

NOM: _____ COGNOM: _____
(Contesteu cada pregunta en el seu lloc. Expliqueu i justifiqueu els càlculs. Totes les preguntes valen igual)

Primer Problema (B4)


Primera part

Suposem que volem portar a terme un test d'hipòtesis bilateral per a comparar una esperança amb un valor concret. Explica els conceptes següents i **il·lustra** la teva explicació amb un gràfic.

1) Definiu *punt-crític*, i expliqueu per a què serveix. Calculeu el punt crític associat a una distribució t-Student amb 10 graus de llibertat, amb un $\alpha=0.01$.

En un test bilateral, el punt crític associat a un nivell de significació α és el valor de l'eix de les x's que deixa una probabilitat acumulada igual a $1-\alpha/2$.

Els punts crítics severixen per a decidir si es rebutja o no la hipòtesi nul·la del test. Concretament, si $t_{\{n, 1-\alpha/2\}}$ és el punt crític, es rebutjarà H_0 quan el valor de l'estadístic corresponent sigui, en valor absolut, superior al punt crític.

El punt crític associat a una t-d'Student amb 10 graus de llibertat i un $\alpha=0.01$ és **3.161**. 

2) Definiu *p-valor* i expliqueu per a què serveix. Calculeu el p-valor que correspon a un valor de 3.169 en una distribució t-Student amb 10 graus de llibertat.

El p-valor associat a un nombre real x_0 , i corresponent a un test bilateral, és igual a dues vegades la probabilitat acumulada en x_0 si x_0 és negatiu, i dues vegades la unitat menys la probabilitat acumulada en x_0 , si x_0 és positiu.

El p-valor serveix també per a decidir si es rebutja o no la hipòtesi nul·la. A un nivell de significació α , es rebutjarà H_0 quan el p-valor sigui inferior a α .

El p-valor corresponent a 3.169 en una distribució t-d'Student amb 10 graus de llibertat és igual a 0.01.

Segona part

La diferència entre el temps de processament de dues implementacions diferents d'un algorisme es va mesurar en 7 workloads diferents. Les diferències de temps obtingudes varen ser de: 1.5, 2.6, -1.8, 1.3, -0.5, 1.7 i 2.4.

1) Explica com creus que s'han obtingut aquestes dades. Creus que provenen de dues mostres independents o aparellades? Perquè?

Es tracta de les diferències dels valors obtinguts en dues mostres aparellades. Això és degut a que, per a cada workload, s'ha mesurat el temps de processament de cadascuna de les implementacions. Per tant, els temps venen aparellats per mitjà del workload.

2) Quina és la distribució de probabilitat de la v.a. Y corresponent a la diferència de temps de processament de les dues implementacions? Com comprovaries que és l'acertada?

La distribució natural d'una variable corresponent al temps d'execució d'un programa és la Normal.

Com que la diferència de variables aleatòries amb distribució Normal té també distribució Normal, el correcte és assumir que la diferència dels temps d'execució de les dues implementacions diferents, és la Normal.

Per a comprobar que les dades són normalment distribuïdes, dibuixariem l'histograma i el qqnorm. Respecte a l'histograma, mirariem que fos simètric i que s'assemblés a la campana de Gauss. Respecte al qqnorm, intentariem veure que els punts cauen sobre la recta.

3) El nostre objectiu és portar a terme un test d'hipòtesis per a testar si les diferències de processament de les dues implementacions són o no estadísticament significatives.

Quin és el test d'hipòtesis escaient en aquest cas? Especifiqueu-ne les hipòtesis.

Si anomenem μ al valor esperat de la variable Y, el que volem veure és si el seu valor esperat és o no estadísticament diferent de zero.

El test és el que té com a hipòtesis:

$H_0: \mu=0$

$H_1: \mu \neq 0$

4) Quin és el valor de l'estadístic pel nostre conjunt de dades?

Grandària mostral $n=7$

mitjana aritmètica $\bar{x}=7.2/7=1.03$

Variància Mostral $S^2=(22.84-7.2*7.2/7)/6=2.57$

desviació estàndard mostral $s=\sqrt{2.57}=1.6$

Estadístic $t=\bar{x}/(S/\sqrt{7})=1.07/(1.6/\sqrt{7})=1.07/0.6047=1.77$



5) Conclueu si la hipòtesi nul.la pot o no ser rebutjada. Primer feu-ho per mitjà del punt crític, i després per mitjà del p-valor (calculeu un p-valor aproximat amb les taules).

Prenent $\alpha=0.05$,

El punt crític associat a una probabilitat de 0.975 en una distribució t-d'Student amb 6 g.ll és $t_{\{6, 0.975\}}=2.447$. Atès que $|1.77| < 2.447$. No es pot rebutjar H_0 .

Mirant les taules, veiem que la probabilitat acumulada en 1.7 és aproximadament igual a 0.93. Així doncs, el p-valor per un test bilateral en aquest punt és igual a

p- valor: $2(1-0.93)=0.14$ aprox. Atès que $0.14 > 0.05$, No podem rebutjar H_0 .

Podem afirmar que no hi ha diferències estadísticament significatives entre les dues implementacions del algorisme. Es pot fer servir indistintament qualsevol de les dues implementacions.

6) Calculeu un interval de confiança al 95% per l'esperança de la diferència dels temps. Us porta a la mateixa conclusió que heu tret abans respecte al test d'hipòtesis?

IC és igual a:

$$\bar{x} \pm t_{\{0.975, 6\}} * S/\sqrt{n} = 1.03 \pm 2.447 * 0.6047 = 1.03 \pm 1.4797$$

per tant, l'interval és el (-0.4497, 2.509)

Atès que l'interval conté el valor zero, no podem rebutjar la hipòtesi nul.la.

7) Calculeu un interval de confiança per al valor esperat de Y menys 1 ($\mu-1$).

L'IC per a $\mu-1$, s'obté de restar-li una unitat al IC per a μ . Així doncs, aquest és igual a:

(-1.4497, 1.509).

8) A partir de l'interval de confiança del punt anterior, compareu si el valor esperat de Y és o no diferent de la unitat.

El test que té per hipòtesis:

$H_0: \mu=1$

$H_1: \mu \neq 1$

és equivalent a:

$H^*_0: \mu-1=0$

$H^*_1: \mu-1 \neq 0$

Com que l'IC per a $\mu-1$ conté el zero, No podem rebutjar la hipòtesi nul.la H_0^* , el que equival a no poder rebutjar H_0 .

També es veu que no es pot rebutjar H_0 perquè l'IC per a μ trobat en l'apartat 7) conté la unitat.

B5. Efecte d'una actualització del hardware a l'empresa informàtica

PCaccount.com

El director de l'empesa PCaccount.com està interessat en demostrar l'efectivitat d'una important actualització del hardware existent en aquesta companyia. Per fer-ho es pren una mostra de 61 observacions abans de l'actualització i 61 més després de l'actualització. Basat en aquestes dades, el temps mitjà de funcionament és de $\bar{y}_1 = 8.5$ minuts abans de l'actualització i $\bar{y}_2 = 7.2$ minuts després de l'actualització. A més, es sap que $s_1^2 = 3.24$ i $s_2^2 = 2.25$.

- 1) Quina distribució de probabilitat segueix el quocient de variàncies. Raoneu la vostra resposta (1.5 punts)

Les dades originals venen d'una distribució normal, i el quocient entre les variàncies mostrals (numerador) i poblacionals (denominador), segueixen una distribució χ_{n-1}^2 , i χ_{m-1}^2 , és a dir

$$\frac{nS_1^2}{\sigma_1^2} \sim \chi_{n-1}^2 \quad \text{i} \quad \frac{mS_2^2}{\sigma_2^2} \sim \chi_{m-1}^2$$

El quocient d'aquestes dues χ^2 segueix una distribució F amb $n - 1$ graus de llibertat en el numerador i $m - 1$ en el denominador, és a dir,

$$\frac{\frac{nS_1^2}{\sigma_1^2}}{\frac{mS_2^2}{\sigma_2^2}} \Rightarrow F_{(n-1)(m-1)}$$

o de forma equivalent,

$$\frac{nS_1^2\sigma_2^2(m-1)}{mS_2^2\sigma_1^2(n-1)}$$

Aquest quocient segueix una distribució $F_{n-1,m-1}$

En aquest cas, els graus de llibertat son 60 en el numerador i 60 en el denominador

- 2) Plantegeu i resoleu la prova d'hipòtesi d'igualtat de variàncies amb una confiança del 95%, sent l'alternativa $\sigma_1^2 > \sigma_2^2$ (1.5 punts)

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

$$F_{1,2} = \frac{S_1^2}{S_2^2} = \frac{3.24}{2.25} = 1.44$$

El punt crític, amb una probabilitat acumulada del 95% és 1.53

Com que $F_{1,2} = 1.44 < 1.53$, no podem rebutjar la hipòtesi nul.la H_0 i per tant, $\sigma_1^2 = \sigma_2^2$

- 3) Obtindríeu el mateix resultat de la pregunta 2 si l'alternativa a la prova d'hipòtesi d'igualtat de variàncies for $\sigma_1^2 \neq \sigma_2^2$? Trebal·leu amb una confiança del 95%, (1.5 punts)

Ara el punt crític és 1.67 (mirar les taules de la F) , que està relacionat amb la hipòtesi alternativa $\sigma_1^2 \neq \sigma_2^2$

Com que $\hat{F} = 1.44 < 1.67$, en aquest cas tampoc podem rebutjar la hipòtesi H_0 d'igualtat de variàncies

- 4) Si assumim que les variàncies poblacionals son iguals ($\sigma_1^2 = \sigma_2^2$), calculeu la variància "pooled" mostral així como la desviació estàndard "pooled", mostral (1.5 punts)

$$S_{pooled}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{60 * 3.24 + 60 * 2.25}{120} = 2.745$$

$$S_{pooled} = \sqrt{2.745} = 1.66$$

- 5) Pel que fa a la comparació de mitjanes, hem obtingut els següents valors per les mitjanes mostrals $\bar{y}_1 = 8.5$ i $\bar{y}_2 = 7.2$ i els valors de les variàncies mostrals son: $S_1^2 = 2$ i $S_2^2 = 4$. Calculeu un interval de confiança al 95% per a la diferencia de mitjanes. (1.5 punts)

IC(95%) per la diferencia de mitjanes

$$(\bar{y}_1 - \bar{y}_2) \pm 1.96 * S_{pooled} * \sqrt{\frac{1}{61} + \frac{1}{61}} = (8.5 - 7.2) \pm 1.96 * 1.66 * 0.18 = 1.3 \pm 0.56 = (0.74, 1.86)$$

- 6) Ara, el director de l'empresa PCaccount.com se n'ha adonat que potser treballar amb dues mostres independents no és la millor opció i pensa que és millor recollir un únic conjunt de dades i aplicar dos algorismes diferents per veure si realment un és més efectiu que l'altre. Quines avantatges/inconvenients representa treballar amb dades aparellades? Com es calcula, en aquest cas, la variància de la mostra? Raoneu la vostra resposta (1 punt)

Treballar amb dades aparellades, en general, disminueix la variabilitat entre les dues mostres.

La variable diferència es calcula a partir de $d_i = y_{1i} - y_{2i}$, per $i=1, \dots, n$, on y_{1i} son els temps de l'algorisme "1" i y_{2i} els de l'algorisme "2". Ambdós algorismes treballen amb el mateix conjunt de dades originals. És per aixó que aquest procediment s'anomena "treballar amb dades aparellades".

La variància mostral que s'emprarà ara és la variància de la diferència de respostes d_i .

- 7) Finalment, aquesta companyia decideix treballar amb dos conjunts de dades i està interessada en poder estimar la proporció de ventes de cadascun d'aquests algorismes. Pel que fa al primer algorisme la proporció és 0.8 i pel segon algorisme és 0.5. El nombre d'observacions utilitzada pel primer algorisme és 50 i pel segon és 30. Es pot assumir que aquests dos algorismes tenen un rendiment similar? Raoneu la vostra resposta (1.5 punts)

$$p_1 = 0.8 \quad p_2 = 0.5$$

$$n_1 = 50 \quad n_2 = 30$$

$$H_0: \pi_1 = \pi_2$$

$$H_1: \pi_1 > \pi_2$$

$$p = \frac{n_1 * p_1 + n_2 * p_2}{n_1 + n_2} = \frac{50 * 0.8 + 30 * 0.5}{80} = \frac{40 + 15}{80} = 0.6875$$

Com que n_1 i n_2 son prou grans, es pot utilitzar l'estadístic z.

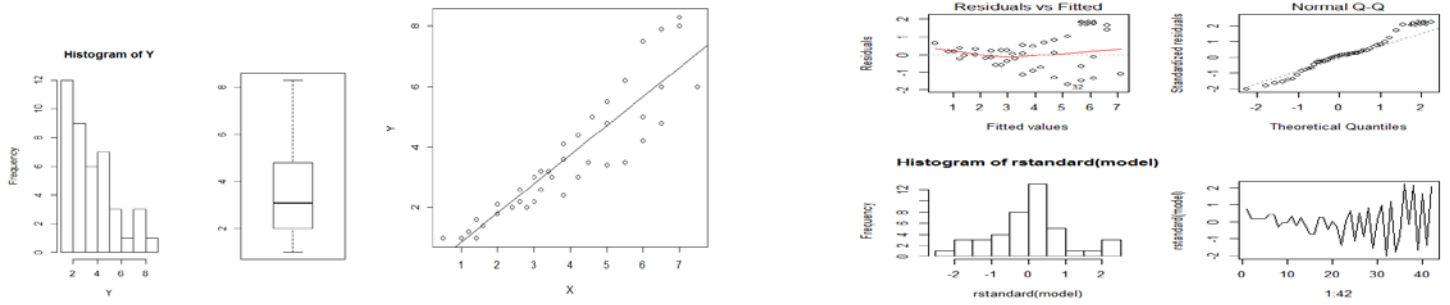
En aquest cas

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{0.8 - 0.5}{\sqrt{\frac{0.6875 * 0.3125}{50} + \frac{0.6875 * 0.3125}{30}}} = \frac{0.3}{\sqrt{0.004 + 0.007}} = \frac{0.3}{0.11} = 2.727$$

Com que $P(Z \geq 2.727) = 0.0032$, es rebutja H_0 i per tant $\pi_1 > \pi_2$

Problema 3 (B6)

Volem estudiar la relació lineal entre les notes de dos parcials d'una assignatura. Es recullen les notes del primer parcial (X) i del segon parcial (Y) de 42 estudiants, i obtenim els següents resultats:



Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.10195	0.28491	-0.358	0.722
X	0.96198	0.06718	14.320	<2e-16

Residual standard error: 0.8467 on 40 degrees of freedom
 Multiple R-squared: 0.8368, Adjusted R-squared: 0.8327

1.- Indiqueu l'equació de la recta de regressió estimada i interpreteu-ne els coeficients

$$Y = -0.10195 + 0.96198 X$$

L'ordenada a l'origen (valor de Y quan X val 0) és molt propera al 0 (-0.1) i la pendent és positiva i propera a 1 (0.96) indicant una gran equivalència de la nota del segon parcial respecte la del primer

2.- Calculeu un interval de confiança al 95% pels paràmetres β_0 i β_1

$$b_0 \pm t_{40,0.975} S_{b_0} = -0.10195 \pm 2.021 * 0.28491 = [-0.678, 0.474]$$

$$b_1 \pm t_{40,0.975} S_{b_1} = 0.96198 \pm 2.021 * 0.06718 = [0.826, 1.098]$$

3.- Comenteu què diuen els intervals anteriors respecte a la interpretació dels coeficients

IC de β_0 : amb una confiança del 95% es pot afirmar que β_0 està entre -0.678 i 0.474. Com que 0 hi pertany indica que no hi ha evidència per rebutjar que la recta passa per l'origen (una nota de 0 del primer parcial correspon també a 0 al segon parcial)

IC de β_1 : amb una confiança del 95% es pot afirmar que β_1 està entre 0.826 i 1.098. Cada unitat que augmenta X implica que Y augmenta entre 0.826 i 1.098 (entorn del valor de 1 indicant una equivalència de 1 a 1 entre la nota del primer parcial i el segon)

4.- Calculeu el coeficient de determinació i el coeficient de correlació i interpretant-los

$$R^2 = 0.8368 \quad \text{bona part (> 80\%) de la variabilitat de Y queda explicada per X}$$

$$r = 0.91 \quad \text{relació lineal positiva i força intensa (proper a 1)}$$

5.- Enuncieu les premisses o hipòtesis de la regressió lineal i comenteu si es compleixen o no per aquest cas concret. Especifiqueu de quins resultats i/o gràfics es dedueixen els vostres comentaris.

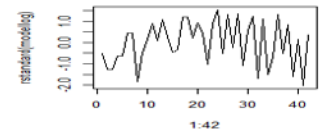
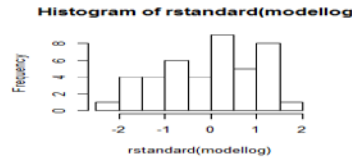
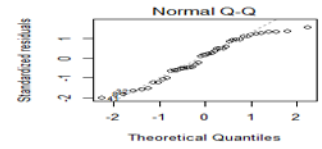
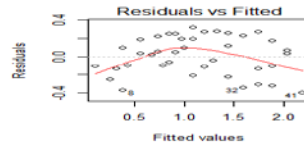
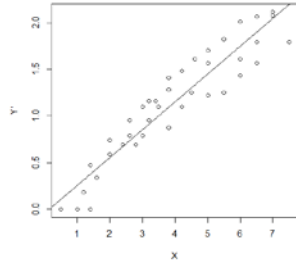
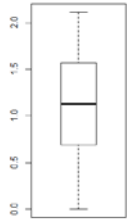
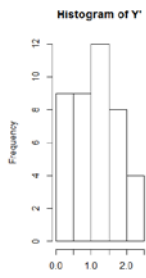
Linealitat: ok, valors força ajustats a la recta en el plot entre X i Y

Independència: hi ha certa dependència ja que a l'augmentar X augmenta el valor dels residus

Homocedasticitat: no es compleix ja que els gràfics dels residus presenten menys variabilitat a l'inici i va augmentant. També s'observa al plot entre X i Y una primera zona amb els punts més ajustats a la recta i després es van separant més.

Normalitat: l'ajust a una normal és força correcte però falla una mica com indicaria la part final del NormalQ-Q no ajustant-se a una recta

Apliquem una transformació logarítmica a les notes del segon parcial (Y) obtenint unes noves dades (Y') pels 42 estudiants i n'estudiem la relació lineal amb X:



$$\sum x_i = 158.3$$

$$\sum y'_i = 45.427$$

$$\sum x_i^2 = 755.49$$

$$\sum (y'_i)^2 = 65.056$$

$$\sum x_i y'_i = 218.617$$

6.- Calculeu la covariància i la correlació entre X i Y'

$$\text{Mean}(x) = 158.3/42 = 3.77$$

$$s_x^2 = (755.49 - (158.3)^2/42) / 41 = 3.87$$

$$s_x = 1.97$$

$$\text{Mean}(y') = 45.427/42 = 1.08$$

$$s_y^2 = (65.056 - (45.427)^2/42) / 41 = 0.39$$

$$s_y = 0.62$$

$$S_{XY'} = (218.617 - 158.3 * 45.427/42) / 41 = 1.16$$

$$r = 1.16 / (1.97 * 0.62) = 0.95$$

$$(R^2 = 0.90)$$

$$s^2 = (41 * 0.39 * 0.10) / 40 = 0.04$$

$$s = 0.20$$

7.- Calculeu la recta de regressió i interpreteu-ne els coeficients

$$b_1 = 1.16/3.87 = 0.30$$

$$b_0 = 1.08 - 0.3 * 3.77 = -0.05$$

$$Y' = -0.05 + 0.30 X$$

L'ordenada a l'origen continua sent molt propera al 0 (-0.05) i la pendent també és positiva però inferior (0.30)

8.- Poseu a prova si la recta de regressió es pot considerar plana amb un risc del 5%

$$H_0: \beta_1 = 0 \quad (H_1: \beta_1 <> 0)$$

$$t = (b_1 - 0) / S_{b_1} = (0.30 - 0) / 0.016 = 18.7$$

$$(S_{b_1} = \text{sqrt}(0.04 / (41 * 3.87))) = 0.016$$

amb risc del 5% l'estadístic (18.7) està més enllà dels punts crítics (-2.021 i 2.021 ja que $t_{40,0.975} = 2.021$) per tant hi ha evidència per rebutjar que β_1 sigui 0 és a dir recta plana.

9.- Calculeu una predicció puntual pel cas de X=5 i calculeu un interval de confiança d'aquesta predicció

$$Y' = -0.05 + 0.3 * 5 = 1.45$$

$$1.45 \pm t_{40,0.975} 0.20 \text{sqrt}(1 + 1/42 + (5 - 3.77)^2 / (1.97^2 * 41)) = 1.45 \pm 2.021 * 0.2 = [1.04, 1.86]$$

10.- Compareu els dos models en quant al coeficient de determinació, la variabilitat residual i les premisses. Quin model creieu que és més encertat?

$$(\text{en aquest segon cas: } R^2 = (r)^2 = 0.90 \quad s^2 = (41 * 0.39 * 0.10) / 40 = 0.04 \quad s = 0.20)$$

Coefficient de determinació semblants (superior a 80%): en ambdós casos bona part de la variabilitat de la resposta queda explicada per X

La variabilitat residual en el primer cas (0.8467) és molt més gran que en el segon (0.2)

La premissa d'homocedasticitat en el primer cas no es compleix i en segon si (a més de més normalitat també en segon cas i menys dependència)

Per això el segon model és més encertat