NOM:_____COGNOM:_____

(*Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs*)

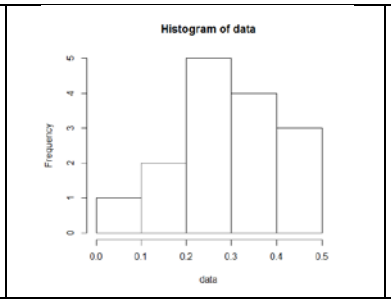
Problema 1 (B4)

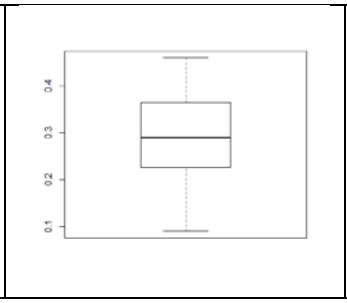
Un alumne de la FIB està estudiant com detectar els accessos fraudulents a comptes protegits per contrasenya. El primer mètode que està provant consisteix en detectar els accessos fraudulents a partir de mesurar el temps entre les pulsacions de les tecles a l'hora d'introduir l'usuari i la contrasenya.

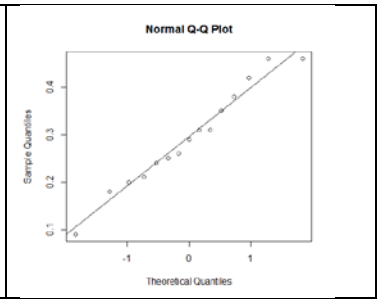
En un cas sospitós ha recollit els temps en segons entre pulsacions de les tecles i ha obtingut les següents dades:

$$\sum_{i=1}^{15} x_i = 4.41$$

$$\sum_{i=1}^{15} x_i^2 = 1.4551$$







1. Calculeu les estimacions puntuals del temps mitjà i de la desviació tipus (1 punt)

   $\bar{x}$= 0.294 s
   s_X= 0.106422 s

2. A partir dels gràfics argumenteu si podem suposar que el temps entre pulsacions segueix una distribució normal. (1 punt)

   A partir de l'histograma, del boxplot i del qqnorm podem suposar que el temps entre pulsacions segueix una distribució normal.

   Per estudis previs, se sap que una persona real utilitza 0.3653 segons de temps mitjà entre pulsacions per introduir les claus d'accés.

3. Per estudiar si el cas sospitós és un possible accés fraudulent es vol contrastar si el temps mitjà entre pulsacions és 0.3653 s (mitjana per una persona real) o no amb un risc de l'1%.
   Indiqueu:
   a) Les hipòtesis, les premisses, la fórmula de l'estadístic i quina és la distribució d'aquest sota la hipòtesi nul·la (1 punt)

   $$\begin{cases} H_0: \mu_X = 0.3653 \\ H_1: \mu_X \neq 0.3653 \end{cases}$$

   La premissa és X~ Normal

   L'estadístic és $\hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$ i $\hat{t}$~ t_{14}

   b) Calculeu el valor de l'estadístic (0.5 punts)

   $$\hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} = -2.594791$$

   c) Representeu gràficament els punts crítics, les zones d'acceptació i de rebuig i el valor de l'estadístic (1 punt)

   Els punts crítics són -2.977 i 2.'977

   Hem de buscar a les taules P(T < k)= 0.995 amb T~t_{14}. Trobem que k=2.977

   La zona d'acceptació és [-2.977 , 2.977]
   La zona de rebuig és (-∞,-2.977) ∪ (2.977, +∞)

   Gràficament.

d) A partir de l'estudi i del càlculs realitzats, interpreteu els resultats de la prova d'hipòtesi aplicada sobre el cas sospitós de ser fraudulent. (0.5 punts)

El valor de l'estadístic -2.594791 pertany a la zona d'acceptació, per tant podem concloure que amb un risc de l'1% no tenim evidències significatives que el cas sospitós sigui un cas d'accés fraudulent al compte de l'usuari.

4. a) Calculeu un interval bilateral amb 90% de confiança per la mitjana de temps entre pulsacions en l'usuari fraudulent (2 punts)

   Tenim que X~ Normal i que n=15 i α = 0.1, per tant,

   $(\bar{x} - t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}, \bar{x} + t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}})$ =

   En el nostre cas tenim que $t_{n-1,1-\frac{\alpha}{2}} = t_{14,0.95} = 1.761$

   $= (0.294 - 1.761 \cdot \sqrt{\frac{0.1066^2}{15}}, 0.294 + 1.761 \cdot \sqrt{\frac{0.1066^2}{15}})$ = (0.2456, 0.3424)

   b) A partir de l'interval de confiança obtingut, interpreteu els resultats sobre el cas sospitós de ser fraudulent. (0.5 punts)

   El valor de 0.3653 s no pertany a l'interval amb 90% de confiança de la mitjana poblacional del nostre cas sospitós (0.2456, 0.3424) i per tant, amb un 90% de confiança concloem que tenim evidències que ha estat una entrada fraudulenta.

5. Compareu els resultats dels apartats 3 i 4 (0.5 punts)

   Amb un risc de l'1% i amb els resultats de l'apartat 3 hem conclòs que no tenim evidències que l'accés ha estat fraudulent.

   Amb una 90% de confiança i amb els resultats de l'apartat 4 hem conclòs que tenim evidències que ha estat una entrada fraudulenta.

   En aquest cas tenim resultats diferents depenent del nivell de risc/confiança que establim en el nostre estudi

6. Un segon mètode que volen explorar és detectar els accessos fraudulents a partir del temps total en introduir l'usuari i la contrasenya. Experimentalment l'alumne de la FIB ha establert que els accessos fraudulents triguen més de 5s en accedir. Dels darrers 80 accessos, 20 han trigat més de 5s en accedir.
   a) Calculeu un IC 95% per a aquesta proporció d'accessos que triguen més de 5s en accedir (1.5 punts)

   Emprant estimació: $IC(\pi, 95\%) = p \pm z_{0.975} \cdot \sqrt{p \cdot \frac{1-p}{n}} = 0.25 \pm 1.96 \cdot \sqrt{0.25 \cdot \frac{0.75}{80}} =$ [0.1551, 0.3449]

   (També es pot calcular amb máxima indeterminació: $IC(\pi, 95\%) = p \pm z_{0.975} \cdot \sqrt{\frac{0.5^2}{n}} = 0.25 \pm 1.96 \cdot \sqrt{\frac{0.5^2}{80}} =$ [0.1404, 0.3596]

   b) Interpreteu el resultat obtingut en l'apartat anterior (0.5 punts)

   Amb una confiança del 95% la proporció d'accessos que triguen més de 5s en introduir l'usuari i la contrasenya està entre 16% i 34% (o 14% i 36%)

Al laboratori de càlcul on treballem es monitoritza la temperatura dels servidors allotjats. S'està estudiant com es comporta la temperatura a 10 dels ordinadors seleccionats a l'atzar, concretament el canvi després de dotze hores de funcionament. S'ha observat una variació de 46.3 a 52.2 graus en la temperatura mitjana; les diferències entre temperatura inicial i final presenten una desviació tipus de 9 graus.

Per altra banda, sabem que 4 dels 10 servidors escollits disposen d'un sistema de refrigeració instal·lat recentment: en aquests l'increment mitjà de la temperatura ha estat de 3.2 graus.

1. Determina a partir de la informació disponible en quant ha augmentat en mitjana la temperatura als altres sis servidors.

D = $T_{final} - T_{inicial}$. Mitjana global (m(D)): 5.9 = 52.2 – 46.3. Noteu que són dades de la mostra. El canvi al grup de 4 servidors amb nou sistema té mitjana: $m(D_N) = 3.2$. Volem saber la mitjana del canvi a l'altre grup: $m(D_V) = x$

$5.9 = (4\times3.2 + 6\times x)/10$; $x = (59 - 4\times3.2)/6 = 7.7$ graus

2. Considerant la mostra dels 10 servidors, creieu que es pot dir que després de dotze hores ha augmentat la temperatura mitjana dels servidors? Justifiqueu la resposta i estimeu amb interval de confiança del 95% l'increment esperat de la temperatura.

És un disseny aparellat i hem observat la temperatura a l'inici i al cap de 12h en cada servidor (i d'aquests valors hem tret les diferències). Tenim m(D)=5.9 (mitjana de la diferència de temperatura), i s=9 (desviació tipus), dades mostrals per una variable D observada a n=10 servidors. Plantegem una prova de hipòtesis:

H_0: E(D) = 0
H_1: E(D) > 0 (ha augmentat?)

Sota H_0, l'estadístic $t = \frac{m(D)-E(D)}{s/\sqrt{n}}$ segueix una t de Student amb 9 graus de llibertat. Si $t > t_{9,\,0.95} = 1.833$ (taules) rebutjarem la hipòtesi nul·la. En aquest cas, t=2.073, i es pot dir que la temperatura mitjana augmenta amb el temps, amb el risc d'error del 5%.

L'interval (bilateral) de confiança és: $m(D) \pm t_{9,0.975}\, s/\sqrt{n}$ = (–0.54, 12.34)

►*Cal evitar utilitzar l'interval de confiança per respondre la qüestió de si la temperatura mitjana augmenta, i fer una prova d'hipòtesis.*

3. La desviació tipus de l'increment de temperatura al grup de 4 servidors ha estat de 5 graus, i de 7.5 graus a l'altre grup. Justifiqueu formalment si hi ha motius per pensar que els servidors amb el sistema de refrigeració més recent tenen menor variabilitat en el canvi de temperatura que els altres servidors. Aclariu les premisses necessàries per la prova que utilitzeu.

Farem una prova d'hipòtesis per veure si les variàncies poblacionals de les dues variables (D_N: sistema recent; D_V: sistema previ) són iguals o no:

H_0: $V(D_N) = V(D_V)$
H_1: $V(D_N) < V(D_V)$ (menor al nou?)

Sota H_0, l'estadístic $F = \frac{s_V^2}{s_N^2}$ segueix una distribució F amb 5 i 3 graus de llibertat. Les premisses són: dues mostres independents, i distribució Normal de la variable "canvi". Si F és major de 9.01 (taules del quantil 0.95, per unilateral) veurem evidències de que una variància és major que l'altra. No és el cas en aquesta mostra, perquè F val 2.25 ($7.5^2/5^2$).

4. Es vol trobar resposta també a la qüestió de si el nou sistema de refrigeració és més efectiu que l'antic. És a dir: l'increment de temperatura és igual en ambdós sistemes, o un permet moderar millor la temperatura? Independentment del que heu trobat a la pregunta anterior, assumiu que els dos sistemes comparteixen una variància comuna en la variable "increment de temperatura".

La prova de hipòtesis es farà per demostrar, si es pot, que les mitjanes poblacionals de D_N y D_V no són iguals.

H_0: $E(D_N) = E(D_V)$
H_1: $E(D_N) \neq E(D_V)$ (si es vol demostrar inferioritat, hem de posar <, i el valor crític per rebutjar serà –1.860)

Sota H_0, l'estadístic $t = \frac{m(DN) - m(DV)}{s\sqrt{1/4 + 1/6}}$ segueix una distribució t de Student amb 8 graus de llibertat. Com les mostres són petites, la variable "canvi" hauria de ser Normal. Es rebutjarà H_0 si $|t| > 2.306$.

L'estimació de la variància comuna s^2 val $\frac{3\,5^2 + 5\,7.5^2}{8}$ = 44.53 (s=6.673). La diferència de mitjanes val 3.2–7.7 = –4.5 graus. L'error tipus de la diferència de mitjanes és 4.308, i *t* val –1.045. No es pot rebutjar la igualtat de mitjanes.

5. Expliqueu els resultats trobats als apartats anteriors, segons l'evidència que proporciona la mostra disponible.

Hem observat indicis clars que la temperatura als servidors és major al cap de 12 hores que a l'inici (interval de confiança al 95% per el valor esperat del canvi de temperatura: –0.54, 12.34).

No hem vist cap evidència que el nou sistema de refrigeració controli més l'augment de temperatura que l'anterior sistema: les observacions són compatibles amb un mateix valor esperat de l'increment de temperatura en els dos sistemes. Tampoc hem pogut demostrar que la variància sigui diferent en el nou sistema.

Donada la variabilitat de la variable temperatura, o de la variable diferència entre temperatura inicial i final, seria convenient disposar d'una mostra considerablement més gran.

6. El fabricant dels sistemes de refrigeració disposa de molta més informació dels seus clients, i manté una base de dades dels resultats de les proves efectuades, encara que només pot saber si l'increment de temperatura ha estat superior o inferior a 8 graus, després de 12 hores d'observació. En resum, aquesta és la informació recollida:

	Increment de T inferior a 8º	Increment de T superior a 8º
Sistema anterior	270	115
Sistema nou	95	28

Podem dir amb les dades del fabricant si hi ha cap diferència entre els sistemes? Feu la prova basada en l'estadístic chi-quadrat. Justifiqueu i interpreteu la conclusió.

Marginals	Increment de T inferior a 8º	Increment de T superior a 8º	
Sistema anterior	270	115	385
Sistema nou	95	28	123
	365	143	508

Sumem files i columnes per trobar els marginals.

Valor esperat	Increment de T inferior a 8º	Increment de T superior a 8º	
Sistema anterior	276.624	108.376	385
Sistema nou	88.376	34.624	123
	365	143	508

El valor esperat a cada cel·la és el valor proporcional als marginals corresponents, com si l'increment fos independent del sistema, la hipòtesi nul·la que s'assumeix. Per exemple: 385 365 / 508 = 276.624.

Sota la hipòtesi d'independència entre sistema i increment (major/menor de 8º), l'estadístic X^2 segueix una distribució chi-quadrat amb 1 grau de llibertat.

X^2 = (270-276.624)^2/276.624+(115-108.376)^2/108.376+(95-88.376)^2/88.376+(28-34.624)^2/34.624 = 2.327

Veiem que el punt crític per rebutjar és 3.84, per tant no podem descartar que els dos sistemes funcionin de forma equivalent, almenys respecte a un increment de 8 graus. No hem pogut demostrar que el sistema nou aconsegueixi moderar millor l'increment de temperatura als servidors.

NOM:_____COGNOM:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

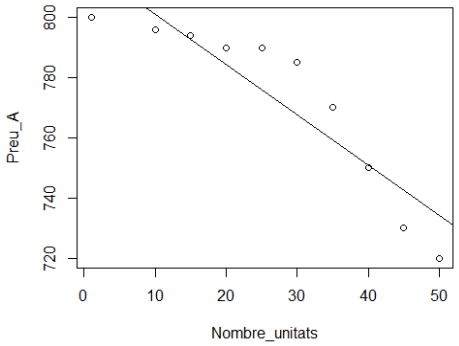
Problema 3 (B6)

Un empresa té dos proveïdors principals (A i B) de PCs, amb unes rebaixes del **preu per unitat** segons el nombre de PCs que es compren. Es decideix analitzar el preu per unitat que va aplicar cada proveïdor en 10 comandes de l'any passat a cadascun i que van ser de 1, 10, 15, 20, 25, 30, 35, 40, 45, i 50 unitats.

Un equip de l'empresa analitza les dades del proveïdor A i presenta els següents resultats:

Nombre d'unitats comprades en 10 comandes del darrer any: Nombre_unitats <- c(1, 10, 15, 20, 25, 30, 35, 40, 45, 50**)**

Preu unitari aplicat en cadascuna de les 10 comandes indicades: Preu_A <- c(800,796,794,790,790,785,770,750,730,720)



	Mitjana	Desviació tipus	Min	Max
Preu_A	772.5	29.09	720	800
Nombre_unitats	27.1	15.84	1	50

cov(Preu_A,Nombre_unitats) = $S_{\text{Preu\_A,Nombre\_unitats}}$ = -419.72

- Calculeu els coeficients de la recta de regressió del Preu_A en funció del Nombre d'unitats. Indiqueu i interpreteu l'equació de la recta ajustada. Calculeu i interpreteu el coeficient de determinació i la desviació residual (2 punts)

b_1 = -419.72 / 15.84^2 = -1.67

b_0 = 772.5 – (-1.67) 27.1 = 817.76

Preu_A = 817.76 – 1.67 Nombre_unitats

Des de 816.1 eur que costaria un PC, per cada unitat més que es compri el preu de cadascun baixa en 1.67 eur

Corr(G,ltemps) = -419.72 / (29.09 15.84) = -0.91 R^2 = sqr(-0.91) = 0.83 (coeficient de determinació)

R^2 : el 83% de la variabilitat del preu unitari és explicable pel nombre d'unitats comprades

S^2 = (9 29.09^2 (1-0.83)/8 = 161.8 → s=12.7

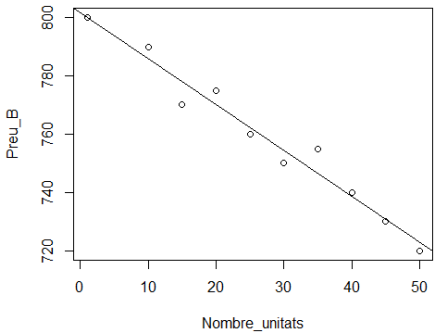
S^2 = (9 (29.09^2 – (-1.67)(-419.72))/8 = 163.4

Desviació residual val 12.7. La desviació en les prediccions serà de l'ordre de 13 euros

Un altre equip de l'empresa analitza les dades del proveïdor B i presenta els següents resultats:

Nombre d'unitats comprades en 10 comandes del darrer any: Nombre_unitats <- c(1, 10, 15, 20, 25, 30, 35, 40, 45, 50)

Preu unitari aplicat en cadascuna de les 10 comandes indicades: Preu_B <- c(800,790,770,775,760,750,755,740,730,720)




```
lm(formula = Preu_B ~ Nombre_unitats)
Coefficients:
                Estimate    Std. Error    t value        Pr(>|t|)
(Intercept)     _______      3.3566        238.87     < 2e-16 ***
Nombre_unitats  -1.5792      0.1083        ________    4.81e-07 ***
Residual standard error: 5.146 on 8 degrees of freedom
Multiple R-squared:  0.9637
```

- Indiqueu i interpreteu l'equació de la recta ajustada del Preu_B en funció del Nombre d'unitats. Compareu les dues regressions en quant a la recta ajustada, al coeficient de determinació i a la desviació residual (2 punts)

Preu_ = 801.8 – 1.58 Nombre_unitats

Des de 800.22 eur que costaria un PC, per cada unitat més que es compri el preu de cadascun baixa en 1.58 eur

Les equacions de les rectes són molt semblants; en el cas A la pendent una mica més pronunciada

Els coeficients de determinació també són semblants; en el cas B una mica superior

Les desviacions residuals són més diferents, 12.7 i 5.146 respectivament. En el cas A la desviació residual és més gran degut a que els punts estan menys ajustats a la recta.

- En al cas de la segona regressió amb el Preu_B, calculeu i interpreteu un IC 95% del pendent de la recta, i plantegeu la prova d'hipòtesis de si la recta és plana o no indicant el valor de l'estadístic i la conclusió (2 punts)

$IC_{95\%}(\beta_1) = b_1 \pm t_{8,0.975}\, Sb_1 = -1.5792 \pm 2.306 \cdot 0.1083 \approx -1.5792 \pm 0.25 \approx [-1.83; -1.33]$

Amb un 95% de confiança la reducció del preu unitari per cada unitat més comprada és entre 1.33 i 1.83 eur

$H_0: \beta_1 = 0$

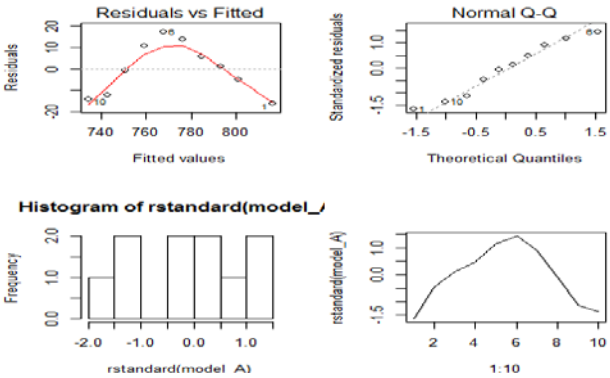
$H_1: \beta_1 <> 0$

t = -1.5792 / 0.1083 = -14.58 (p_value = 4.81e-07)

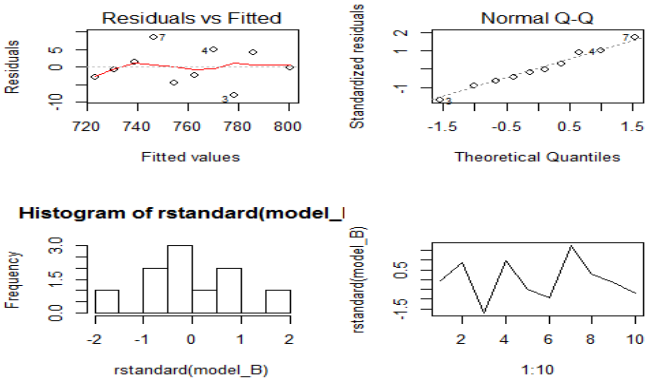
Hi ha evidència per rebutjar H_0,
ja que el valor posat a prova (0) no pertany al IC
o bé l'estadístic (-14.58) està fora dels punts crítics ($\pm$ 2.306,)
o bé el p-value és < que risc del 5%
Per tant, hi ha prou evidència per dir que la recta no és plana

- Enuncieu les premisses de la regressió lineal i compareu-les en els dos models (indiqueu els gràfics d'on es dedueixen) (2 punts)

`Preu_A ~ Nombre_unitats`



`Preu_B ~ Nombre_unitats`



Linealitat: clara al plot entre Preu_B i Nombre_unitats del gràfic de l'enunciat, i amb poca dispersió. No clara en el del Preu_A

Homoscedasticitat: raonable en el cas B (no hi ha patró, no hi ha grans zones amb més i menys variabilitat en gràfic dalt-esquerra i baix-dreta). Menys clara en cas A en que la variabilitat va augmentant i disminuint (tb en punts propers o no a la recta ajustada en gràfic enunciat)

Normalitat: l'ajust a una normal és força correcte en el NormalQQ,i histograma en el cas B. Menys clara en cas A

Independència: molt raonable en el cas B (no hi ha patró que indiqui dependència en gràfic dalt-esquerra i baix-dreta). En cas A sí hi ha dependència amb dues zones clares en gràfic dalt-esquerrra

- Feu una valoració global dels dos models de regressió en quant a quins resultats són semblants i quins no entre els dos proveïdors, i en quant a quin preferirieu i perquè (2 punts)

L'equació de la recta ajustada i el coeficient de determinació són semblants. La variabilitat residual és més gran en el cas A i les premisses fallen en el cas A (sobretot linealitat, independència i homocedasticitat)

En el cas B el model lineal és més adequat, el cas A fallen premisses. En el cas B hi ha una reducció lineal del preu per unitat conforme augmenten el nombre d'unitats (la recta ajusta al llarg de tot el rang de valors del Nombre d'unitats demanadaes). En el cas A semblaria ajustar una recta de pendent més plana en la primera part (fins 30 unitats) i una altra de pendent més pronunciada en la segona part. Per tant fins 30 unitats el proveïdor A és pitjor (reducció de preu unitari menys accentuada) però a partir de 35 el proveïdor A accentua més el descens en el preu unitari. De totes maneres al final arriben a reduccions semblants per tant en global és millor el B