

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

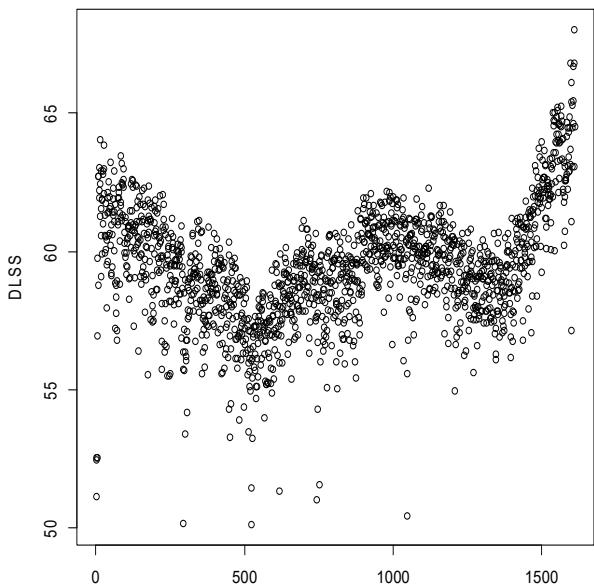
## Problema 2 (Bloc C)

*Contesteu cada pregunta en el seu lloc, amb lletra clara. Expliciteu i justifiqueu els càlculs*

Uns companys han comparat el rendiment d'un videojoc quan utilitzem una targeta RTX de NVIDIA amb tecnologia DLSS, i quan no s'utilitza, i han recollit la resposta FPS (fotogrames renderitzats per segon [fps]).

	Min.	1r Qu.	Mediana	Mitjana	3er Qu.	Max
Sense DLSS	25.42	29.91	32.07	32.24	34.08	42.91
Amb DLSS	50.11	58.57	60.03	62.34	63.08	80.40

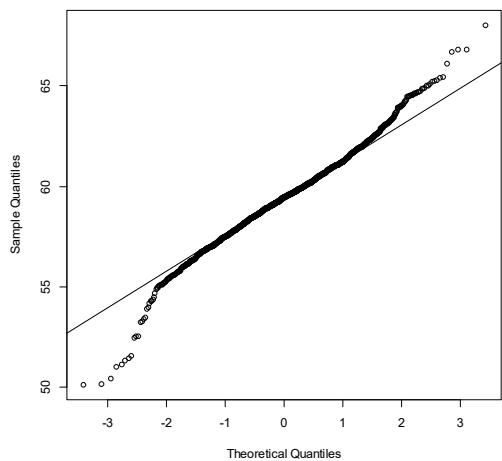
- 1) Trobeu a faltar alguna informació rellevant amb aquesta primera descriptiva? Quina? **(0.5 punts)**



- 2) Aquest gràfic mostra els valors de FPS pel grup amb DLSS en funció de l'ordre de la recollida de les dades. Raoneu perquè s'aprecia certa variabilitat en FPS (és a dir, perquè obtenim valors diferents). Discutiu també si aquesta seqüència de valors de la resposta us sembla adequada per l'estudi i anàlisi posterior. **(1 punt)**

- 3) Inventeu un valor raonable de la desviació tipus per FPS amb DLSS, i interpreteu. **(0.5 punts)**

- 4) Interpreteu el QQPLOT de la variable FPS amb DLSS (dreta). **(0.5 punts)**



Altres company han evaluat quant més lent és un llenguatge interpretat com Python (P) respecte a un de compilat com C++ (C). Han generat 100 vectors de entre 30.000 i 50.000 elements i els han ordenat amb els dos llenguatges. Finalment han mesurat els respectius temps, i els han transformat, amb el logaritme base 2, creant les variables  $\ell_2 P$  i  $\ell_2 C$ . La correlació entre les dues mesures val  $\text{Corr}(\ell_2 P, \ell_2 C)=0.9$ .

5) Expliqueu si la descripció anterior correspon a un disseny aparellat o de dues mostres independents, i quin és l'avantatge estadístic que ofereix el tipus de disseny aplicat **(1.5 punt)**

```
> D <- log2py - log2cmasmas
> summary(lm(D~1))

Call:
lm(formula = D ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.43852 -0.03019 -0.01762  0.00004  0.97831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.3405     0.0132   480.3   <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: [redacted] on [redacted] degrees of freedom
```

A la esquerra teniu el codi i els resultats de R de les dades transformades [log base 2].

6) Estimeu puntuallment l'avantatge de C respecte a P en termes de log-2, desfeu la transformació i interpreteu **(1.5 punt)**

7) Estimeu per interval del 95% l'avantatge de C respecte a P en termes de log-2 i desfeu la transformació **(1 punt)**

8) Trobeu el valor que hem eliminat (“residual standard error”). Expliqueu en què es diferencien aquest indicador i el anomenat “Std. Error” (què vol dir cadascun?). **(1.5 punt)**

9) Interpreteu els resultats. **(1 punt)**

10) Heu trobat un article vell que reproduceix un estudi com el vostre, però els autors comparen Python amb C utilitzant dues mostres, de grandàries respectives 45 i 55. El article no inclou l'anàlisi estadística, només les mitjanes (10 i 3.5, en termes de logaritme en base 2) i les desviacions tipus (3.2 i 1.3). Es demana que calculeu el quocient senyal/soroll amb les dades de l'article i que comenteu sobre el resultat obtingut. **(1 punt)** *(Responeu aquesta qüestió en un full apart)*

```
> qt(0.95, c(49,50,98,99,100))
[1] 1.676551 1.675905 1.660551 1.660391 1.660234
```

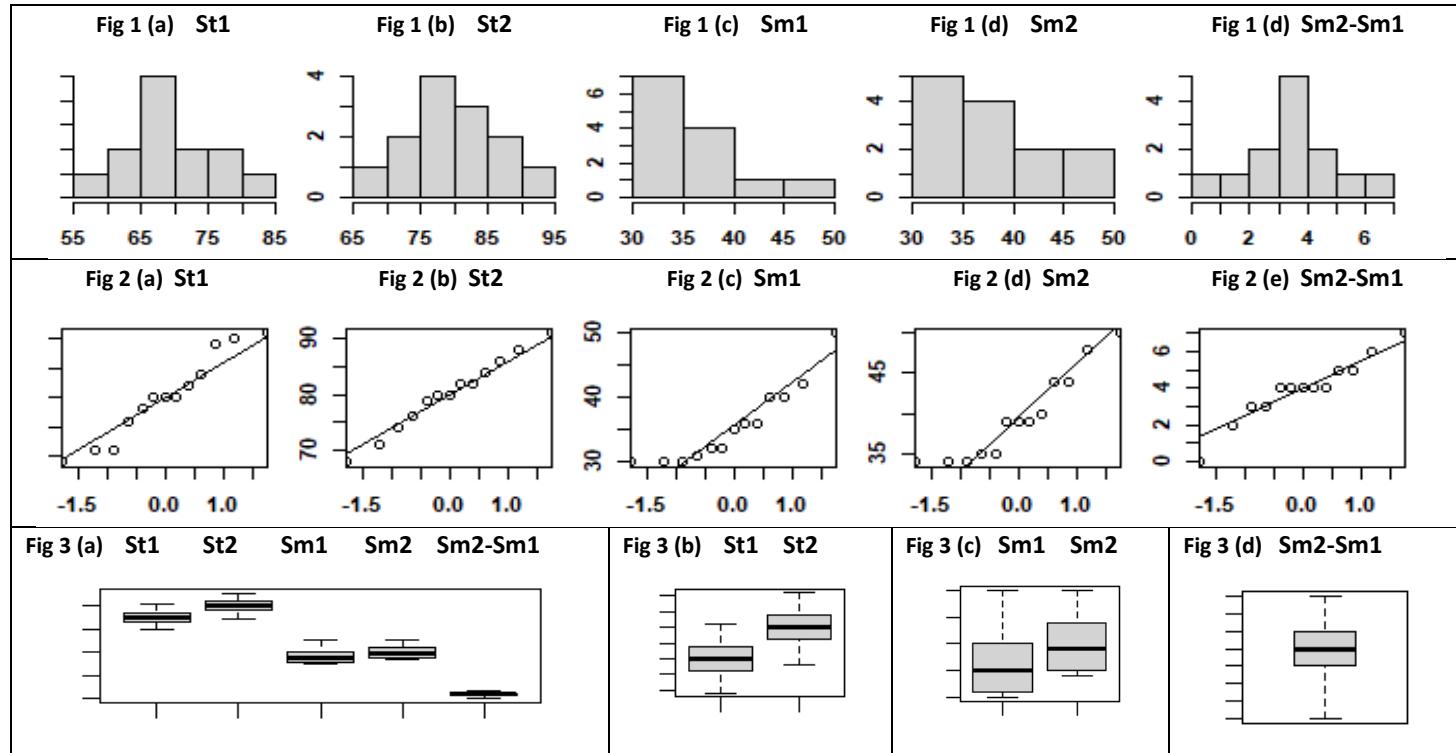
```
> qt(0.975, c(49,50,98,99,100))
[1] 2.009575 2.008559 1.984467 1.984217 1.983972
```

NOM: \_\_\_\_\_

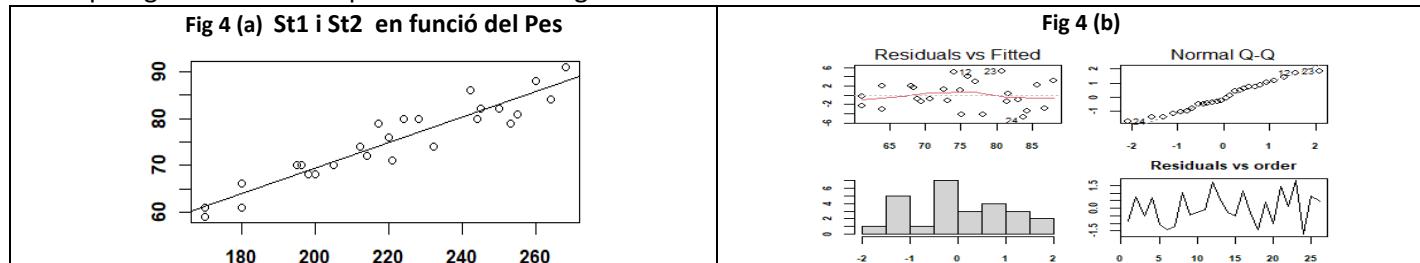
## Problema D

Per comparar 2 aparells (1 i 2) d'extreure Suc a taronges tenim m.a.s. dels grams extrets en taronges senceres i partides: St1 i St2 són els grams en 26 taronges senceres (13 en cada aparell); i Sm1 i Sm2 en 13 taronges partides (una meitat a cada aparell). També es té el pes de les 26 taronges no partides. Alguns dels resultats per St1, St2, Sm1, Sm2, i Sm2-Sm1 són:

$\text{mean(St1)}=70.1 \text{ sd(St1)}=7.2 \text{ mean(St2)}=80.1 \text{ sd(St2)}=6.6 \text{ mean(Sm1)}=35.7 \text{ sd(Sm1)}=6.0 \text{ mean(Sm2)}=39.6 \text{ sd(Sm2)}=5.4 \text{ mean(Sm2-Sm1)}=3.9$



I resultats pels grams i també el pes de les 26 taronges senceres:



I resultats per diversos models a partir de les dades recollides:

### MODEL A

```
lm(Y~1) # Y és Sm2-Sm1
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.9231    0.4865   8.064   3.47e-06 ***
Residual standard error: 1.754 on 12 degrees of freedom
```

### MODEL B

```
lm(Y~Aparell) # Y és St1 i St2 junts. "Aparell" és 1 o 2 segons l'utilitzat per extreure
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.077     1.918   36.535 <2e-16 ***
Aparell2     10.000     2.713    3.687  0.00116 **
Residual standard error: 6.916 on 24 degrees of freedom
Multiple R-squared:  0.3615,    Adjusted R-squared:  0.3349
```

### MODEL C

```
lm(Y~Pes) # Y és St1 i St2 junts
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.01486    4.41767   3.399  0.00236 **
Pes          0.27192    0.01983  13.711 7.55e-13 ***
Residual standard error: 2.912 on 24 degrees of freedom
Multiple R-squared:  0.8868,    Adjusted R-squared:  0.8821
predict(modelC,data.frame("Pes"=250),interval="prediction") : fit=82.99386 lwr=76.75405 upr=89.23367
predict(modelC,data.frame("Pes"=250),interval="confidence") : fit=82.99386 lwr=81.31765 upr=84.67007
```

Contesteu, en un full apart, les següent preguntes, justificant la resposta (si cal indicant en quina figura o model dels resultats anteriors us baseu).

1.- (1 punt) Indiqueu i justifiqueu si la comparació de mitjanes de grams de suc extret en taronges senceres és un estudi de mostres aparellades o independents. I també quan la comparació és de mitjanes de grams de suc extret en mitges taronges.

2.- (1.5 punts) Per concloure si els dos aparells extreuen una quantitat de suc comparable o no, indiqueu i quantifiqueu la diferència mitjana de grams de suc extrets (amb el seu error tipus) pel cas de taronges senceres i pel cas de mitges taronges. Interpreteu i justifiqueu en quin dels resultats donats baseu la resposta.

3.- (1.5 punts) Indiqueu —tant pel cas del model aplicat a mitges taronges com a senceres— les premisses i si es compleixen o no justificant-ho segons els resultats i figures donades.

4.- (1 punt) Comenteu globalment la comparativa entre els dos aparells d'extreure suc de taronges segons les dues proves. Indiqueu i justifiqueu, en quina de les 2 proves, l'interval de confiança de la mitjana seria més precís.

Per altra part, a partir del model de relació lineal entre els grams de suc extrets i el pes de les taronges senceres:

5.- (1 punt) Indiqueu quina és l'equació de la recta que relaciona els grams de suc extrets amb el pes de la taronja. Interpreteu els coeficients de la recta i justifiqueu en quin resultat anterior baseu la resposta.

6.- (1.5 punt) Per quantificar quant bo és aquest model per fer prediccions, indiqueu quin és el coeficient de determinació del model, i interpreteu-lo relacionant-lo amb el càlcul de la correlació entre pes i grams extrets i justifiqueu en quin resultat anterior baseu la resposta. I també indiqueu la desviació residual del model

7.- (1 punt) Calculeu, i indiqueu-ne el càlcul, una predicció del suc extret per una taronja de 250 grams. I doneu uns intervals, amb límit inferior i superior al 95% de confiança, per a la predicció com a puntual o com a valor esperat o mitjana, i compareu-los

8.- (1.5 punts) Per aquest model, indiqueu quines són les premisses i si es compleixen o no justificant-ho segons els resultats i figures anteriors.

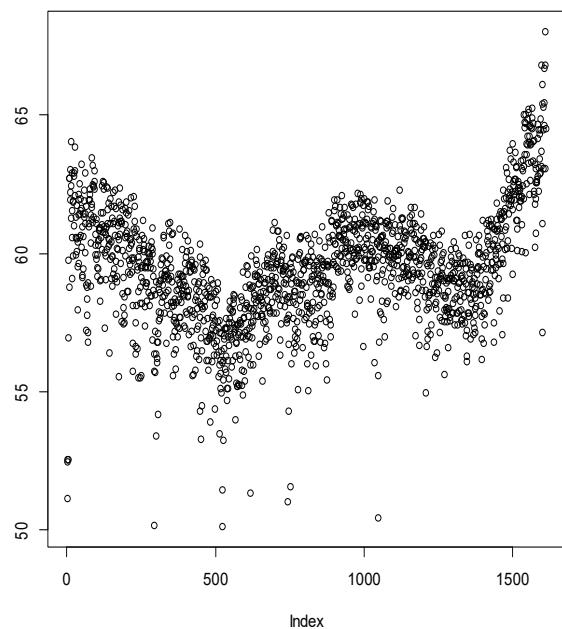
**Problema Bloc C***Contesteu cada pregunta en el seu lloc. Explicieu i justifiqueu els càlculs***Solució esperada o plantilla.** I explicació docent.

Uns companys han comparat el rendiment d'un videojoc quan utilitzem una targeta RTX de NVIDIA amb tecnologia DLSS, i quan no s'utilitza, i han recollit la resposta FPS (fotogrames renderitzats per segon [fps]).

	Min.	1r Qu.	Mediana	Media	3er Qu.	Max
Sense DLSS	25.42	29.91	32.07	32.24	34.08	42.91
Amb DLSS	50.11	58.57	60.03	62.34	63.08	80.40

- 1) Trobeu a faltar alguna informació rellevant amb aquesta primera descriptiva? Quina? **(0.5 punts)**

**El nombre de casos per grup.** [R no ho dona perquè, en un estudi ben planificat i executat, no és un resultat, sinó una decisió dels investigadors. Però en general, és la primera informació que volem saber: de quantes dades disposem? Mal si SE ja que no és descriptiva, sinó inferència. Mal si SD. Ja que hi dona els quartils i es pot obtenir el IQS, mesura robusta alternativa a la SD.]



deixant un temps entre execucions]

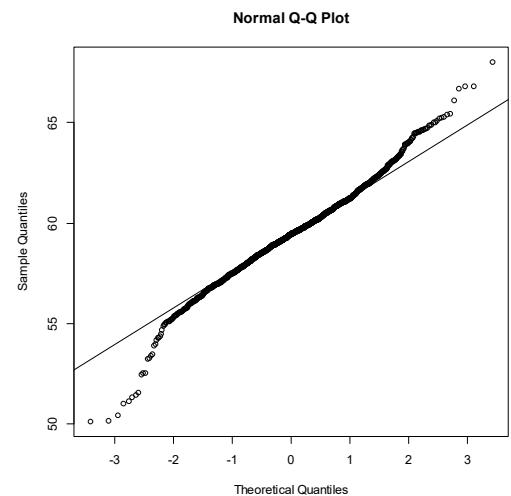
- 3) Inventeu un valor raonable de la desviació tipus pels FPS cas DLSS i interpreteu. **(0.5 punts)**

[la mitjana estaria al voltant de 60 fps (62 segons la descriptiva anterior) i ] la desviació tipus estaria al voltant de 2. Vol dir que la distància mitjana amb el centre val aproximadament 2 fps. [Exactament, en lloc de la distància mitjana, és l'arrel quadrada de la mitjana de les distàncies quadrades amb la mitjana.]

- 4) Interpreteu el QQPLOT de la variable FPS amb DLSS (dreta). **(0.5 punts)**

Els extrems mostren que les cues no ajusten bé al model normal (cues massa pesades amb molts casos).

[Hi ha dades extremes, massa distants. Convindria estudiar perquè.]



Altres companys han evaluat quant més lent és un llenguatge interpretat com Python (P) respecte a un de compilat de C++ (C). Han generat 100 vectors de entre 30.000 i 50.000 elements i els han ordenat amb els dos llenguatges. Finalment han mesurat els respectius temps, i els han transformat, amb el logaritme base 2, creant les variables  $\ell_2 P$  i  $\ell_2 C$ . La correlació entre les dues mesures val  $\text{Corr}[\ell_2 P, \ell_2 C] = 0.9$ .

5) Expliqueu si la descripció anterior correspon a un disseny aparellat o de dues mostres independents, i quin és l'avantatge estadístic que ofereix el tipus de disseny aplicat (**1.5 punt**)

La variància de la diferència com si fossin independents:  $V(I_2P - I_2C) = V(I_2P) + V(I_2C)$

I si fossin aparellades:  $V(I_2P - I_2C) = V(I_2P) + V(I_2C) - 2 \text{ Cov}(I_2P, I_2C) = V(I_2P) + V(I_2C) - 2 \text{ Corr} \cdot SD(I_2P) \cdot SD(I_2C)$

Com la correlació és positiva, reduirien la variabilitat de la diferència i també de l'error estàndard de la estimació de la diferència mitjana.

→ Les comparacions de les mitjanes serà més precises, tindran menys incertesa.

```
> D <- log2py - log2cmasmas
> summary(lm(D~1))

Call:
lm(formula = D ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.43852 -0.03019 -0.01762  0.00004  0.97831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.3405     0.0132   480.3   <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 ' ' 1

Residual standard error      on 99 degrees of freedom
```

A la esquerra teniu el codi i els resultats de R de les dades logotransformades [log base 2].

6) Estimeu puntualment l'avantatge de C respecte a P en termes de log-2, desfeu la transformació i interpreteu (**1.5 punt**)

En termes log-2, l'avantatge de C és 6.34 amb un error estàndard de 0.01.

Podem desfer la transformació per la mitjana de les diferències i trobar la mitjana geomètrica del ratis P/C:

$$2^{(6.34)} \approx 81$$

C es 81 vegades més ràpid.

7) Estimeu per interval del 95% l'avantatge de C respecte a P en termes de log-2 i desfeu la transformació (**1 punt**)

$$\text{mean}(D) \pm t_{99,0.975} \cdot \text{sd}(D) / \sqrt{n} = \text{mean}(D) \pm t_{99,0.975} \cdot \text{s.e.} \approx [6.314, 6.367]$$

$$\text{Exp}_{\text{base}2}(6.367) \approx 82,5 \quad \text{Exp}_{\text{base}2}(6.314) \approx 79,6$$

[OK si aproximeu  $t_{99,0.975} \approx 1.98$  per la normal  $\approx 1.96$ ]

8) Trobeu el valor que hem eliminat ("residual standard error"). Expliqueu en què es diferencien aquest indicador i el anomenat "Std. Error" (què vol dir cadascun?). (**1.5 punt**)

$\text{Std.error} = 0.0132 = s_D / \sqrt{n}$ , perquè el model (el més simple de tots, no conté variables explicatives) és només estimar la mitjana de D. Llavors,  $s_D = \sqrt{100} \times 0.0132 = 0.132$

El *residual standard error* en aquest cas és la desviació tipus de D, i mesura la dispersió al voltant del valor mitjà [que depèn molt de les particularitats del vector ordenat]. El *Std. Error* és l'error tipus de la mitjana de D, i estima la variabilitat de la mitjana mostra a una mostra de 100 elements com aquesta [les variacions de la mitjana són degudes a l'atzar, dels elements concrets seleccionats per formar part de la mostra].

9) Interpreteu els resultats. (**1 punt**)

Amb una confiança del 95%, C triga entre 79,6 i 82,5 vegades menys.

És un interval prou estret que sembla suficient.

10) Heu trobat un article vell que reproduueix un estudi com el vostre, però els autors comparen Python amb C utilitzant dues mostres, de grandàries respectives 45 i 55. El article no inclou l'anàlisi estadística, només les mitjanes (10 i 3.5, en termes de logaritme en base 2) i les desviacions tipus (3.2 i 1.3). Es demana que calculeu el quotient senyal/soroll amb les dades de l'article i que comenteu sobre el resultat obtingut. (**1 punt**)

La s pooled val 2.351 (encara que les desviacions semblen massa diferents entre Python i C per obtenir un valor comú).

L'indicador t (senyal/soroll) val  $(10 - 3.5) / 2.351 \sqrt{1/45 + 1/55} = 13.75$

És a dir: la diferència entre mitjanes (6.5) és quasi 14 vegades més gran que l'error mostral de la diferència de mitjanes mostrals (0.47)

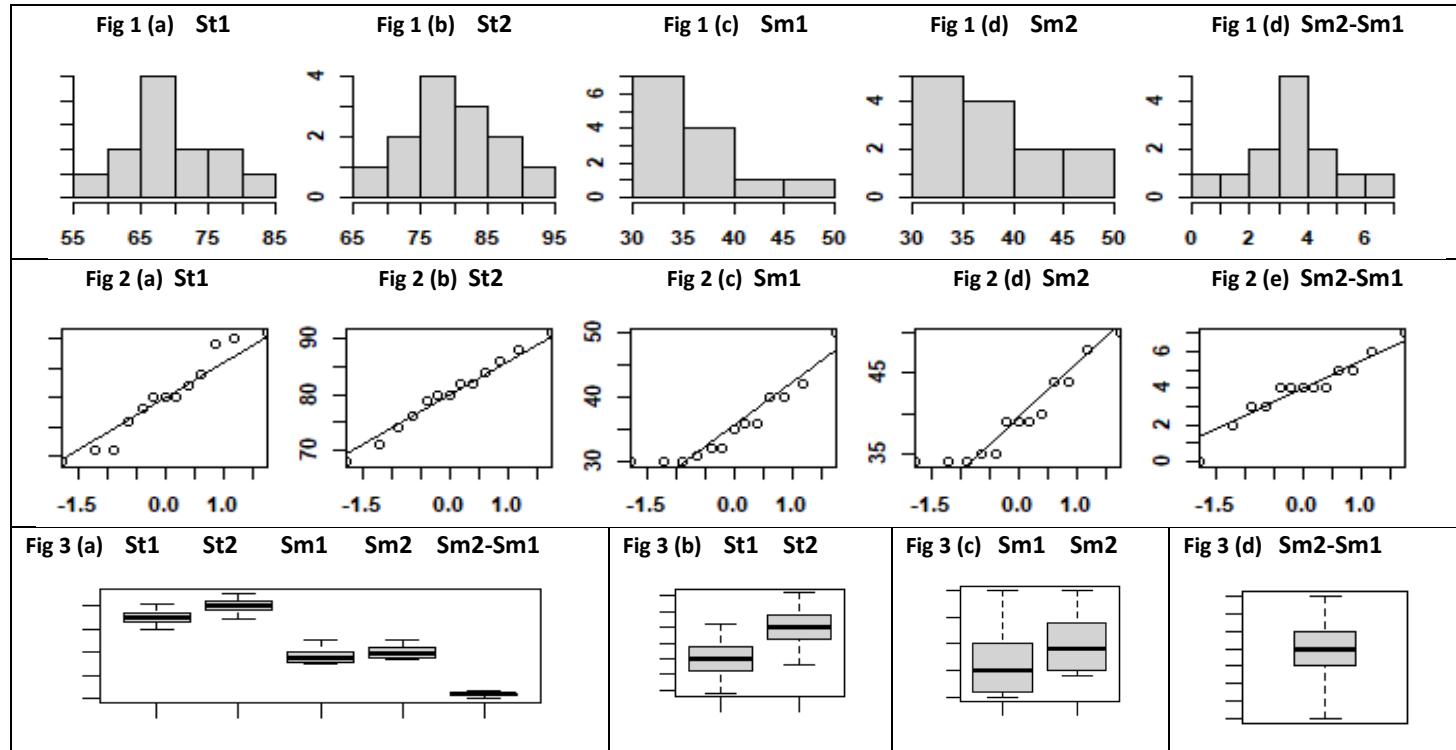
Ja sabíem que C és molt més ràpid que Python, per tant aquest resultat no és sorprenent. Si obtenim l'intervall de confiança d'aquestes dades, seria aproximadament  $6.5 \pm 1$  que, comparat amb el nostre disseny aparellat, és bastant menys precís.

NOM: \_\_\_\_\_

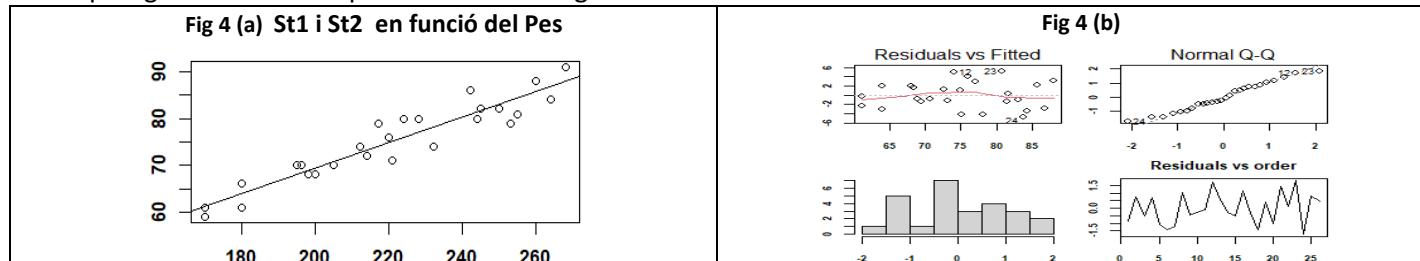
## Problema D

Per comparar 2 aparells (1 i 2) d'extreure Suc a taronges tenim m.a.s. dels grams extrets en taronges senceres i partides: St1 i St2 són els grams en 26 taronges senceres (13 en cada aparell); i Sm1 i Sm2 en 13 taronges partides (una meitat a cada aparell). També es té el pes de les 26 taronges no partides. Alguns dels resultats per St1, St2, Sm1, Sm2, i Sm2-Sm1 són:

$\text{mean(St1)}=70.1 \text{ sd(St1)}=7.2$   $\text{mean(St2)}=80.1 \text{ sd(St2)}=6.6$   $\text{mean(Sm1)}=35.7 \text{ sd(Sm1)}=6.0$   $\text{mean(Sm2)}=39.6 \text{ sd(Sm2)}=5.4$   $\text{mean(Sm2-Sm1)}=3.9$



I resultats pels grams i també el pes de les 26 taronges senceres:



I resultats per diversos models a partir de les dades recollides:

### MODEL A

```
lm(Y~1) # Y és Sm2-Sm1
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.9231 0.4865 8.064 3.47e-06 ***
Residual standard error: 1.754 on 12 degrees of freedom
```

### MODEL B

```
lm(Y~Aparell) # Y és St1 i St2 junts. "Aparell" és 1 o 2 segons l'utilitzat per extreure
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.077 1.918 36.535 <2e-16 ***
Aparell2 10.000 2.713 3.687 0.00116 **
Residual standard error: 6.916 on 24 degrees of freedom
Multiple R-squared: 0.3615, Adjusted R-squared: 0.3349
```

### MODEL C

```
lm(Y~Pes) # Y és St1 i St2 junts
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.01486 4.41767 3.399 0.00236 **
Pes 0.27192 0.01983 13.711 7.55e-13 ***
Residual standard error: 2.912 on 24 degrees of freedom
Multiple R-squared: 0.8868, Adjusted R-squared: 0.8821
predict(modelC,data.frame("Pes"=250),interval="prediction"): fit=82.99386 lwr=76.75405 upr=89.23367
predict(modelC,data.frame("Pes"=250),interval="confidence"): fit=82.99386 lwr=81.31765 upr=84.67007
```

Contesteu les següent preguntes, justificant la resposta (si cal indicant en quina figura o model dels resultats anteriors us baseu).

1.- (1 punt) Indiqueu i justifiqueu si la comparació de mitjanes de grams de suc extret en taronges senceres és un estudi de mostres aparellades o independents. I també quan la comparació és de mitjanes de grams de suc extret en mitges taronges.

Les dades de suc extret de taronges senceres són independents ja que són 13 taronges diferents que es passen per cada aparell. Les dades de suc de taronges partides són aparellades ja que cada taronja es parteix i cadascuna de les dues meitats es passen per un dels aparells.

2.- (1.5 punts) Per concloure si els dos aparells extreuen una quantitat de suc comparable o no, indiqueu i quantifiqueu la diferència mitjana de grams de suc extrets (amb el seu error tipus) pel cas de taronges senceres i pel cas de mitges taronges. Interpreteu i justifiqueu en quin dels resultats donats baseu la resposta.

En el model de mostres independents en taronges senceres (resultats del MODEL B):

**10 gr** és l'estimació del canvi de més suc extret en mitjana per l'aparell 2 (canvi a Aparell2 enlloc del 1)

**2.713 gr** és l'error tipus d'aquesta estimació

En el model de mostres aparellades en mitges taronges (resultats del MODEL A):

**3.9 gr** és l'estimació del canvi de més suc extret en mitjana per l'aparell 2 (la diferència és de l'aparell 2 menys el 1)

**0.4865 gr** és l'error tipus d'aquesta estimació

3.- (1.5 punts) Indiqueu —tant pel cas del model aplicat a mitges taronges com a senceres— les premisses i si es compleixen o no justificant-ho segons els resultats i figures donades.

En el model de mostres independents en taronges senceres, una primera premissa és que St1 i St2 siguin **mostres aleatòries simples** (m.a.s. de l'enunci). També que compleixin **normalitat**, que es veu en les figures 1(a) i 1(b) amb histogrames força semblants al model Normal, i en les figures 2(a) i 2(b) amb qqnorm força alineats seguint el model normal. I també les dues mostres independents tenen una **variabilitat semblant** i força simetria (valors de sd i gràfics 3(a) i 3(b))

En el model de mostres aparellades en mitges taronges, una primera premissa és que la mostra de 13 taronges que partim sigui **aleatòria simple** (m.a.s. de l'enunci). I també que compleixi **normalitat**, que es veu en la figura 1(e) amb historgrama força semblant al model Normal reforçat en la figura 3(d) amb boxplot molt simètric, i en les figures 2(e) amb qqnorm força alineat. Les premisses cal que es compleixen en la diferència i no en cadascuna

4.- (1 punt) Comenteu globalment la comparativa entre els dos aparells d'extreure suc de taronges segons les dues proves. Indiqueu i justifiqueu, en quina de les 2 proves, l'interval de confiança de la mitjana seria més precís.

Tant la prova aparellada com la independent compleixen les premisses, i mostren uns resultats de **més extracció de suc usant l'aparell 2**. En el cas de taronges s'extreu entorn uns 10 gr més de suc, i en taronges partides entorn uns 4 gr més.

L'**IC de la prova aparellada seria més precís**. L'amplada és  $2 * \text{Std\_error} * qt()$ , i Std\_error o error tipus en aparellada és molt menor ( $0.4865 << 2.713$ ), tot i que el valor de qt és una mica major per menys graus de llibertat ( $qt(0.975, 12) > qt(0.975, 24)$ ) ja que en un cas són 13 taronges partides i en l'altre 26 taronges senceres

Per altra part, a partir del model de relació lineal entre els grams de suc extrets i el pes de les taronges senceres:

5.- (1 punt) Indiqueu quina és l'equació de la recta que relaciona els grams de suc extrets amb el pes de la taronja.

Interpreteu els coeficients de la recta i justifiqueu en quin resultat anterior baseu la resposta.

La recta és (resultats a MODEL C com a estimadors de l'intercept o ordenada a l'origen i del coeficient que afecta al Pes):

**Suc = 15.01486 + 0.27192 \* Pes**

El coeficient del terme independent ( $b_0=15.015$  gr) indicaria els grams extrets quan Pes=0, tot i que en aquest cas no té sentit interpretar la recta per aquests valors de Pes molt allunyats del rang de pesos usats per ajustar-la.

El coeficient del pendent de la recta ( $b_1=0.27$  gr) indica que per cada gram més de pes, s'extreuen uns 0.27 gr més de suc.

6.- (1.5 punt) Per quantificar quant bo és aquest model per fer prediccions, indiqueu quin és el coeficient de determinació del model, i interpreteu-lo relacionant-lo amb el càcul de la correlació entre pes i grams extrets i justifiqueu en quin resultat anterior baseu la resposta. I també indiqueu la desviació residual del model

**R<sup>2</sup> = 88.68 %** És el coeficient de determinació (a MODEL C Multiple R-squared)

i indica que quasi un 90% de la variabilitat dels grams de suc extrets s'expliquen pel pes.

**$\sqrt{0.8868} = 0.94$**  (+0.94 ja que el pendent de la recta és positiu a la figura 4(a)) és la correlació entre Pes i Suc

i indica la forta relació positiva entre els grams de Pes i de Suc

**2.912** és la desviació residual del model (a MODEL C Residual Standard Error)

7.- (1 punt) Calculeu, i indiqueu-ne el càlcul, una predicció del suc extret per una taronja de 250 grams. I doneu uns intervals, amb límit inferior i superior al 95% de confiança, per a la predicció com a puntual o com a valor esperat o mitjana, i compareu-los

$$\text{Suc\_250} = 15.01486 + 0.27192 * 250 = 82.99 \text{ grams}$$

[76.75405, 89.23367] és l'interval per a la predicció puntual (lwr i upr a MODEL C predict amb interval="prediction")

[81.31765, 84.67007] és l'interval per a la predicció puntual (lwr i upr a MODEL C predict amb interval="confidende")

En el cas de l'interval per a la predicció com a valor puntual és més ample, ja que hi ha més incertesa, comparat amb la predicció com a valor esperat o mitjana que té menys error per calcular l'interval

8.- (1.5 punts) Per aquest model, indiqueu quines són les premisses i si es compleixen o no justificant-ho segons els resultats i figures anteriors.

En el model lineal entre Pes i Suc, unes primeres premisses són que es tracti d'una **mostra aleatòria simple** (m.a.s. de l'enunci) i que hi hagi **linealitat** entre el Pes i el Suc (comprovable a la figura 4(a) on els punts s'ajusten força al llarg de la recta)

Per altra part cal que es compleixin les premisses de normalitat, homocedasticitat i independència. La **normalitat** es pot comprovar a la figura 4(b) en l'histograma força simètric i forma de campana, i en el Normal QQ amb força alineació. En els dos altres gràfics de Residuals de la figura 4(b) es pot comprovar la **homocedasticitat** (en els residus no hi ha zones indicant més proximitat a la recta i altres allunyats, com també es veu a la figura 4(a), per tant la variabilitat al llarg de la recta és homogènia). I també la **independència** doncs no hi ha cap patró entre els valors de residus que n'indiqués alguna dependència.

## Problema 1 (Bloc C)

Nota: el separador decimal en tot l'exercici és el punt (".")

El *Cyber Monday* és un dilluns on les empreses que tenen servei de venda online fan descomptes dels seus productes. S'han recollit dades de preus de productes de 2 grans tendes ("Amazones" i "Market") abans i durant el *Cyber Monday* (A i C respectivament). Algunes dades de 2 productes (mòbils i ordinadors portàtils) de gamma intermèdia són:

		(A) Preus abans Cyber Monday		(C) Preus Cyber Monday		
Tenda	Producte	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$
Market	mòbil	12	5395	2622825	5214	2455166
Amazones	mòbil	12	4631	1963139	4568	1929330
Market	ordinador	12	5925	3174325	5782	3085308
Amazones	ordinador	12	6273	3534247	5979	3266463

		(A) Preus abans Cyber Monday		(C) Preus Cyber Monday	
Tenda	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$
Market					
Amazones					

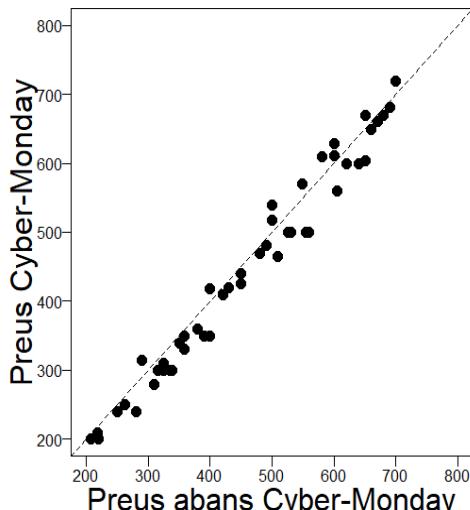
- 1.- (0.5 punts) Indiqueu els valors a la taula de la dreta per a les dues tendes (sense distingir per tipus de productes)  
 2.- (1.5 punts) Doneu una estimació puntual de la mitjana de **preus al Cyber Monday** a la tenda **Market** (indistintament que siguin mòbils o ordinadors). Quin és l'error estàndard d'aquesta estimació?

- 3.- (1 punt) Calculeu un interval de confiança (IC) del 99% per a la mitjana de **preus al Cyber Monday** a la tenda **Market** (indistintament que siguin mòbils o portàtils) suposant una desviació poblacional de 150 euros

- 4.- (1 punt) L'IC del 95% de la mitjana dels **preus de mòbils al Cyber Monday** amb  $\sigma$  desconeguda a la tenda **Amazones** és [297, 464] amb els 12 productes recollits (per tant l'amplada és 167). Amb una mostra de 24 productes de la mateixa tenda i mateixos indicadors, quin dels següents 4 intervals de confiança del 95% seria més plausible obtenir? Per què?

- a) [310, 451]      b) [340, 421]      c) [322, 439]      d) [332, 449]

- 5.- (1.5 punts) El següent gràfic representa els preus al Cyber Monday versus els preus abans del Cyber Monday pels 48 productes. La línia puntejada representa la recta bisectriu ( $y = x$ ). Observant el gràfic, indiqueu la proporció de productes que pugen de preu el Cyber Monday i calculeu i interpreteu un IC95% d'aquesta proporció



6.- (1 punt) Argumenteu quines característiques han de tenir les dades per tal de calcular un interval de confiança per a la diferència de mitjanes de preus entre les dues tendes (*Amazones* versus *Market*) com a mostres aparellades o independents

7.- (1 punt) A partir només de la següent sortida de R, calculeu un interval de confiança del 90% per a la diferència de preus del Cyber-Monday menys els preus abans del Cyber Monday.

```
Paired t-test
data: preus_cyber(C) and preus_abans(A)
t = -4.1278, df = 47, p-value = 0.0001487
alternative hypothesis: true mean difference is not
equal to 0
95 percent confidence interval:
-21.102015 -7.272985
sample estimates:
mean difference -14.1875
```

8.- (1.5 punts) Calculeu un interval de confiança del 95% per a la diferència de mitjanes de **preus abans del Cyber Monday entre mòbils (m) i ordinadors portàtils (o)** a la tenda **Amazones** assumint variàncies iguals. Interpreteu l'interval, argumenteu si podem concloure que els mòbils i els portàtils tenen diferent preu (quantificant la diferència si en tenen)

9.- (1 punt) Troba la grandària mostra necessària per a que l'anterior interval de confiança del 95% tingui una amplada de 150€ assumint que la  $\sigma$  comuna als 2 grups és 100€ i que volem el mateix nombre de portàtils que de mòbils.

qnorm(0.900) = 1.282	qnorm(0.975) = 1.960	qt(0.950,22)=1.717	qt(0.975,22)=2.074	qt(0.950,46)=1.679	qt(0.975,46)=2.013
qnorm(0.925) = 1.440	qnorm(0.990) = 2.326	qt(0.950,23)=1.714	qt(0.975,23)=2.069	qt(0.950,47)=1.678	qt(0.975,47)=2.012
qnorm(0.950) = 1.645	qnorm(0.995) = 2.576	qt(0.950,24)=1.711	qt(0.975,24)=2.064	qt(0.950,48)=1.677	qt(0.975,48)=2.011

**BD NOM:** \_\_\_\_\_ **COGNOMS:** \_\_\_\_\_

Sigueu concisos i feu lletra lleigible. *Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs. Cada apartat val 1 punt*

S'analitza si la diferència en el temps d'execució [ms] entre C i C++ a l'hora de trobar els valors propis d'una matriu depèn de la mida de la matriu (aquí, les dimensions seran  $n \times n$  on  $n$  és 10, 20, ..., 190, 200). Cada operació amb una matriu a l'atzar d'una mida determinada es fa amb cada llenguatge, obtenint-ne els temps d'execució (C i + respectivament). Per conveniència, treballarem amb el logaritme del temps, i la seva diferència “ $D = \ln(+) - \ln(C)$ ” equival al logaritme del rati: “ $\ln(+/C)$ ”. A sota veieu les sortides de R per a dos models que hem provat (dades\$Y és el logaritme dels temps; dades\$L és llenguatge: C o +).

Call: lm(formula = D ~ 1) Residuals: Min 1Q Median 3Q Max -0.41595 -0.34077 -0.05871 0.24805 0.77256 Coefficients: Estimate Std. Error t value Pr(> t ) (Intercept) 1.23682 0.08229 15.03 5.32e-12 Residual standard error: 0.368 on 19 degrees of freedom	Call: lm(formula = Y ~ L, data = dades) Residuals: Min 1Q Median 3Q Max -0.82690 -0.38031 -0.04163 0.33268 1.30651 Coefficients: Estimate Std. Error t value Pr(> t ) (Intercept) 0.5784 0.1169 4.950 1.55e-05 LC -1.2368 0.1653 -7.484 5.50e-09 Residual standard error: 0.5226 on 38 degrees of freedom. Multiple R-squared: 0.5958, Adjusted R-squared: 0.5852 F-statistic: 56.01 on 1 and 38 DF, p-value: 5.496e-09
--	---

A) Descriu quins són els dissenys i les premisses que hi ha darrera de cada opció (1, esquerra; 2, dreta).

B) Digueu si són dues opcions vàlides; si ho són, justifiqueu la resposta; si no ho són, digueu quina és l'opció apropiada.

C) Independentment de la resposta anterior, interpreteu el resultat obtingut amb l'opció 1.

D) El mateix amb l'opció 2 (a aquesta pregunta i a l'anterior, no cal considerar aspectes menors).

E) Què representa el “Residual Standard error” en cada cas?

F) Expliqueu de què ens informa el valor de “Multiple R-squared”. Perquè una opció inclou el coeficient de determinació i l’altra no?

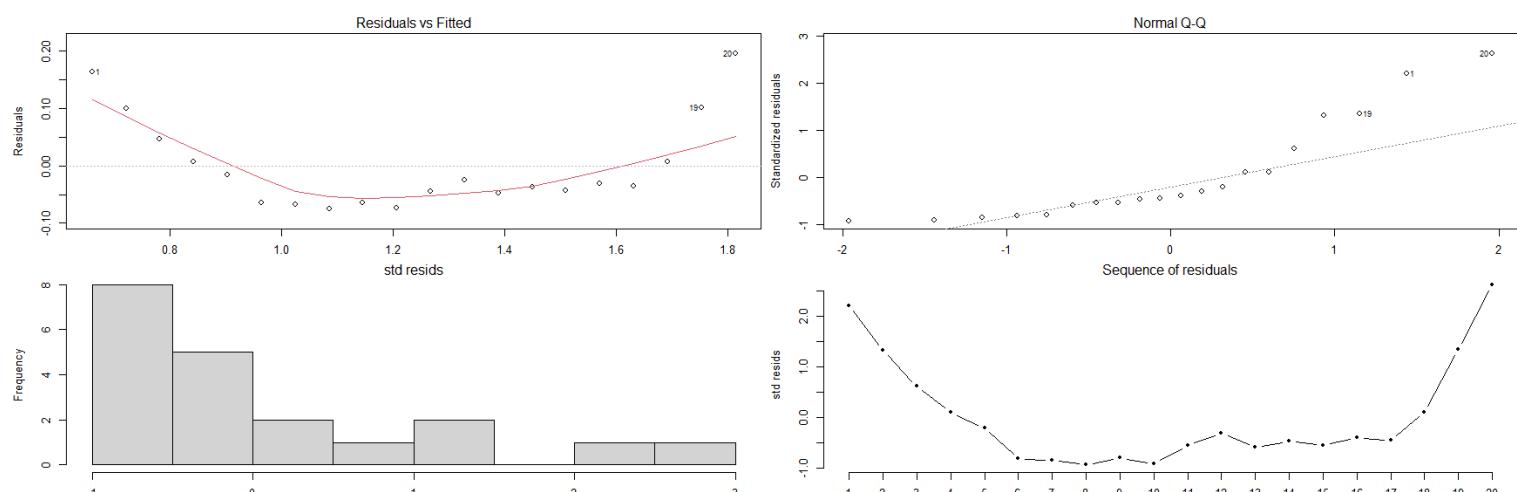
G) Calculeu un IC al 95% de confiança per al paràmetre del model 1, i explica què ens diu respecte als dos llenguatges.

H) Seguidament teniu un fragment de la sortida del model “D en funció de la mida (N)”. Comenteu què representen els dos valors de la columna “Estimate”, d’acord amb el model estadístic que hem aplicat en aquesta ocasió.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.5993784	0.0383419	15.63	6.45e-12
N	0.0060709	0.0003201	18.97	2.40e-13
Residual standard error:	0.08254	on 18 degrees of freedom		
Multiple R-squared:	0.9523,	Adjusted R-squared:	0.9497	

I) Segons la sortida anterior, utilitzeu un interval de confiança 95% per a dir com afecta la mida de la matriu en la relació entre C++ i C.



J) A la vista d’aquests gràfics, com us sembla que ha estat el compliment de les premisses del model emprat? Justifiqueu en detall la resposta.

qt(0.95, 18)=	1,7341	qt(0.975, 18)=	2,1009	qt(0.95, 19)=	1,7291	qt(0.975, 19)=	2,0930
qt(0.95, 38)=	1,6860	qt(0.975, 38)=	2,0244	qt(0.95, 39)=	1,6849	qt(0.975, 39)=	2,0227

## Problema 1 (Bloc C)

Nota: el separador decimal en tot l'exercici és el punt (".")

El *Cyber Monday* és un dilluns on les empreses que tenen servei de venda online fan descomptes dels seus productes. S'han recollit dades de preus de productes de 2 grans tendes ("Amazones" i "Market") abans i durant el *Cyber Monday* (A i C respectivament). Algunes dades de 2 productes (mòbils i ordinadors portàtils) de gama intermèdia són:

			(A) Preus abans Cyber Monday		(C) Preus Cyber Monday	
Tenda	Producte	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$
Market	mòbil	12	5395	2622825	5214	2455166
Amazones	mòbil	12	4631	1963139	4568	1929330
Market	ordinador	12	5925	3174325	5782	3085308
Amazones	ordinador	12	6273	3534247	5979	3266463

			(A) Preus abans Cyber Monday		(C) Preus Cyber Monday	
Tenda	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$	
Market	24	11320	5797150	10996	5540474	
Amazones	24	10904	5497386	10547	5195793	

- 1.- (0.5 punts) Indiqueu els valors a la taula de la dreta per a les dues tendes (sense distingir per tipus de productes)  
 2.- (1.5 punts) Doneu una estimació puntual de la mitjana de **preus al Cyber Monday** a la tenda **Market** (indistintament que siguin mòbils o ordindors). Quin és l'error estàndard d'aquesta estimació?

$$\bar{C}_M = \frac{10996}{24} = 458.17 \text{ €}$$

$$s_M^2 = \frac{(5540474) - \frac{(10996)^2}{24}}{23} = 147.8^2 \rightarrow s_M = 147.8 \text{ €}$$

$$SE_M = \frac{147.8}{\sqrt{24}} = 30.2 \text{ €}$$

- 3.- (1 punt) Calculeu un interval de confiança (IC) del 99% per a la mitjana de **preus al Cyber Monday** a la tenda **Market** (indistintament que siguin mòbils o portàtils) suposant una desviació poblacional de 150 euros

$$IC(\mu_M, 99%) = \bar{C}_M \pm z_{0.995} \cdot \frac{150}{\sqrt{24}} = 458.17 \pm 2.576 \cdot \frac{150}{\sqrt{24}} = [379.3, 537.04]$$

- 4.- (1 punt) L'IC del 95% de la mitjana dels **preus de mòbils al Cyber Monday** amb  $\sigma$  desconeguda a la tenda **Amazones** és [297, 464] amb els 12 productes recollits (per tant l'amplada és 167). Amb una mostra de 24 productes de la mateixa tenda i mateixos indicadors, quin dels següents 4 intervals de confiança del 95% seria més plausible obtenir? Per què?

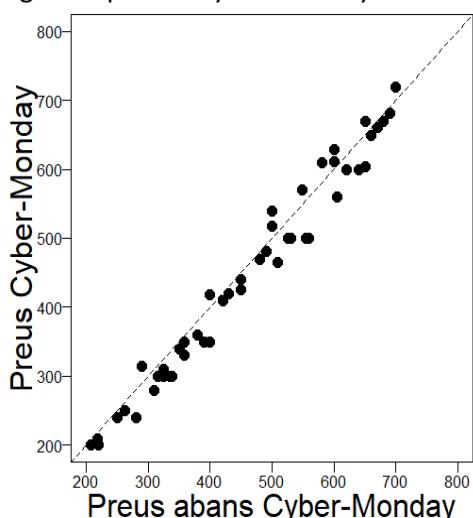
- a) [310, 451]      b) [340, 421]      c) [322, 439]      d) [332, 449]

L'amplada de l'interval que ens donen és 167. Si doblem la mostra i assumint que la desviació roman més o menys constant, esperem que l'interval de confiança tingui una amplada de  $167/\sqrt{2} \approx 118$ .

Les amplades dels intervals són 141, 83, 117 i 117 (aquest últim amb el centre molt diferent).

Per tant, l'interval més plausible és el c)

- 5.- (1.5 punts) El següent gràfic representa els preus al Cyber Monday versus els preus abans del Cyber Monday pels 48 productes. La línia puntejada representa la recta bisectriu ( $y = x$ ). Observant el gràfic, indiqueu la proporció de productes que pugen de preu el Cyber Monday i calculeu i interpreteu un IC95% d'aquesta proporció



Com que hi ha 10 punts per sobre de la bisectriu, 10 productes pugen de preu

$$\hat{p} = p = \frac{10}{48} = 0.21$$

$$IC(\pi, 95%) = p \pm z_{0.975} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0.21 \pm 1.96 \cdot \sqrt{\frac{0.21 \cdot 0.79}{48}} =$$

$$= 0.21 \pm 1.96 \cdot 0.06 = 0.21 \pm 0.12 = [0.09, 0.33]$$

Amb 95% de confiança es pot esperar entre un 9% i un 33% de productes de la categoria estudiada, que puguen de preu pel Cyber Monday

6.- (1 punt) Argumenteu quines característiques han de tenir les dades per tal de calcular un interval de confiança per a la diferència de mitjanes de preus entre les dues tendes (*Amazones* versus *Market*) com a mostres aparellades o independents

Només tindria sentit tractar-les com a aparellades si els 24 productes mostrejats a cada tenda fossin els mateixos.

Llavors els podríem aparellar per model i calcular l'interval de confiança de la diferència aparellada (en aquest cas la diferència cal que sigui normal).

En cas contrari, no es podria fer i s'hauria de calcular un interval de confiança de la diferència de mitjanes per mostres independents (en aquest cas les dues mostres han de ser normals).

7.- (1 punt) A partir només de la següent sortida de R, calculeu un interval de confiança del 90% per a la diferència de preus del Cyber-Monday menys els preus abans del Cyber Monday.

```
Paired t-test
data: preus_cyber (C) and preus_abans (A)

t = -4.1278, df = 47, p-value = 0.0001487
alternative hypothesis: true mean difference
is not equal to 0

95 percent confidence interval:
-21.102015 -7.272985
sample estimates:
mean difference -14.1875
```

Hem de deduir  $\bar{d}$  i  $s_d$  de la sortida:

$$\bar{d} = -14.1875$$

Emprant l'amplada de l'interval es pot trobar  $s_d$  i se:

$$(21.102015 - 7.272985)/2 \rightarrow 6.91 \quad 6.91 = t_{47,0.975} \cdot \frac{s_d}{\sqrt{48}} = 2.012 \cdot \frac{s_d}{\sqrt{48}}$$

$$s_d = \frac{\sqrt{48} \cdot 6.91}{2.012} = 23.8 \quad se = \frac{23.8}{\sqrt{48}} = 3.44$$

O bé amb l'estadístic calcular se:  $-4.1278 = -14.1875/se \quad se = 3.44$

$$IC(\mu_d, 90\%) = -14.1875 \pm t_{47,0.95} \cdot se \\ = -14.1875 \pm 1.678 \cdot 3.44 = [-19.96, -8.42]$$

8.- (1.5 punts) Calculeu un interval de confiança del 95% per a la diferència de mitjanes de **preus abans del Cyber Monday entre mòbils (m) i ordinadors portàtils (o)** a la tenda **Amazones** assumint variàncies iguals. Interpreteu l'interval, argumenteu si podem concloure que els mòbils i els portàtils tenen diferent preu (quantificant la diferència si en tenen)

$$\bar{m}_A = \frac{4631}{12} = 385.92 \quad s_m^2 = \frac{1963139 - \frac{(4631)^2}{12}}{11} = 126.48^2 \quad \rightarrow \quad s_m = 126.48$$

$$\bar{o}_A = \frac{6273}{12} = 522.75 \quad s_o^2 = \frac{3534247 - \frac{(6273)^2}{12}}{11} = 152.27^2 \quad \rightarrow \quad s_o = 152.27$$

$$S_{pooled}^2 = \frac{11 \cdot 126.48^2 + 11 \cdot 152.27^2}{22} = 19591.6 \quad \rightarrow \quad S_{pooled} = 139.97 \quad \rightarrow \quad se = 139.97 \cdot \sqrt{\frac{1}{12} + \frac{1}{12}} = 57.14$$

$$IC(\mu_m - \mu_o, 95\%) = (\bar{m}_A - \bar{o}_A) \pm t_{22,0.975} \cdot s_{pooled} \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_o}} = (385.92 - 522.75) \pm 2.074 \cdot 139.97 \cdot \sqrt{\frac{1}{12} + \frac{1}{12}} = \\ = (-136.83) \pm 57.14 = (-136.83) \pm 118.51 = [-255.34, -18.33]$$

Amb un 95% de confiança la diferència de mitjanes de preus està entre 18.33 i 255.34 €.

Com que el 0 no està inclòs a l'interval tenim evidència per dir que els ordinadors portàtils són en mitjana més cars que els mòbils a Amazones abans del Cyber Monday.

Abans del Cyber Monday, a Amazones els ordinadors portàtils són en mitjana entre 18.33 i 255.34 € més cars que els mòbils.

9.- (1 punt) Troba la grandària mostra necessària per a que l'anterior interval de confiança del 95% tingui una amplada de 150€ assumint que la  $\sigma$  comuna als 2 grups és 100€ i que volem el mateix nombre de portàtils que de mòbils.

$$Z_{0.975} \cdot \sigma \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_o}} = 75 \xrightarrow{n_M=n_O=n} 1.96 \cdot 100 \cdot \sqrt{\frac{2}{n}} = 75 \rightarrow n = \frac{2}{\left(\frac{75}{1.96 \cdot 100}\right)^2} = 13.66 \rightarrow \\ n = 14$$

Necessitarérem com a mínim 14 productes de cada tipus enllot de 12 per assolir aquesta precisió

qnorm(0.900) = 1.282	qnorm(0.975) = 1.960	qt(0.950,22)=1.717	qt(0.975,22)=2.074	qt(0.950,46)=1.679	qt(0.975,46)=2.013
qnorm(0.925) = 1.440	qnorm(0.990) = 2.326	qt(0.950,23)=1.714	qt(0.975,23)=2.069	qt(0.950,47)=1.678	qt(0.975,47)=2.012
qnorm(0.950) = 1.645	qnorm(0.995) = 2.576	qt(0.950,24)=1.711	qt(0.975,24)=2.064	qt(0.950,48)=1.677	qt(0.975,48)=2.011

**Model de evaluació i explicació docent. Les interpretacions genèriques (“anumèriques”) → ½ p**

S’analitza si la diferència en el temps d’execució [ms] entre C i C++ a l’hora de trobar els valors propis d’una matriu depèn de la mida de la matriu (aquí, les dimensions seran  $n \times n$  on  $n$  és 10, 20, ..., 190, 200). Cada operació amb una matriu a l’atzar d’una mida determinada es fa amb cada llenguatge, obtenint-ne els temps d’execució (C i + respectivament). Per conveniència, treballarem amb el logaritme del temps, i la seva diferència “ $D = \ln(+) - \ln(C)$ ” equival al logaritme del rati: “ $\ln(+/C)$ ”. A sota veieu les sortides de R per a dos models que hem provat (dades\$Y és el logaritme dels temps; dades\$L és llenguatge: C o +).

<pre> Call: lm(formula = D ~ 1) Residuals:     Min      1Q   Median     3Q    Max  -0.411595 -0.34077 -0.05871  0.24805  0.77256  Coefficients:             Estimate Std. Error t value Pr(&gt; t )     (Intercept) 1.23682   0.08229 15.03 5.32e-12 *** Residual standard error: 0.368 on 19 degrees of freedom </pre>	<pre> Call: lm(formula = Y ~ L, data = dades) Residuals:     Min      1Q   Median     3Q    Max  -0.82690 -0.38031 -0.04163  0.33268  1.30651  Coefficients:             Estimate Std. Error t value Pr(&gt; t )     (Intercept) 0.5784    0.1169   4.950 1.55e-05 *** LC          -1.2368   0.1653   -7.484 5.50e-09 *** Residual standard error: 0.5226 on 38 degrees of freedom. Multiple R-squared:  0.5958, Adjusted R-squared:  0.5852  F-statistic: 56.01 on 1 and 38 DF,  p-value: 5.496e-09 </pre>
---	---

A) Descriu quins són els dissenys i les premisses que hi ha darrera de cada opció (1, esquerra; 2, dreta).

1) **disseny aparellat**, tenim per a cada matriu el temps obtingut amb C, el temps obtingut amb C++ i treballem amb la diferència. Les principals premisses són m.a.s; independència entre matrius; i normalitat de la Diferència (no cal per als temps de cada opció)

2) **disseny de dues mostres independents**. Premisses: les mostres són m.a.s., Y per C és independent de Y per C++ (no s’acompleix), mateixa variància amb C i amb C++, distribució normal amb C i amb C++, per separat.

[Malament si confon model amb disseny: el disseny és com s’obtenen les dades; el model és la construcció que ens agradarà que expliqués les dades.]

Per cert: anomenarem *model lineal* als models que tenen predictors quantitatius expressats en forma lineal. Si el predictor és C++ o C no pot intervenir linealment, no podem dir “si s’incrementa el predictor un 1 ...”. Si el model no té predictors, difícilment es pot dir *lineal*]

B) Digueu si són dues opcions vàlides; si ho són, justifiqueu la resposta; si no ho són, digueu quina és l’opció apropiada.

Només és vàlida l’opció 1, perquè la resposta no s’ha recollit de forma independent (amb matrius diferents per a C i per a C++) sinó de forma aparellada.

C) Independentment de la resposta anterior, interpreteu (no llegiu) el resultat obtingut amb l’opció 1.

Estima la mitjana del logaritme del rati +/C com 1.24 (error tipus=0.08):

en mitjana, C++ triga 3.46 vegades més que C

[És molt greu no saber interpretar correctament el resultat, i dir que C++ és més ràpid, o més eficient, que C. Tampoc és admissible dir que les unitats de l’estimació 1.23682, o de la seva transformada 3.46, són ms.]

D) El mateix amb l’opció 2 (a aquesta pregunta i a l’anterior, no cal considerar aspectes menors).

La mitjana estimada per a  $\log(\text{temps C++})$  és 0.5784, error tipus=0.12; aproximadament, 1.78 ms en mitjana. Però quan el llenguatge és C, l’estimació decreix en -1.24 (error tipus=0.17).

Això vol dir que, en mitjana (geomètrica), el temps amb C és  $\exp(-1.24)=0.29$  vegades **més petit**.

E) Què representa el “Residual Standard error” en cada cas?

1) S=0.368 és la desviació tipus estimada de les diferències entre  $\log(+)$  i  $\log(C)$

2) S=0.5226 és la desviació tipus estimada en comú (pooled) dels logaritmes del temps per cada llenguatge

La primera és més petita perquè, al ser per una mateixa matriu, les diferències entre matrius ja no hi contribueixen al soroll .

F) Expliqueu de què ens informa el valor de “Multiple R-squared”. Perquè una opció inclou el coeficient de determinació i l’altra no?

Informa de la proporció de la variabilitat de la variable resposta que poden explicar els factors predictors.

Perquè el model 2 inclou un predictor (el llenguatge, C++/C), i el model 1 no en té cap.

G) Calculeu un IC al 95% de confiança per al paràmetre del model 1, i explica què ens diu respecte als dos llenguatges.

$IC = 1.23682 \pm t_{0.975, 19} 0.08229 = (1.064, 1.409)$ . Com és escala logarítmica:

$IC(\text{rati } +/C) = (2.90, 4.09)$ : en mitjana (geomètrica), C++ triga 3.46 (entre 2.90 i 4.09) vegades més que C

H) Seguidament teniu un fragment de la sortida del model “D en funció de la mida (N)”. Comenteu què representen els dos valors de la columna “Estimate”, d’acord amb el model estadístic que hem aplicat en aquesta ocasió.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.5993784	0.0383419	15.63	6.45e-12
N	0.0060709	0.0003201	18.97	2.40e-13
Residual standard error:	0.08254	on 18 degrees of freedom		
Multiple R-squared:	0.9523,	Adjusted R-squared:	0.9497	

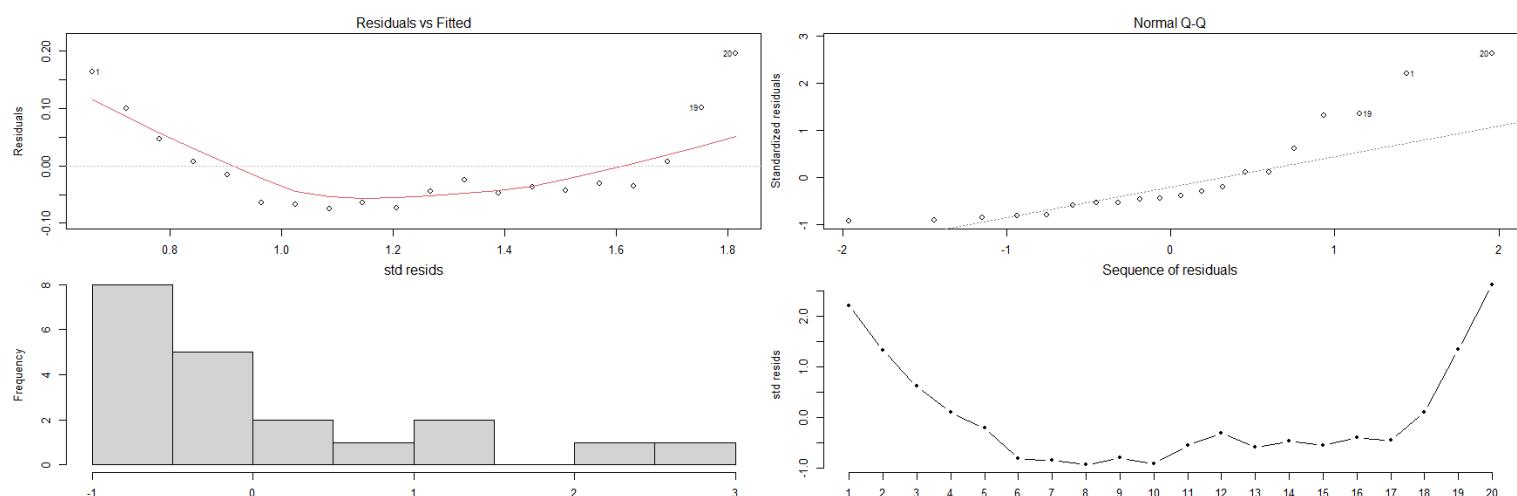
El model estadístic és  $D = \log(+/C) = 0.60 + 0.006 N$ . Equival a  $+/C = e^{0.60} e^{0.006 N} = e^{0.60} (e^{0.006})^N = 1.822 1.006^N$

0.60 és el terme independent. Diríem que *com a base* (per a matrius N=0), C++ triga 1.822 vegades més.

Si la matriu és més gran, la diferència augmenta. Concretament, per a un increment de 1 en N, el rati  $+/C$  augmenta un 0.6%.

I) Segons la sortida anterior, utilitzeu un interval de confiança 95% per a dir com afecta la mida de la matriu en la relació entre C++ i C.

$0.0060709 \pm t_{0.975, 18} 0.0003201 = (0.005398395, 0.006743405)$ . L’increment relatiu està entre 0.54% i 0.68% per qualsevol increment en N de 1 (*qualsevol*, perquè suposadament aplica un model lineal: després es veu a l’anàlisi de les premisses que realment no és així).



J) A la vista d’aquests gràfics, com us sembla que ha estat el compliment de les premisses del model emprat? Justifiqueu en detall la resposta.

Clarament, la premissa que no és correcta és la linealitat, perquè els residus dibuixen una forma corba, senyal de que el rati no varia només linealment respecte n. Per tant, el model es millorable.

Com els residus no són únicament variació aleatòria, sinó que depenen no linealment de la mida, és difícil valorar si els residus són homoscedàstics, o segueixen el model Normal (sembla que no) o són independents.

NOM: \_\_\_\_\_

COGNOM: \_\_\_\_\_

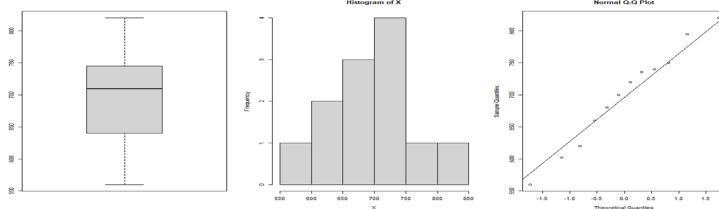
(Contesteu cada pregunta en el seu lloc. Explicieu i justifiqueu els càlculs)

**Problema 1 (B4)**

Una Universitat ha iniciat un pla de recollida de dades monitoritzant la concentració de CO<sub>2</sub> a les aules, en ppm (parts per milió).

Obtenim, a l'atzar, 12 mesures de CO<sub>2</sub> en aules amb característiques, horari i ocupacions equivalents:

$$X \sim c(560, 750, 660, 740, 620, 720, 680, 700, 820, 736, 602, 795) \quad \sum X_i = 8383 \quad \sum X_i^2 = 5923025$$



Les mesures de CO<sub>2</sub> presenten una fluctuació natural, amb desviació tipus de 70 ppm. Ens diuen també que 750 ppm és un valor de referència com a llindar superior per considerar òptima la qualitat de l'aire.

1.- Calculeu l'estimació puntual de la mitjana i la desviació de la concentració de CO<sub>2</sub> (1 punt)

2.- Amb els valors anteriors calculats, i els anteriors gràfics descriptius, comenteu la informació que donen sobre la qualitat de l'aire i sobre si es compleixen les premisses per calcular intervals de confiança (1 punt)

3.- Assumint el valor de la fluctuació natural de les mesures de CO<sub>2</sub> com a desviació poblacional, calculeu un interval de confiança al 95% per a la concentració mitjana de CO<sub>2</sub> (1 punt)

4.- I calculeu l'interval anterior si no assumim el valor anterior com a poblacional (1 punt)

5.- Interpreteu i compareu els dos intervals anteriors (2 punt)

Ara ens centrarem en unes dades d'una inspecció un dia i hora concrets en la que es prenen les mesures a 30 aules, i es tenen els dos resultats (A i B) següents:

A: <b>t = -2.04, df = 29</b> alternative hypothesis: true mean is not equal to 750 95 percent confidence interval: 677.1729 750.0937 sample estimates: mean of x 713.6333	B: <b>t = -2.04, df = 29</b> alternative hypothesis: true mean is less than 750 95 percent confidence interval: -Inf 743.9237 sample estimates: mean of x 713.6333
--	---

Una de les dues proves aporta evidència que la mitjana de ppm de les aules és inferior al llindar de 750 amb una confiança del 95%. Indiqueu quina és la prova i indiqueu hipòtesis, conclusió de la prova i interpretació de l'interval de confiança (2 punts)

Seguint amb aquestes dades de la inspecció, s'obté que de les 30 aules en 23 no es supera el llindar de 750 ppm. Indiqueu un interval de confiança al 95% pel percentatge d'aules que no superen el llindar (2 punts)

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

## Problema 2 (B5)

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

Volem comparar el temps d'execució de dos algorismes per ordenar vectors amb la mateixa complexitat de temps  $O(n^2)$ : Bubble sort, i Insertion sort. Hem generat a l'atzar 50 vectors amb una mida entre  $10^1$  i  $10^5$  i hem calculat el temps que triga cada algoritme. La taula següent proporciona la mitjana i la dispersió (desviació típica o estàndard) per a cada algoritme i per a la seva diferència. Per la resposta 'temps', esquerra; i pel seu logaritme natural, dreta. **Cada pregunta 1 punt.**

Temps en segons			Log(temps)		
Var	Mitjana	Dispersió	Var	Mitjana	Dispersió
B	193'4	175'7	In(B)	4'2	2'4
I	91'6	83'1	In(I)	3'4	2'4
B-I	101'8	98'4	In(B)-In(I)	0'74	0'03

1.- Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents

2.- Comenteu què implica cada disseny (independent o aparellat) en quant a la variància de la diferència. Dona això alguna pista sobre el grau d'aparellament (dependència) de les dades?

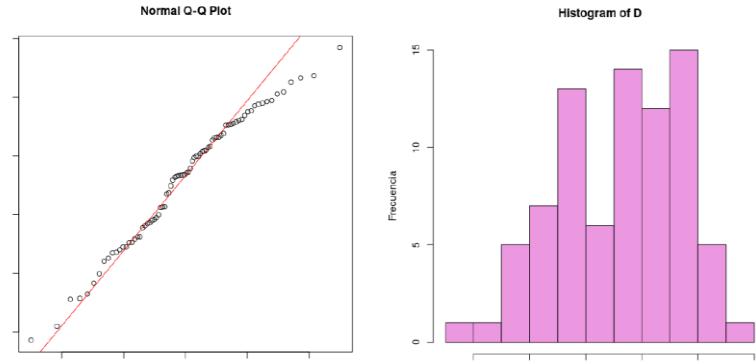
Si es tracten com a mostres independents (assumint normalitat i igualtat de variàncies poblacionals), calculeu:

3.- la desviació pooled i l'error estàndard de la diferència de mitjanes

4.- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans)

5.- Opineu sobre les premisses

6.- Es considera ara la diferència dels logaritmes  $D=\ln(B)-\ln(I)$ , obtenint aquests dos gràfics. Interpreteu i indiqueu de què ens informen aquests dos gràfics

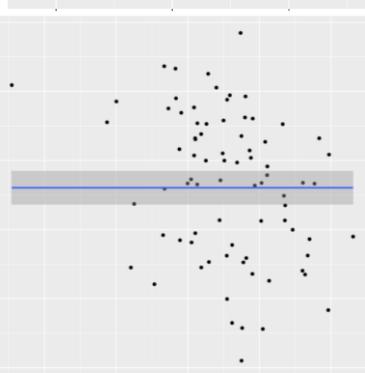
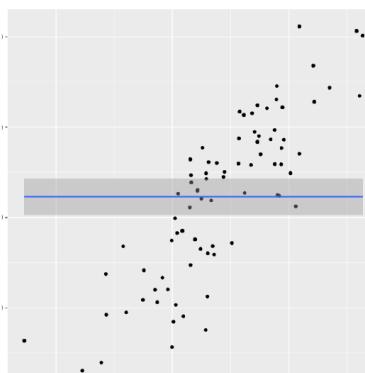


7.- Interpreteu els resultats numèrics descriptius (mitjana i desviació) de la diferència dels logaritmes. Quin triga menys? Quin és més ràpid? Quant més ràpid?

8.- Sigui T una v.a. amb distribució t de Student amb 49 graus de llibertat i  $P(T<0'68)= 0'75$ ; doneu un interval simètric de confiança per a la diferència (0.5 punts) i indiqueu amb quina confiança s'haurà calculat (0.5 punts).

9.- Interpreteu el interval de confiança en la escala del temps (desfeu els logs)

10.- Els següents dos gràfics mostren les diferències  $B-I$  per cada fitxer en ordenades en funció de les mitjanes  $[(B+I)/2]$  en abscisses. Primer, el gràfic inicial, sense transformar; i després, el gràfic amb la transformació logarítmica. Sabent que ordenar fitxers grans pot resultar en diferències mes grans, interpreteu aquests gràfics. Té sentit estimar una diferència única per aplicar a tots els casos amb les dades sense transformar (primer gràfic); i amb les dades transformades (segon)?



**B6** NOM: COGNOMS:

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

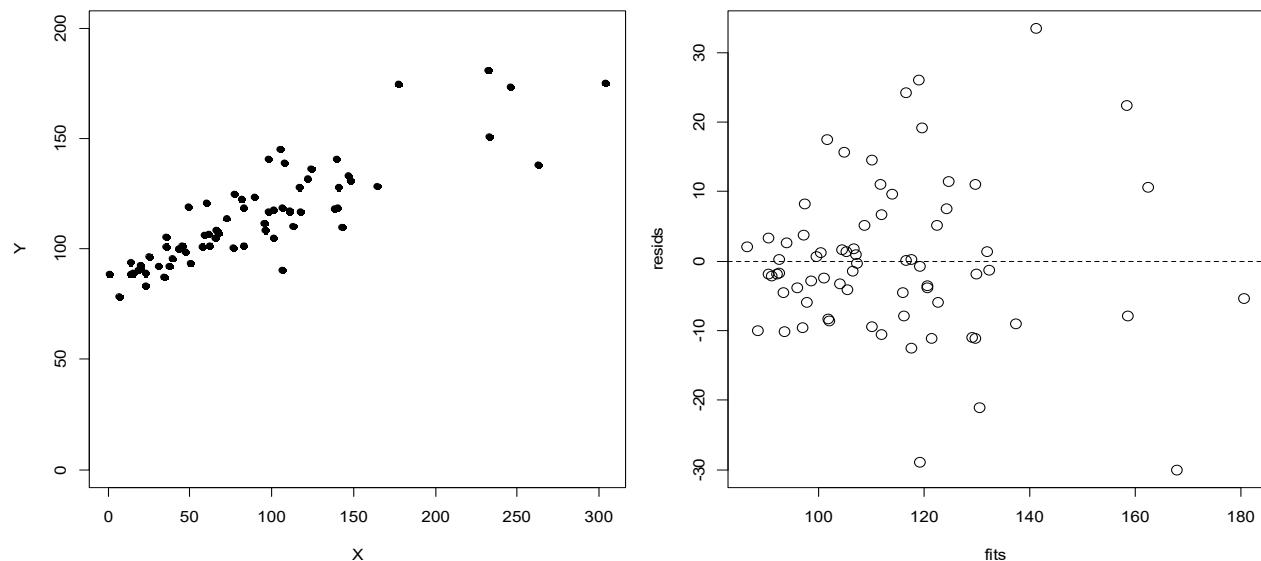
El temps emprat per a visualitzar una pàgina web a un navegador es pot descompondre en el temps destinat a la connexió amb el servidor remot, i el temps de processament del codi de la pàgina (descàrrega i mostrar a pantalla). Hem dissenyat un petit estudi, les dades del qual es troben a la dreta.

	1	2	3	4	5	6	$\sum v^2$	$\sum v$	Covar
Temps (cs)	84	96	135	112	98	136	75141	661	1365,6
Mida dades (KB)	16	45	125	76	22	190	60266	474	

La *mida de les dades* és l'espai que ocupa el fitxer HTML en kilobytes, i el *temps* és el temps mesurat en centèsimes de segon des de que es llença la petició fins a que la pàgina apareix completa al navegador (per tant, el temps total).

- [2pts] El primer objectiu de la recerca és el temps de la primera part, el destinat a la connexió, i que no depèn de la mida de les dades. L'anàlisi estadístic definit serà un model lineal amb les variables de la taula. Heu de trobar el valor de les estimacions pels paràmetres del model: 1) terme independent, 2) terme lineal, 3) desviació residual.
  - [1.5pts] Expliqueu el significat de les estimacions anteriors (sigueu curosos amb les unitats corresponents a cada cas).
  - [1.5pts] A partir de les estimacions resultants del model anterior, calculeu un interval de confiança al 95% per al temps esperat que es precisa per a connectar amb el servidor remot, i doneu una interpretació per a complementar el resultat.
  - [1.5pts] Si es vol trobar un interval més estret per al paràmetre anterior, comenteu sobre l'eficàcia de les següents estratègies, justificant les respostes (preferiblement de manera formal):
    - Empraria una mostra més gran (per exemple, 12 observacions)
    - No augmentaria la mida de mostra, però augmentaria la mida de les pàgines
    - No augmentaria la mida de mostra, però disminuiria la mida de les pàgines

5. [1.5 pts] En base a la mida de la pàgina, quina és la capacitat de predicció del temps total del model anterior? Expliqueu com l'heu deduït. Com es coneix a l'indicador que heu utilitzat i què mesura?



Hem replicat l'estudi amb moltes més dades, els resultats del qual es mostren a les figures superiors.

6. [1 pt] Expliqueu breument cada un dels dos gràfics.

7. [1 pt] Amb l'ajut dels mateixos, valideu el model lineal aplicat a aquest cas: quines premisses es podrien valorar? Quines semblen admissibles, i perquè (o perquè no)? Si veieu alguna que no admetrieu, comenteu les possibles causes i solucions.

NOM: \_\_\_\_\_

COGNOM: \_\_\_\_\_

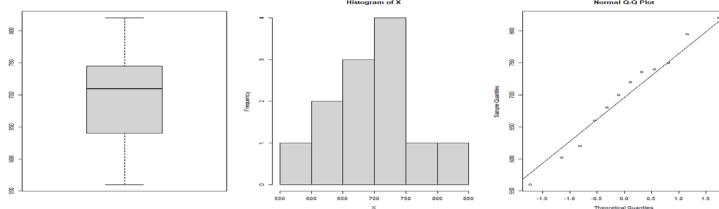
(Contesteu cada pregunta en el seu lloc. Explicieu i justifiqueu els càlculs)

**Problema 1 (B4)**

Una Universitat ha iniciat un pla de recollida de dades monitoritzant la concentració de CO<sub>2</sub> a les aules, en ppm (parts per milió).

Obtenim, a l'atzar, 12 mesures de CO<sub>2</sub> en aules amb característiques, horari i ocupacions equivalents:

$$X \sim c(560, 750, 660, 740, 620, 720, 680, 700, 820, 736, 602, 795) \quad \sum X_i = 8383 \quad \sum X_i^2 = 5923025$$



Les mesures de CO<sub>2</sub> presenten una fluctuació natural, amb desviació tipus de 70 ppm. Ens diuen també que 750 ppm és un valor de referència com a límit superior per considerar óptima la qualitat de l'aire.

1.- Calculeu l'estimació puntual de la mitjana i la desviació de la concentració de CO<sub>2</sub> (1 punt)

$$\text{mean}(X) \quad 698.58 \quad 8383 / 12 = 698.58$$

$$\text{sd}(X) \quad 77.93 \quad \sqrt{(5923025 - (8383 * 8383 / 12)) / 11} = \sqrt{6072.811} = 77.93$$

2.- Amb els valors anteriors calculats, i els anteriors gràfics descriptius, comenteu la informació que donen sobre la qualitat de l'aire i sobre si es compleixen les premisses per calcular intervals de confiança (1 punt)

La mitjana és inferior al límit de 750, però la desviació és prou gran (superior a la que es pot esperar com a fluctuació natural)

La premissa de normalitat es compleix amb un boxplot força simètric i el Normal plot força alineat, tot i que l'histograma no presenta les cues ben bé simètriques

3.- Assumint el valor de la fluctuació natural de les mesures de CO<sub>2</sub> com a desviació poblacional, calculeu un interval de confiança al 95% per a la concentració mitjana de CO<sub>2</sub> (1 punt)

$$658.98 \quad 738.19$$

$$698.58 - 1.96 * (70 / \sqrt{12}) = 698.58 - 39.61 = 658.97$$

$$698.58 + 1.96 * (70 / \sqrt{12}) = 698.58 + 39.61 = 738.19$$

4.- I calculeu l'interval anterior si no assumim el valor anterior com a poblacional (1 punt)

$$649.07 \quad 748.10$$

$$698.58 - 2.201 * (77.93 / \sqrt{12}) = 698.58 - 49.51 = 649.07$$

$$698.58 + 2.201 * (77.93 / \sqrt{12}) = 698.58 + 49.51 = 748.10$$

$$(t_{11,0.975} = 2.201)$$

5.- Interpreteu i compareu els dos intervals anteriors (2 punt)

Amb un 95% de confiança el valor de la mitjana de CO<sub>2</sub> esperable estarà entre els valors de l'interval (658.98 i 738.19 assumint el valor de sigma, i 649.07 i 748.1 si no l'assumim)

L'interval amb sigma desconeguda és més ample ja que, al no conèixer el valor de sigma, s'aproxima per l'estimador s (que té un valor una mica superior respecte si s'assumeix el valor de 70) i s'usa la distribució t enlloc de la Normal, que porten a menys precisió

Ara ens centrarem en unes dades d'una inspecció un dia i hora concrets en la que es prenen les mesures a 30 aules, i es tenen els dos resultats (A i B) següents:

A: $t = -2.04$ , df = 29 alternative hypothesis: true mean is not equal to 750 95 percent confidence interval: 677.1729 750.0937 sample estimates: mean of x 713.6333	B: $t = -2.04$ , df = 29 alternative hypothesis: true mean is less than 750 95 percent confidence interval: -Inf 743.9237 sample estimates: mean of x 713.6333
--	---

Una de les dues proves aporta evidència que la mitjana de ppm de les aules és inferior al llindar de 750 amb una confiança del 95%. Indiqueu quina és la prova i indiqueu hipòtesis, conclusió de la prova i interpretació de l'interval de confiança (2 punts)

La prova B pq és unilateral, i permet contrastar un valor respecte valors inferiors

$$H_0: \mu = 750$$

$$H_1: \mu < 750$$

Punt crític  $t_{29,0.05}$  és **-1.699** i el valor de l'estadístic (-2.04) està a la zona de rebutig (de -infinit a -1.699). Per tant hi ha evidència per rebutjar la hipòtesis nul·la de 750 ppm, i acceptar que és inferior al llindar de 750 ppm

L'interval indica que amb una confiança del 95% la mitjana de CO<sub>2</sub> en aquestes aules serà inferior a 743.92 ppm

Seguint amb aquestes dades de la inspecció, s'obté que de les 30 aules en 23 no es supera el llindar de 750 ppm. Indiqueu un interval de confiança al 95% pel percentatge d'aules que no superen el llindar (2 punts)

$$P = 23/30 = 0.767$$

$$se = \sqrt{0.767 * 0.233 / 30} = 0.08 \quad (\text{o bé } se = \sqrt{0.5 * 0.5 / 30} = 0.09)$$

$$P - 1.96 * 0.08 = 0.767 - 0.157 = 0.61 \quad (\text{o bé } 0.767 - 0.176 = 0.59)$$

$$P + 1.96 * 0.08 = 0.767 + 0.157 = 0.92 \quad (\text{o bé } 0.767 + 0.176 = 0.94)$$

- ➔ (1) [0.61, 0.92]    o (2) [0.59, 0.94]

## Problema 2 (B5)

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

Volem comparar el temps d'execució de dos algorismes per ordenar vectors amb la mateixa complexitat de temps  $O(n^2)$ : Bubble sort, i Insertion sort. Hem generat a l'atzar 50 vectors amb una mida entre  $10^1$  i  $10^5$  i hem calculat el temps que triga cada algoritme. La taula següent proporciona la mitjana i la dispersió (desviació típica o estàndard) per a cada algoritme i per a la seva diferència. Per la resposta 'temps', esquerra; i pel seu logaritme natural, dreta. **Cada pregunta 1 punt.**

Temps en segons			Log(temps)		
Var	Mitjana	Dispersió	Var	Mitjana	Dispersió
B	193'4	175'7	ln(B)	4'2	2'4
I	91'6	83'1	ln(I)	3'4	2'4
B-I	101'8	98'4	ln(B)-ln(I)	0'74	0'03

1.- Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents

D'acord amb l'enunci, es tracta de dades aparellades perquè s'ordenen els mateixos 50 vectors amb cada algoritme.

2.- Comenteu què implica cada disseny (independent o aparellat) en quant a la variància de la diferència. Dona això alguna pista sobre el grau d'aparellament (dependència) de les dades?

Si les dades fossin independents, aleshores la variància de la diferència hauria de ser la suma de les variàncies, és a dir  $\text{Var}(B-I) = \text{Var}(B) + \text{Var}(I) = 175'7^2 + 83'1^2 = 30870'5 + 6905'6 = 37776'1 = 194'4^2$

En ser les dades aparellades, la variància de la diferència segueix la relació  $V(B-I) = V(B) + V(I) - 2 \cdot \text{Cov}(B,I)$ :

$$\text{Cov}(B,I) = [V(B) + V(I) - V(B-I)]/2 = [175'7^2 + 83'1^2 - 98'4^2]/2 = 14046'8$$

$$|\text{Corr}(B,I)| = \text{Cov}(B,I)/S_B S_I = 14046'8 / 175'7 \cdot 83'1 = 0'96$$

→ B,I tenen una relació directe molt intensa (dades 'molt' aparellades)

Si es tracten com a mostres independents (assumint normalitat i igualtat de variàncies poblacionals), calculeu:

3.- la desviació pooled i l'error estàndard de la diferència de mitjanes

$$S_{\text{pooled}}^2 = \frac{(n_S - 1) \cdot S_S^2 + (n_H - 1) \cdot S_H^2}{(n_S + n_H - 2)} = \frac{49 \cdot 175'7^2 + 49 \cdot 83'1^2}{50 + 50 - 2} = 18880'0 \rightarrow S = 137'4$$

$$s.e = \sqrt{\frac{S_{\text{pooled}}^2}{n_S} + \frac{S_{\text{pooled}}^2}{n_H}} = \sqrt{\frac{18880}{50} + \frac{18880}{50}} \approx 27'5$$

4.- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans)

$$IC(95\%, \mu_H - \mu_S) = (\bar{y}_H - \bar{y}_S) \mp z_{0.975} \cdot s.e = (193'4 - 91'6) \mp 1'96 \cdot 27'5 \approx 101'8 \mp 53'9 \approx [48'9, 156'7]$$

5.- Opineu sobre les premisses

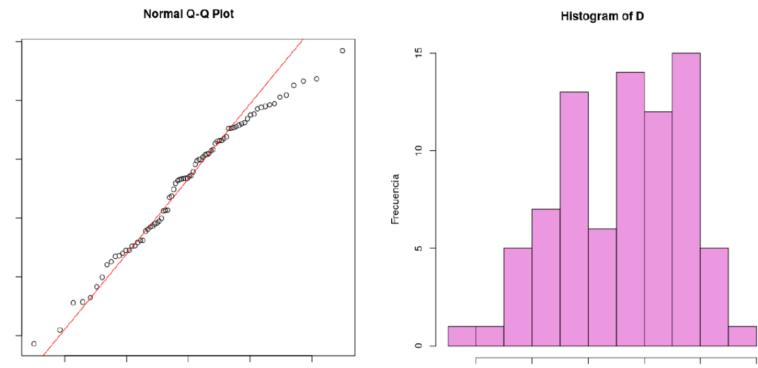
Sense gràfics de les seves distribucions (histograma i QQPLOT), només puc opinar sobre la igualtat de les variàncies, molt dubtosa, donats els resultats:  $175'7^2 = 30870'$  i  $83'1^2 = 6905'6$ ;  $\text{VAR}(B) = 4'5 \cdot \text{VAR}(I)$

6.- Es considera ara la diferència dels logaritmes  $D = \ln(B) - \ln(I)$ , obtenint aquests dos gràfics. Interpreteu i indiqueu de què ens informen aquests dos gràfics

El quantil-quantil (primer) i l'histograma (segon) ens informen de la forma de la distribució.

L'histograma és més intuïtiu, però la seva forma depèn de l'amplitud dels intervals. El quantil-quantil és més informatiu perquè reflecteix cada punt, sense necessitat de talls arbitraris.

Tots dos apunten a una distribució simètrica amb cues aplanades, que es podria modelar amb la D. Normal de Gauss-Laplace. És assenyat assumir aquesta distribució per a la inferència estadística.



7.- Interpreteu els resultats numèrics descriptius (mitjana i desviació) de la diferència dels logaritmes. Quin triga menys? Quin és més ràpid? Quant més ràpid?

En l'escala logaritme natural les diferències es distribueixen al voltant de 0'74, bastant concentrades, ja que la distància típica a aquesta mitjana val 0.03.

Això indica que B trigà el doble ( $2.09 = e^{0.74}$ ). Per tant, I serà més eficient.

8.- Sigui T una v.a. amb distribució t de Student amb 49 graus de llibertat i  $P(T < 0'68) = 0'75$ ; doneu un interval simètric de confiança per a la diferència (0.5 punts) i indiqueu amb quina confiança s'haurà calculat (0.5 punts).

$$IC(\mu_D, ??) \approx 0'74 \pm 0'68 \cdot 0'03/\sqrt{50} \approx [0'736, 0'744]$$

Com  $P(T < 0'68) = 0'75$ ,  $P(-0'68 < T < 0'68) = 0'5 \rightarrow$  Amb una confiança del 50%

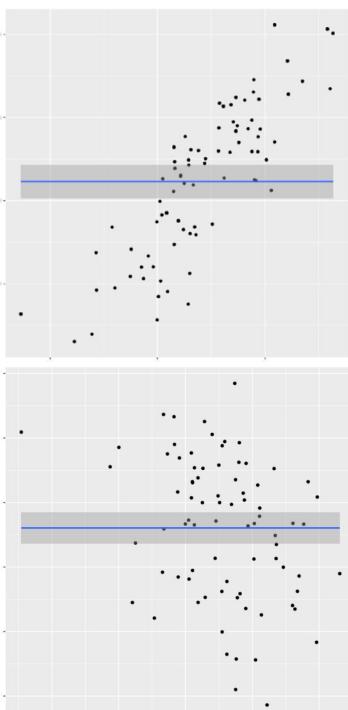
9.- Interpreteu el interval de confiança en la escala del temps (desfeu els logs)

Com  $\ln(B) - \ln(I) = \ln(B/I)$ , aleshores  $B/I = e^D$ ,

$$\text{ i } IC(B/I, 0.50) \approx e^{0.736}, e^{0.744} \approx 2.09, 2'10$$

Per vectors de mida entre  $10$  i  $10^5$ , amb una confiança del 50%, B triga entre  $2'09$  y  $2'10$  vegades més.

10.- Els següents dos gràfics mostren les diferències B-I per cada fitxer en ordenades en funció de les mitjanes  $[(B+I)/2]$  en absisses. Primer, el gràfic inicial, sense transformar; i després, el gràfic amb la transformació logarítmica. Sabent que ordenar fitxers grans pot resultar en diferències mes grans, interpreteu aquests gràfics. Té sentit estimar una diferència única per aplicar a tots els casos amb les dades sense transformar (primer gràfic); i amb les dades transformades (segon)?



En el primer gràfic, el núvol apunta a que la diferència és més gran com més gran és la mitjana: relació directa entre diferències i mitjanes. Aquells fitxers en què es triga més (potser per ser més grans?), la diferència és més gran. No té sentit proporcionar un únic valor de la diferència, una mitjana de  $101'8$  s, per a tots els vectors entre  $10$  i  $10^5$  elements..

En el segon gràfic es mostra que aplicar logaritmes (naturals) ha solucionat el problema: si té sentit proporcionar un únic valor pel rati dels temps.

El temps emprat per a visualitzar una pàgina web a un navegador es pot descompondre en el temps destinat a la connexió amb el servidor remot, i el temps de processament del codi de la pàgina (descàrrega i mostrar a pantalla). Hem dissenyat un petit estudi, les dades del qual es troben a la dreta.

(Contesteu cada pregunta en el seu lloc. Explicieu i justifiqueu els càlculs)									
	1	2	3	4	5	6	$\sum v^2$	$\sum v$	Covar
Temps (cs)	84	96	135	112	98	136	75141	661	1365,6
Mida dades (KB)	16	45	125	76	22	190	60266	474	

La *mida de les dades* és l'espai que ocupa el fitxer HTML en kilobytes, i el *temps* és el temps mesurat en centèsimes de segon des de que es llença la petició fins a que la pàgina apareix completa al navegador (per tant, el temps total).

1. [2pts] El primer objectiu de la recerca és el temps de la primera part, el destinat a la connexió, i que no depèn de la mida de les dades. L'anàlisi estadístic definit serà un model lineal amb les variables de la taula. Heu de trobar el valor de les estimacions dels paràmetres del model: 1) terme independent, 2) terme lineal, 3) desviació residual.

Y=temps; X=mida; estimarem els valors dels coeficients de  $Y=b_1X + b_0$ , i s, desviació dels residus.

$$s_x^2 = \frac{60266 - 6(474/6)^2}{5} = 4564 \quad s_y^2 = \frac{75141 - 6(661/6)^2}{5} = 464,17$$

$$b_1 = 1356,6/4564 = 0,2992 \quad (\text{terme lineal})$$

$$b_0 = (661/6) - 0,2992 \cdot (474/6) = 86,53 \quad (\text{terme independent})$$

$$s^2 = 5(464,17 - 0,2992 \cdot 1365,6)/4 = 69,45; s = 8,334$$

2. [1.5pts] Expliqueu el significat de les estimacions anteriors (sigueu curosos amb les unitats corresponents a cada cas).
- 1) 86,53 és el punt de tall de la recta amb l'eix Y, és a dir, quan la X=0. El podem interpretar com que hi ha un mínim de 86,53 cs de temps que sempre hi serà només perquè hi ha que fer la connexió.
  - 2) 0,2992 és el pendent de la recta. Significa que cada vegada que s'incrementa 1 KB la grandària del fitxer descarregat el temps total s'incrementa en 0,3 cs
  - 3) 8,334 cs és la desviació tipus de l'error de mesura. Significa que el temps té una variabilitat típica d'unes 8,3 cs, encara que la pàgina a descarregar fos la mateixa (o una altra amb la mateixa grandària).
3. [1.5pts] A partir de les estimacions resultants del model anterior, calculeu un interval de confiança al 95% per al temps esperat que es precisa per a connectar amb el servidor remot, i doneu una interpretació per a complementar el resultat.

El terme independent de la recta correspon precisament a  $\beta_0 = E(Y | X=0)$ , és a dir, al temps necessari per fer la connexió. Primer hem de trobar el valor de l'error tipus de l'estimador:

$$s_{b_0}^2 = s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) = 69,45 \left( \frac{1}{6} + \frac{79^2}{5 \cdot 4564} \right) = 30,57; \quad \text{l'error tipus és l'arrel quadrada: } 5,53$$

$$IC(\beta_0, 95\%) = 86,53 \pm t_{4, 0,975} 5,53 = [71,18; 101,88] \text{ cs}$$

Creiem amb una confiança del 95% que la mitjana del temps de connexió es situa entre 71 i 102 centèsimes de segon.

4. [1.5pts] Si es vol trobar un interval més estret per al paràmetre anterior, comenteu sobre l'eficàcia de les següents estratègies, justificant les respostes (preferiblement de manera formal):
- a) Empraria una mostra més gran (per exemple, 12 observacions)  
Si la n augmenta, disminuirà l'error tipus (la n és al denominador) i també el factor de la t-student, però la nova mostra hauria de distribuir-se de manera semblant, sense modificar substancialment la mitjana ni la variància de les X
  - b) No augmentaria la mida de mostra, però augmentaria la mida de les pàgines  
La mitjana de les X augmentaria, fent que l'error tipus s'incrementés, per tant no tindriem un IC més estret.
  - c) No augmentaria la mida de mostra, però disminuiria la mida de les pàgines  
La mitjana de les X disminuiria, i en principi també l'error tipus. No obstant, es tindria que procurar tindre la major dispersió possible en les X, perquè si la variància de les X fos molt petita l'error tipus creixeria.

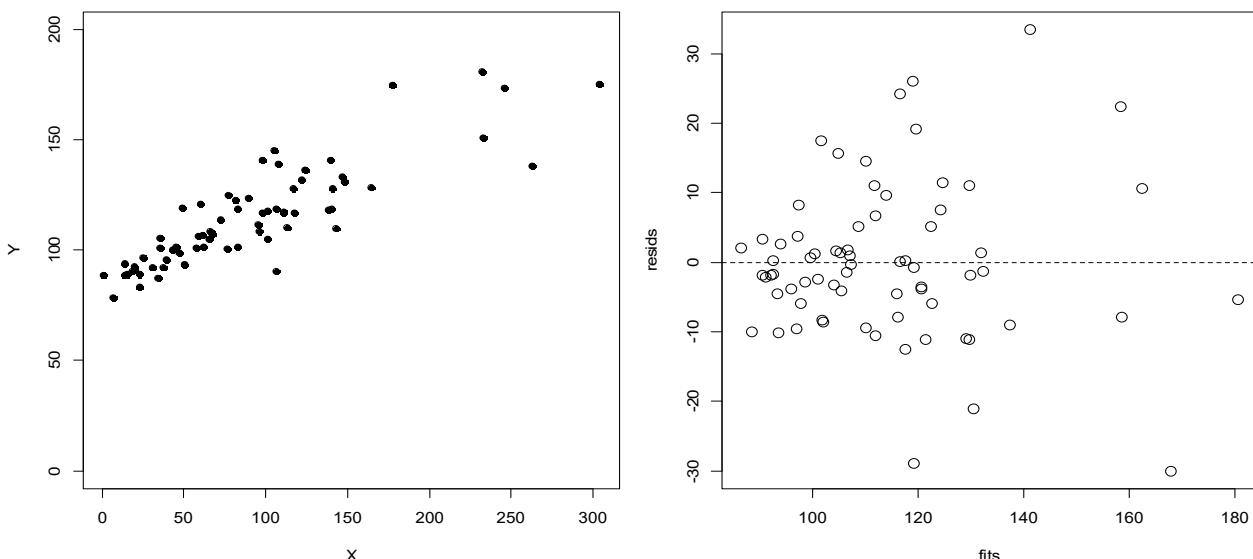
5. [1.5 pts] En base a la mida de la pàgina, quina és la capacitat de predicció del temps total del model anterior? Expliqueu com l'heu deduït. Com es coneix a l'indicador que heu utilitzat i què mesura?

L'indicador apropiat per mesurar la capacitat de predicció d'una variable per una altra és el coeficient de determinació  $R^2$ , que mesura la proporció de variabilitat de la variable resposta que es pot explicar per el model lineal que les relaciona. El coeficient  $R^2$  es pot calcular elevant la correlació de les variables al quadrat. Aquí:

$$r_{XY} = S_{X,Y} / (s_x s_y) = \frac{1365,6}{\sqrt{4564 \cdot 464,17}} = 0,9382$$

$$R^2 = 0,9382^2 = 0,8803$$

El 88% de la variabilitat observada als temps mesurats es pot associar a la mida de les pàgines descarregades. El 12% restant correspon a variabilitat d'origen desconegut (possiblement, soroll aleatori).



Hem replicat l'estudi amb moltes més dades, els resultats del qual es mostren a les figures superiors.

6. [1 pt] Expliqueu breument cada un dels dos gràfics.

A l'esquerra tenim el diagrama de les variables: a l'eix X la mida de les pàgines, a l'eix Y el temps mesurat (ho sabem perquè veiem que a valors de la X propers a 0 tenim valors de Y propers a 100, tal com hem vist als apartats anteriors). També ens mostra una relació positiva bastant forta (compatible amb la correlació trobada).

A la dreta tenim el diagrama dels residus (resids) front als valors ajustats (fits), que són els valors predicts amb la recta de regressió.

7. [1 pt] Amb l'ajut dels mateixos, valideu el model lineal aplicat a aquest cas: quines premisses es podrien valorar? Quines semblen admissibles, i perquè (o perquè no)? Si veieu alguna que no admetrieu, comenteu les possibles causes i solucions.

**Linealitat:** es pot acceptar sense problemes una relació lineal de les variables. A l'esquerra el núvol és recte i a la dreta no es veuen distorsions, sinó que es simètric respecte la línia horitzontal.

**Homoscedasticitat:** no està clar que la variància residual sigui constant, més aviat sembla evident que no ho és. En els dos gràfics es veu que els punts del costat esquerre estan més concentrats que els del altre costat. Això té una explicació senzilla, i es que el temps de descàrrega pot ser més variable a mesura que la pàgina és més gran i, per tant, la pertorbació aleatòria que afecta a les observacions no és independent de la mida, en el sentit que la variància del soroll augmenta quan la mida augmenta. La solució no és simple, si no hi hagués un terme independent gran es podria transformar les dades amb logaritmes, però en aquest cas no funcionaria bé.

**Normalitat:** es podria admetre, no tenim histograma ni QQ-plot però al gràfic dels residus es veuen els punts dispersos de forma simètrica.

**Independència:** aquests gràfics no ens poden donar cap informació.

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

## Problema 1 (B4)

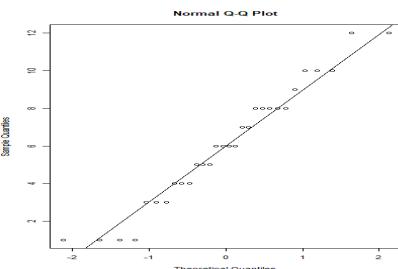
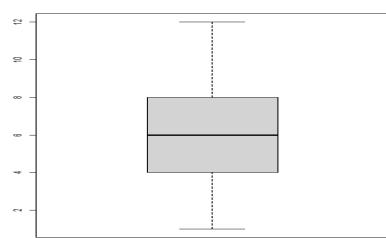
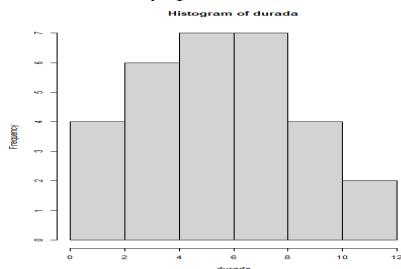
(Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs)

Per avaluar el funcionament d'un aplicatiu de salut, una de les dades que es recullen és la durada en minuts en que els usuaris estan connectats. Just abans de la pandèmia els valors poblacionals dels paràmetres que s'assumien eren de 5 minuts de mitjana i 3 minuts de desviació

Per comprovar si l'ús de l'aplicatiu ha canviat molt, es recull una nova mostra de 30 durades:

$$\sum_{i=1}^{30} \text{durada}_i = 181$$

$$\sum_{i=1}^{30} \text{durada}_i^2 = 1385$$



1.- (0.5 punts) Amb els resultats de la mostra, justifiqueu si es pot considerar que segueixen el model normal.

2.- (0.5 punts) Calculeu una estimació puntual de l'esperança i de la desviació de la durada

3.- (1.5 punt) Calculeu un interval de confiança al 95% de l'esperança de la durada assumint la desviació poblacional de 3 minuts, i interpreteu-lo

4.- (1.5 punts) Calculeu un interval de confiança al 95% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 3

5.- (1.5 punt) Calculeu un interval de confiança al 99% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 4

6.- (2 punts) Uns dels responsables de l'anàlisi d'aquestes dades tenien la sospita que la pandèmia havia fet incrementar la mitjana de la durada de la connexió dels usuaris. Per això van calcular aquests resultats:

```
t.test(durada, mu=5, alternative="greater")
t = 1.7807, df = 29, p-value = 0.04272
alternative hypothesis: true mean is greater than 5
95 percent confidence interval: 5.047337      Inf
sample estimates: mean of x 6.033333
```

Amb aquests resultats plantegeu la prova d'hipòtesis que representa (indiqueu hipòtesis, càlculs i conclusió), i interpreteu l'interval de confiança

7.- (1 punt) Amb aquesta mateixa mostra de 30 durades calculeu un interval de confiança al 95% per a la desviació i interpreteu-lo

8.- Per altra part, en un moment de molts intents d'accés, es vol quantificar el percentatge d'èxits accedint a l'aplicatiu. Per això es recull una nova mostra de 100 intents, obtenint que en 68 sí s'ha aconseguit l'accés.

(1.5 punts) Calculeu un interval de confiança per a la proporció d'intents que sí aconsegueixen accedir-hi i interpreteu-lo relacionant-lo amb el fet de que es voldria assegurar un 80% d'èxit d'accés

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

## Problema 2 (B5)

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

Per comparar la velocitat de transferència de discs HDD (H) i SSD (S), s'han copiat fitxers, de 1 MB a 4 GB, amb un mateix ordinador. Es mesura la velocitat a MB/s de copiar uns mateixos 80 fitxers amb ambdós tipus de disc i decidint a l'atzar quin s'usrà primer. La taula següent proporciona el nombre de rèpliques, la mitjana i la desviació tipus o estàndard, per a cada tecnologia de disc i per a la seva diferència:

		Velocitat en MB/s	
	Nº observacions	Mitjana	Desviació
H	80	28	3
S	80	120	9
S-H	80	92	9

Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents (0.5 punts).

Comenteu en cada cas (o disseny) què implica en quant a la variància de la diferència. (0.5 punts)

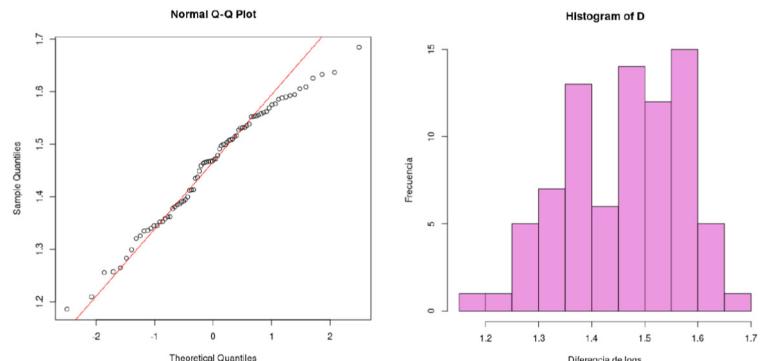
Si es tracten com a mostres independents (assumint normalitat i igualtat de variàncies poblacionals), calculeu:

- la desviació pooled i l'error estàndard de la diferència de mitjanes. (1 punt)

- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans) (1 punt)

Es considera ara la diferència dels logaritmes  $\ln(S)-\ln(H)$ , obtenint aquests dos gràfics:

(1 punt) Interpreteu i indiqueu de què ens informen aquests dos gràfics.



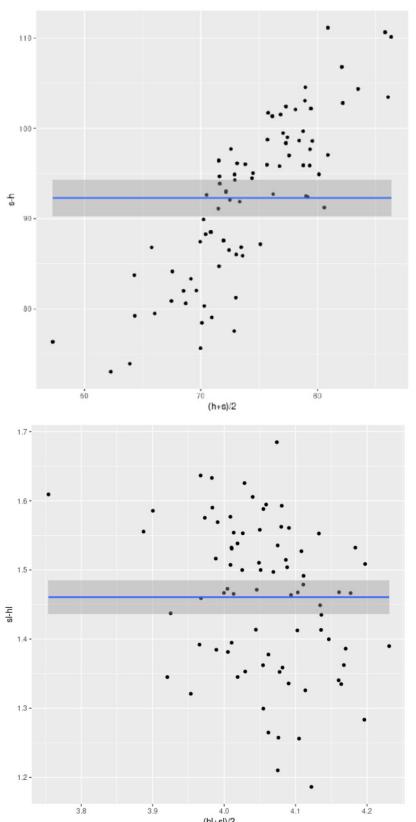
Per la variable “ $\ln(S) - \ln(H)$ ” els resultats han estat  $\sum_{i=1}^{80} x_i = 116'88$  i  $\sum_{i=1}^{80} x_i^2 = 171'72$ .  
(1 punt) Feu una estimació puntual de l’esperança i de la desviació.

(1 punt) Interpreteu les estimacions anteriors.

(2 punts) Sabent que la funció de distribució corresponent a 1'664371 en una t de Student amb 79 graus de llibertat val 0.95, useu aquest valor per donar un interval simètric de confiança per a la diferència i indiqueu amb quina confiança s’haurà calculat.

(1 punt) Interpreteu l’interval anterior en relació a la comparació de les velocitats dels discs H i S.

(1 punt) Els següents dos gràfics mostren les diferències S-H per cada fitxer en ordenades en funció de les mitjanes  $[(H+S)/2]$  en abscisses. Primer, el gràfic inicial, sense transformar; i després, el gràfic amb la transformació logarítmica. Sabent que copiar fitxers grans pot resultar en diferències més grans, interpreteu aquests gràfics. Té sentit estimar una diferència única per aplicar a tots els casos amb les dades sense transformar i amb les dades transformades?

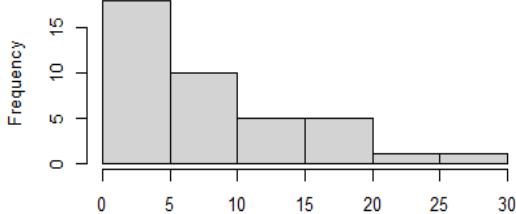
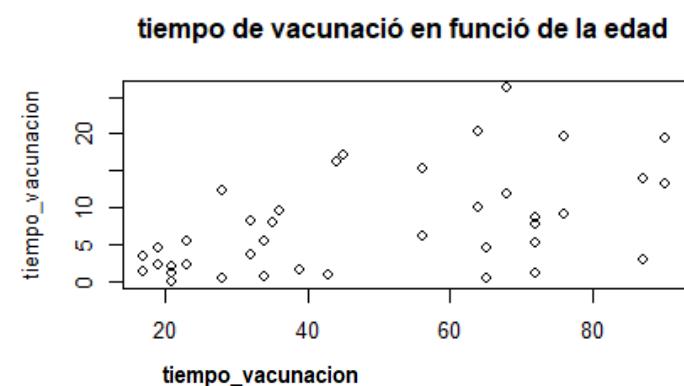
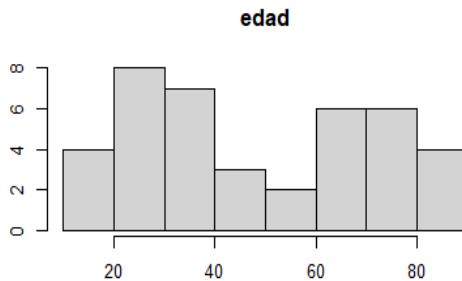


NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

### Problema 3 (B6)

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

El temps de vacunació s'entén com els minuts que passen entre la arribada d'una persona al centre sanitari, amb l'objectiu de vacunar-se, i la sortida amb la vacuna administrada. Donada l'actual situació de la COVID19 el personal sanitari es planteja si el temps de vacunació depèn de l'edat de les persones que s'han de vacunar, ja que s'han observat patrons que poden portar a pensar que sí que hi ha relació. A continuació, es proporciona alguna informació sobre les dades recollides en un centre sanitari que ha fet un estudi observacional:



1 (1 punt) Amb la informació prèvia, creieu que seria raonable plantejar un model lineal per explicar el temps de vacunació en funció de l'edat?

2 (2 punts) Donat la següent sortida de R, comenteu les estimacions de tots els paràmetres del model lineal i què signifiquen:

```

call:
lm(formula = tiempo_vacunacion ~ edad)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.265 -3.495 -1.191  2.466 15.661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.55417   2.10631   0.263 0.793894  
edad        0.14626   0.03917   3.734 0.000615 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.858 on 38 degrees of freedom
Multiple R-squared:  0.2685,    Adjusted R-squared:  0.2492 
F-statistic: 13.95 on 1 and 38 DF,  p-value: 0.0006155

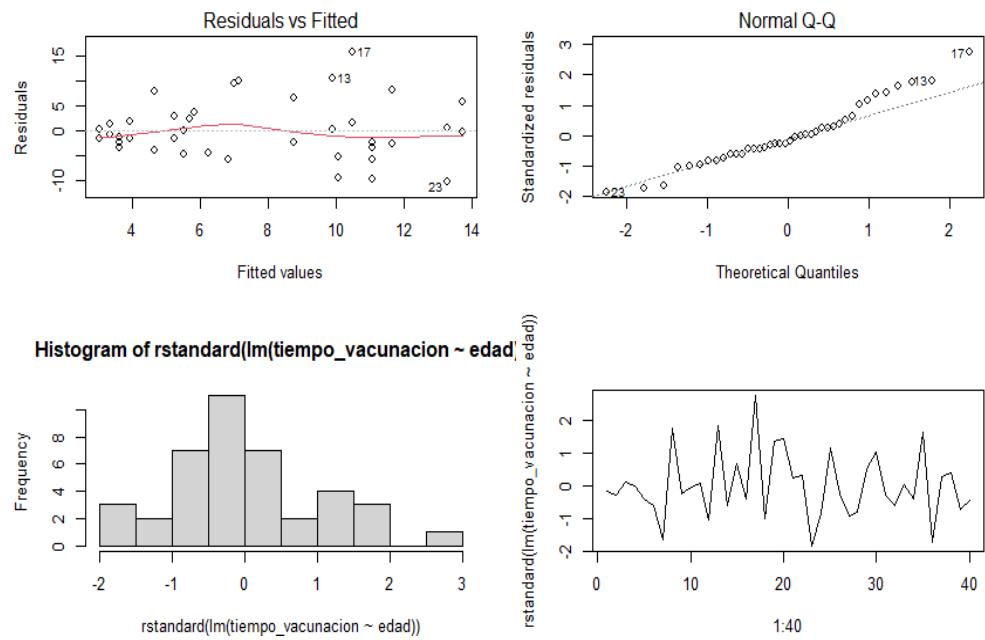
```

3 (1 punt) Valoreu la capacitat del model per explicar la variabilitat de la variable resposta.

4 (1 punts) Calculeu el IC del 90% de confiança per al pendent de la recta i interpreteu-lo.

5 (2 punts) Es vol analitzar si, donat el model anterior, podem afirmar que el temps de vacunació s'incrementa 1 minut per cada 3 anys del pacient (amb un risc del 5%):

6 (2 punts) Donats els següents gràfics de residus, valideu les premisses del model.



7 (1 punt) Si haguéssiu de fer alguna transformació sobre les dades, quina seria apparentment una transformació que podria millorar els resultats? Justifiqueu la proposta.

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

## Problema 1 (B4)

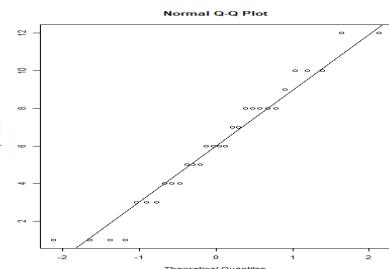
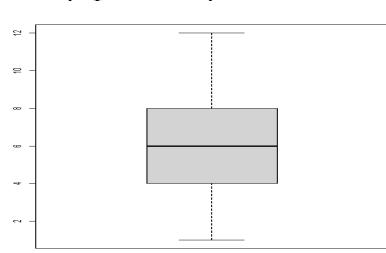
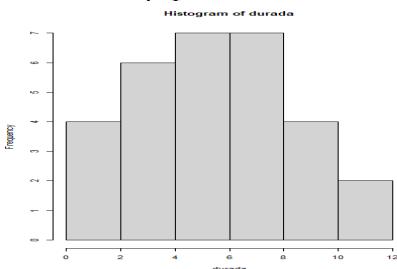
(Contesteu cada pregunta en el seu lloc. Explicieu i justifiqueu els càlculs)

Per avaluar el funcionament d'un aplicatiu de salut, una de les dades que es recullen és la durada en minuts en que els usuaris estan connectats. Just abans de la pandèmia els valors poblacionals dels paràmetres que s'assumien eren de 5 minuts de mitjana i 3 minuts de desviació

Per comprovar si l'ús de l'aplicatiu ha canviat molt, es recull una nova mostra de 30 durades:

$$\sum_{i=1}^{30} \text{durada}_i = 181$$

$$\sum_{i=1}^{30} \text{durada}_i^2 = 1385$$



1.- (0.5 punts) Amb els resultats de la mostra, justifiqueu si es pot considerar que segueixen el model normal.

Si, perquè

l'histograma sembla força la campana de Gauss (encara que cues no totalment simètriques),  
el boxplot és força simètric i centralitzat (tant la caixa com els bigotis),  
el normal\_QQ\_plot s'alinea força a una recta, indicant que els quantils de les dades es corresponen al quantils del model normal

2.- (0.5 punts) Calculeu una estimació puntual de l'esperança i de la desviació de la durada

$$\text{mitjana: } 181 / 30 = 6.03 \text{ min}$$

$$\text{desviació tipus: } s = \sqrt{(1385 - (181^2/30))/29} = 3.18 \text{ min}$$

3.- (1.5 punt) Calculeu un interval de confiança al 95% de l'esperança de la durada assumint la desviació poblacional de 3 minuts, i interpreteu-lo

$$6.03 \pm Z_{0.975} * (\sigma/\sqrt{30}) = 6.03 \pm 1.96 * (3/\sqrt{30}) = [4.96, 7.11]$$

Amb 95% de confiança la mitjana esperada de la durada de les connexions a l'aplicatiu serà d'entre una mica menys de 5 minuts (4.96 minuts) i una mica més de 7 minuts (7.11 minuts). En un 2.5% de casos es pot esperar inferior a 4.96 i en un altre 2.5 % superior a 7.11

4.- (1.5 punts) Calculeu un interval de confiança al 95% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 3

$$6.03 \pm t_{29,0.975} * (s/\sqrt{30}) = 6.03 \pm 2.045 * (3.18/\sqrt{30}) = [4.84, 7.22]$$

És un interval més ample (menys precís) pq usem la desviació estàndard de la mostra i no la desviació poblacional, usem informació de la mostra que té més incertesa que no saber el valor teòric). Usant s i no  $\sigma$  implica usar la distribució t enllloc de la Z que és més ample per la mateixa confiança

5.- (1.5 punt) Calculeu un interval de confiança al 99% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 4

$$6.03 \pm t_{29,0.995} * (s/\sqrt{30}) = 6.03 \pm 2.756 * (3.18/\sqrt{30}) = [4.43, 7.63]$$

És un interval més ample perquè el volem amb més confiança i assumint menys risc; una zona de confiança més ampla porta a valors més extrems de la distribució

6.- (2 punts) Uns dels responsables de l'anàlisi d'aquestes dades tenien la sospita que la pandèmia havia fet incrementar la mitjana de la durada de la connexió dels usuaris. Per això van calcular aquests resultats:

```
t.test(durada, mu=5, alternative="greater")
t = 1.7807, df = 29, p-value = 0.04272
alternative hypothesis: true mean is greater than 5
95 percent confidence interval: 5.047337 Inf
sample estimates: mean of x 6.033333
```

Amb aquests resultats plantegeu la prova d'hipòtesis que representa (indiqueu hipòtesis, càlculs i conclusió), i interpreteu l'interval de confiança

H0:  $\mu=5$

H1:  $\mu>5$  (hipòtesis unilateral)

Estadístic:  $(6.03-5)/(3.18/\sqrt{30})$  **1.78**

Punt crític:  $t_{29,0.95}$  **1.699**

No és raonable acceptar H0 (mitjana esperada de la durada de 5 minuts) sinó la H1 indicant que la mitjana esperada és superior a 5 minuts, ja que l'estadístic > punt crític, el p-value < risc del 5%, el valor 5 queda fora del IC [5.047,Inf] )

**IC [5.047,Inf]** és unilateral i indica que amb confiança del 95% com a mínim la durada mitjana serà una mica superior a 5 minuts (concretament superior a 5.047 minuts)

7.- (1 punt) Amb aquesta mateixa mostra de 30 durades calculeu un interval de confiança al 95% per a la desviació i interpreteu-lo

IC var [6.41, 18.25]  $(10.1*29) / 45.722$  i  $(10.1*29) / 16.047$

IC desv: **[2.53, 4.27]**

Amb un 95% de confiança la desviació esperada en la durada de les connexions és d'entre 2.46 minuts i 4.15 minuts (res s'oposaria a acceptar que el valor de 3 minuts fos una opció vàlida per a la desviació esperada)

8.- Per altra part, en un moment de molts intents d'accés, es vol quantificar el percentatge d'èxits accedint a l'aplicatiu. Per això es recull una nova mostra de 100 intents, obtenint que en 68 sí s'ha aconseguit l'accés.

(1.5 punts) Calculeu un interval de confiança per a la proporció d'intents que sí aconsegueixen accedir-hi i interpreteu-lo relacionant-lo amb el fet de que es voldria assegurar un 80% d'èxit d'accés

$\sqrt{0.68*(1-0.68)/100}$  és 0.047

( o bé  $\sqrt{0.5*0.5/n}$  és 0.05)

IC: **[ 0.59 , 0.77 ]**

$0.68 \pm 1.96*0.047$

IC: **[ 0.58 , 0.78 ]**

$0.68 \pm 1.96*0.05$

Amb un 95% de confiança el percentatge esperable d'èxits en l'accés està entre el 59% i el 77% (o 58% i 78%)

Les dades no mostren evidència que el percentatge d'èxit en l'accés sigui acceptable (no arriba al 80%)

NOM: \_\_\_\_\_ COGNOM: \_\_\_\_\_

## Problema 2 (B5)

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

Per comparar la velocitat de transferència de discos HDD (H) i SSD (S), s'han copiat fitxers, de 1 MB a 4 GB, amb un mateix ordinador. Es mesura la velocitat a MB/s de copiar uns mateixos 80 fitxers amb ambdós tipus de disc i decidint a l'atzar quin s'usrà primer. La taula següent proporciona el nombre de ràpides, la mitjana i la desviació típica o estàndard, per a cada tecnologia de disc i per a la seva diferència:

		Velocitat en MB/s	
	Nº observacions	Mitjana	Desviació
H	80	28	3
S	80	120	9
S-H	80	92	9

Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents (0.5 punts).

D'acord amb l'enunciat, es tracta de dades aparellades perquè es copien els mateixos 80 fitxers en cada tipus de disc (H i S)

Comenteu en cada cas (o disseny) què implica en quant a la variància de la diferència. (0.5 punts)

Si les dades fossin independents, aleshores la variància de la diferència és la suma de les variàncies, és a dir,  $\text{Var}(S-H)=\text{Var}(S)+\text{Var}(H)=3^2+9^2=90$ .

En ser les dades aparellades, la variància de la diferència segueix la relació següent  $V(S-H)=V(S)+V(H)-2\cdot\text{Cov}(S,H)$ , de la qual no tenim informació directe sobre la covariància, però si de  $V(S-H)=9^2$ .

Si es tracten com a mostres independents (assumint normalitat i igualtat de variàncies poblacionals), calculeu:

- la desviació pooled i l'error estàndard de la diferència de mitjanes (1 punt)

$$s_{\text{pooled}}^2 = \frac{(n_S - 1) \cdot S_S^2 + (n_H - 1) \cdot S_H^2}{(n_S + n_H - 2)} = \frac{79 \cdot 9^2 + 79 \cdot 3^2}{80 + 80 - 2} = 45; s_{\text{pooled}} = 6.708$$

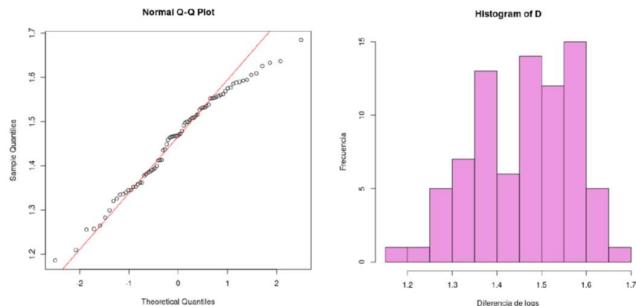
$$s.e = \sqrt{\frac{s_{\text{pooled}}^2}{n_S} + \frac{s_{\text{pooled}}^2}{n_H}} = \sqrt{\frac{45}{80} + \frac{45}{80}} = \sqrt{\frac{9}{8}} \approx 1.06$$

- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans) (1 punt)

$$IC(95\%, \mu_H - \mu_S) = (\bar{y}_H - \bar{y}_S) \mp z_{0.975} \cdot s.e = (28 - 120) \mp 1.96 \cdot 1.06 \approx -92 \mp 2.08 \approx [-94, -90]$$

Es considera ara la diferència dels logaritmes  $\ln(S)-\ln(H)$ , obtenint aquests dos gràfics:

(1 punt) Interpreteu i indiqueu de què ens informen aquests dos gràfics



El quantil-quantil (esquerre) i l'histograma (dreta) ens informen de la forma de la distribució. L'histograma és més intuïtiu, però la seva forma depèn de l'amplitud dels intervals. El quantil-quantil és més informatiu perquè reflecteix cada punt, sense necessitat de talls arbitraris.

Tots dos apunten a una distribució simètrica amb cues aplanades, que es podria modelar amb la D. Normal de Gauss-Laplace. És assenyat assumir aquesta distribució per a la inferència estadística.

Per la variable "ln(S)-ln(H)" els resultats han estat  $\sum_{i=1}^{80} x_i = 116'88$  i  $\sum_{i=1}^{80} x_i^2 = 171'72$ .  
(1 punt) Feu una estimació puntual de l'esperança i de la desviació

$$\bar{x} = 1'461 \text{ i } s_x = 0'110$$

(1 punt) Interpreteu les estimacions anteriors

En l'escala logaritme natural les diferències es distribueixen al voltant de 1.46, bastant concentrades, ja que la distància típica a aquesta mitjana val 0.1. Això indica que S serà unes  $e^{1.46}$  vegades més ràpid.

(2 punts) Sabent que la funció de distribució corresponent a 1'664371 en una t de Student amb 79 graus de llibertat val 0.95, useu aquest valor per donar un interval simètric de confiança per a la diferència i indiqueu amb quina confiança s'haurà calculat

$$IC(\mu_D, ??) \approx 1.461 \pm 1.664371 \cdot 0.110 / \sqrt{80} \approx [1.436521, 1.485479] \approx [1.44, 1.49]$$

90%

(1 punt) Interpreteu l'interval anterior en relació a la comparació de les velocitats dels discs H i S.

Per poder interpretar ahan de fer exponents,

Com  $\ln(S) - \ln(H) = \ln(S/H)$ , aleshores  $S/H = e^D$ ,

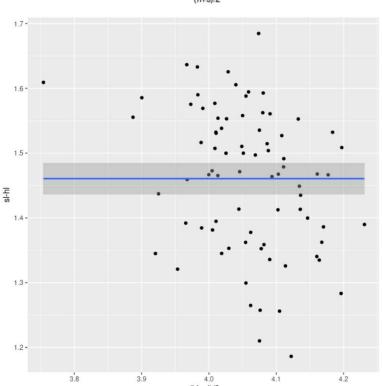
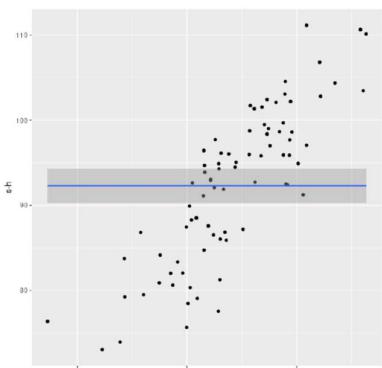
$$\text{i } IC(S/H, 0.90) \approx e^{1.436521}, e^{1.485479} \approx 4.206, 4.417 \approx 4.2, 4.4$$

Per fitxers entre 1 MB a 4 GB, amb una confiança del 90%, el disc S es entre 4.2 y 4.4 vegades més ràpid que el H

(1 punt) Els següents dos gràfics mostren les diferències S-H per cada fitxer en ordenades en funció de les mitjanes  $[(H+S)/2]$  en abscesses. Primer, el gràfic inicial, sense transformar; i després, el gràfic amb la transformació logarítmica. Sabent que copiar fitxers grans pot resultar en diferències mes grans, interpreti aquests gràfics. Té sentit estimar una diferència única per aplicar a tots els casos amb les dades sense transformar i amb les dades transformades?

En el primer gràfic, el núvol apunta a que la diferència és més gran com més gran és la mitjana: relació directa entre diferències i mitjanes. Aquells fitxers en què es triga més (potser per ser més grans?), la diferència és més gran. Fa dubtar si té sentit proporcionar un únic valor, la mitjana de 92 MB/s, per a tots els fitxers.

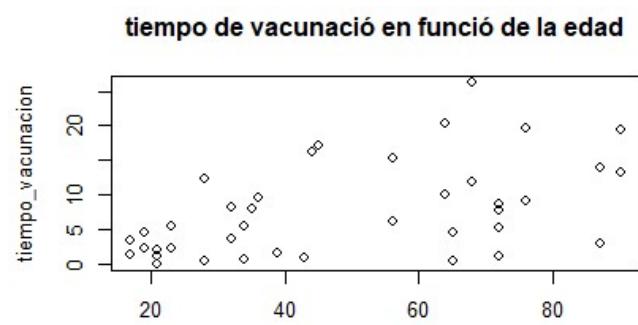
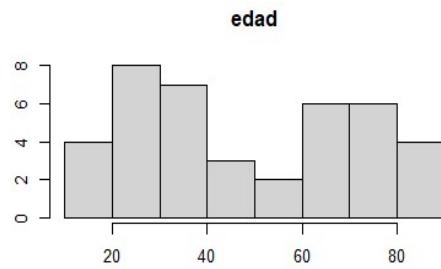
En el segon gràfic es mostra que aplicar logaritmes (naturals) ha solucionat el problema



**Problema 3 (B6)**

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

El temps de vacunació s'entén com els minuts que passen entre la arribada d'una persona al centre sanitari, amb l'objectiu de vacunar-se, i la sortida amb la vacuna administrada. Donada l'actual situació de la COVID19 el personal sanitari es planteja si el temps de vacunació depèn de l'edat de les persones que s'han de vacunar, ja que s'han observat patrons que poden portar a pensar que sí que hi ha relació. A continuació, es proporciona alguna informació sobre les dades recollides en un centre sanitari que ha fet un estudi observacional:



1 (1 punt) Amb la informació prèvia, creieu que seria raonable plantejar un model lineal per explicar el temps de vacunació en funció de l'edat?

En el primer gràfic, aparentment hi ha una correlació positiva encara que feble entre el temps de vacunació i l'edat. La relació sembla monòtona creixent. Destaquen alguns individus amb un temps de vacunació anormalment baix. D'entrada amb la informació disponible no es descartaria un model lineal.

2 (2 punts) Donat la següent sortida de R, comenteu les estimacions de tots els paràmetres del model lineal i què signifiquen:

```

call:
lm(formula = tiempo_vacunacion ~ edad)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.265 -3.495 -1.191  2.466 15.661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.55417   2.10631  0.263  0.793894    
edad        0.14626   0.03917  3.734  0.000615 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.858 on 38 degrees of freedom
Multiple R-squared:  0.2685, Adjusted R-squared:  0.2492 
F-statistic: 13.95 on 1 and 38 DF,  p-value: 0.0006155

```

Intercept o terme constant: és el terme independent de la recta, val 0.55417 (minuts), significa el temps fix que donem a un participant de qualsevol edat.

Pendent de l'edat: és el terme lineal de la recta, val 0.14626 (minuts/any), significa l'increment en el temps de vacunació per cada any d'edat addicional.

Residual Standard error, o desviació residual, val 5.858 (minuts), significa la desviació tipus de l'error aleatori que afecta a cada individu, amb una edat determinada. O l'error típic que es pot esperar a partir de l'estimació de la recta.

3 (1 punt) Valoreu la capacitat del model per explicar la variabilitat de la variable resposta.

A la vista dels resultats obtinguts el model és poc explicatiu ja que el coeficient de determinació  $R^2$  val 0,2686. És a dir, encara que el factor "edad" és estadísticament significatiu, només explica un 27% de la variació que observem en el temps de vacunació.

4 (1 punts) Calculeu el IC del 90% de confiança per al pendent de la recta i interpreteu-lo.

$$IC(\beta_1, 90\%) = 0,14626 \pm 1,684 * 0,03917 = [0,080; 0,212]$$

S'observa que el 0 no pertany a l'interval. El temps de vacunació s'incrementa entre 0.08 i 0.212 minuts en mitjana per cada any de més que tingui el pacient, amb una confiança del 90%.

5 (2 punts) Es vol analitzar si, donat el model anterior, podem afirmar que el temps de vacunació s'incrementa 1 minut per cada 3 anys del pacient (amb un risc del 5%):

Considerem una prova d'hipòtesi bilateral:

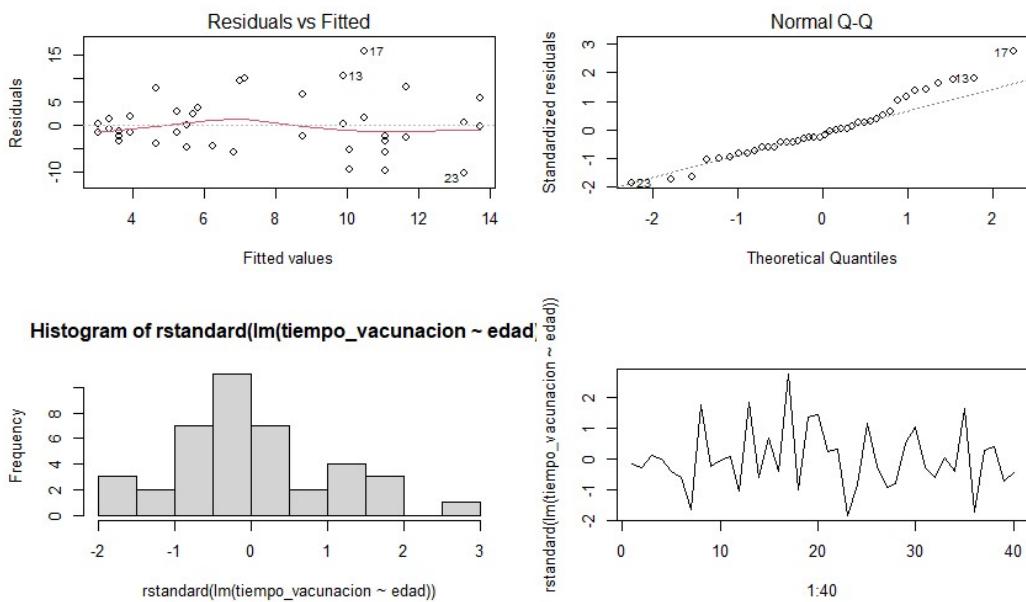
$$H_0: \beta_1 = 1/3$$

$$H_1: \beta_1 \neq 1/3$$

$$t = (0,146 - 0,333)/0,039 = -4,794$$

L'estadístic de la prova pren un valor considerablement extrem. Si fos certa la hipòtesi nul·la, es distribuiria com una t-Student amb 38 graus de llibertat, i el resultat -4.794 estaria molt allunyat de la zona d'acceptació (límits a prop de  $\pm 2$ ). A la vista dels resultats, clarament podem rebutjar la  $H_0$  i afirmar que el temps de vacunació no s'incrementa 1 minut per cada 3 anys d'edat.

6 (2 punts) Donats els següents gràfics de residus, valideu les premisses del model.



Com es veu en el primer gràfic, es pot assumir linealitat, però es presenten clars indicis de heteroscedasticitat en els residus. A banda d'aquest aspecte, la normalitat dels residus és questionable donats els resultats del Q-Q Plot, per la part dreta, que s'allunya més del que hauria de ser. La independència dels residus sembla que no està compromesa. Amb tot això, no es podria validar el model.

7 (1 punt) Si haguéssiu de fer alguna transformació sobre les dades, quina seria apparentment una transformació que podria millorar els resultats? Justifiqueu la proposta.

Una transformació logarítmica sobre el temps de vacunació disminuiria la variància en els individus que presenten més variabilitat (els que tenen temps més alts) i podria millorar el compliment de les premisses.