet que haja un nivell de confiança, és a dir, una probabilitat, NO significa que el paràmetre a estimar siga aleatori. El que és aleatori és la mostra que utilitzem per a construir l'interval. Per açò, és possible que si repetim el mostreig moltes vegades, en un percentatge d'elles (concretament, $(\alpha \times 100)\%$) el verdader valor del paràmetre és fóra de l'interval que generem.

Per a obtenir un interval de confiança en Statgraphics, ens mantindrem en la finestra de resultats de l'anàlisi d'una variable (**Describe > Numeric Data > One-Variable Analysis...**) però ara, a **Tabular Options**, seleccionarem “**Confidence Intervals**” (intervals de confiança; Figura 8), la qual cosa farà que se'ns mostre un panell similar al que es presenta en la Figura 9.

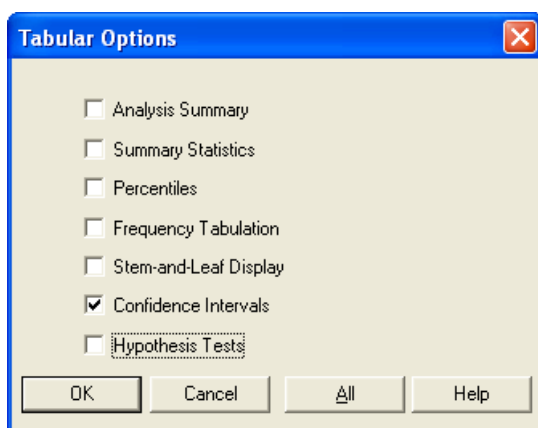


Figura 8. Selecció d'un interval de confiança, dins de les opcions de panell disponibles en l'anàlisi d'una variable.

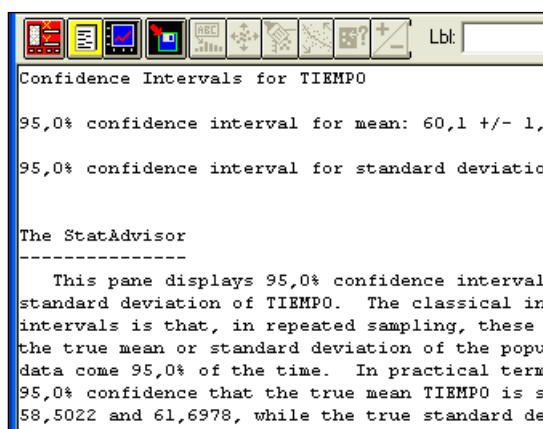


Figura 9. Detall de l'aspecte inicial del panell per a construir intervals de confiança per a paràmetres d'una població normal.

Fent clic amb el botó dret sobre el panell d'intervals de confiança i seleccionat “**Pane Options...**”, podem canviar el nivell de confiança (**Confidence Level**) dels intervals generats.

Pregunta 5. Construïu un interval de confiança per al temps mitjà de depuració dels programes amb un risc de 1^a espècie de 5%. És admissible la hipòtesi que la mitjana teòrica de la població (μ) de la qual provenen les observacions siga 62 minuts?

Pregunta 6. Prenent el mateix nivell de confiança de la pregunta anterior, genera un interval de confiança per a la desviació típica del temps de depuració. És admissible la hipòtesi que la desviació típica teòrica de la població (σ) de la qual provenen les observacions siga 2 minuts?

RECORDEU. Òbviament, si el verdader valor de la desviació típica σ es troba en l'interval $[a, b]$ amb probabilitat 95%, llavors el verdader valor de la variància σ^2 es troba amb la mateixa probabilitat dins de l'interval $[a^2, b^2]$.

Pregunta 7. Prenent un nivell de significació $\alpha=0,01$, genereu un interval de confiança per a la variància del temps de depuració.

Respostes a les preguntes proposades

Pregunta 1

A la vista del paper probabilístic normal de les dades, sí que podem suposar que les dades provenen d'una distribució normal.

Pregunta 2

Coefficient estàndard d'asimetria: $0,57 \in [-2,2]$

Coefficient estàndard de curtosi: $-0,33 \in [-2,2]$

Es pot dir que TEMPS segueix una distribució simètrica i de "apuntament" normal (mesocúrtica), la qual cosa confirma l'afirmació feta 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. (Acceptem H0)
```

Com ens diu l'Statgraphics, NO podem rebutjar la hipòtesi inicial ($p\text{-valor} \geq \alpha$) La distribució de la qual provenen les observacions de TEMPS pot tenir una mitjana de 60,2 minuts, prenent un risc de primera espècie del 5%

Pregunta 4

Com ens diu el Statgraphics, rebutgem la hipòtesi inicial (ja que $p\text{-valor} < \alpha$)

Pregunta 5

L'interval obtingut per a la mitjana és IC m: $[58,5022 ; 61,6978]$.

Com $62 \notin [58,5022 ; 61,6978] \diamond$ Rebutgem la H_0 .

No es pot admetre la hipòtesi que la mitjana teòrica de la població de la qual provenen les observacions siga 62 minuts, amb un nivell de confiança del 95%.

Pregunta 6

L'interval obtingut per a la desviació típica és IC σ : [1,53634;4,07765].

Com $2 \in [1,53634;4,07765] \diamond$ Acceptem la H_0 .

Sí es pot admetre la hipòtesi que la desviació típica teòrica de la població de la qual provenen les observacions siga 2 minuts, amb un nivell de confiança del 95%.

Pregunta 7

L'interval obtingut per a la variància és IC σ^2 : $[(1,37964)^2; (5,08724)^2] = [1,90; 25,88]$.