NOM:_____ COGNOM:_____

Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs.
Realitzeu tots els càlculs (finals i intermedis) amb quatre xifres decimals amb arrodoniment.

Problema 3 (Bloc C)

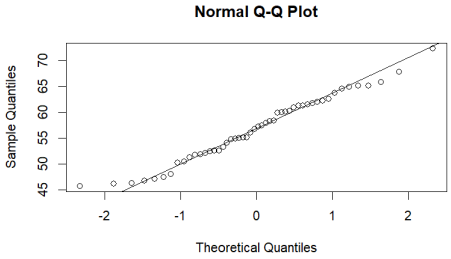
Les acadèmies Serveis d'Anglès SA han començat un pla pilot en dues de les seves seus (A i B) per emprar una aplicació per a l'alumnat entre 10 i 12 anys com a complement a les sessions presencials que realitzen. Després d'un trimestre en funcionament volen fer un estudi per poder valorar-ne el seu ús futur.
La seu de l'acadèmia A ha escollit una mostra aleatòria de 80 alumnes entre 10 i 12 anys i han recollit que 53 d'ells es van descarregar l'aplicació durant els primers quinze dies.
1. [1 punt] Trobeu l'IC de la probabilitat que un estudiant s'hagi descarregat l'aplicació durant els primers quinze dies amb un 95% de confiança i interpreteu-ne el resultat.

2. [1 punt] La seu de l'acadèmia B ha escollit un mostra aleatòria de *nB* alumnes entre 10 i 12 anys i en aquest cas, 78 d'ells es van descarregar l'aplicació durant els primers quinze dies. L'IC amb un 95% de confiança de la diferència de probabilitats entre les dues seus A i B és (-0.3294, -0.0790). Trobeu *nB* i interpreteu-ne el resultat.

Una de les preocupacions de les famílies és que l'alumnat no empri un temps excessiu en l'aplicació i per això es recull el temps (T) en minuts que cada alumne hi està connectat durant una setmana. Per una mostra de 50 alumnes s'han obtingut les següents dades:

$\sum_{i=1}^{50} t_i = 2842.6781$ i $\sum_{i=1}^{50} t_i^2 = 163598.6751$

3. [1 punt] Argumenteu si podeu validar la premissa de normalitat de la variable T



4. [1 punt] A partir de les dades anteriors, doneu una estimació puntual de la mitjana del temps (T) en minuts. Doneu també l'error tipus d'aquesta estimació.

5. [1 punt] Calculeu un interval de confiança al 95% per a la mitjana del temps i interpreta el resultat tenint en compte que no es vol que l'alumnat estigui, de mitjana, més d'una hora a la setmana connectat a l'aplicació.

6. [1 punt] Es vol també estudiar la dispersió del temps de connexió entre l'alumnat. Calcula un interval de confiança al 95% per a la variància i interpreta el resultat per a la desviació dels temps de connexió de l'alumnat a l'aplicació.

Finalment es recull la valoració de l'aplicació de l'alumnat en les dues seus A i B (VA i VB).
7. [0.5 punts] Argumenta les característiques del disseny per a realitzar aquest estudi.

Independentment del que hagis respost a l'apartat 7, considera ara l'estudi amb dades independents.
8.- [0.5 punts] S'ha agafat una mostra a l'atzar de 30 alumnes de la seu A i 30 alumnes de la seu B i s'ha recollit la seva valoració de l'aplicació per estudiar-ne la diferència entre les mitjanes. Indica quines premisses cal tenir en compte per a realitzar l'estudi indicat.

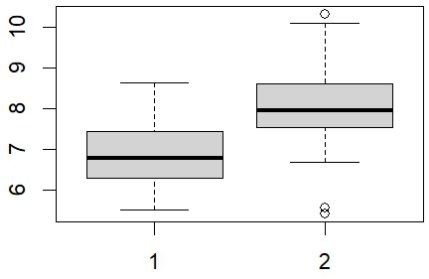
9. [2 punts] Les dades obtingudes són les següents:

Alumnat de la seu A: $\sum_{i=1}^{30} VA_i = 207.0133$ i $\sum_{i=1}^{30} VA_i^2 = 1447.5261$

Alumnat de la seu B: $\sum_{i=1}^{30} VB_i = 240.6079$ i $\sum_{i=1}^{30} VB_i^2 = 1967.0462$

Calculeu un IC de la diferència de mitjanes amb una confiança del 95% i interpreteu-ne el significat.

10. [1 punt] Anomena el gràfic següent i argumenta quina/es premissa/es es poden o no validar de l'estudi anterior. Relaciona les dades del gràfic amb les donades i amb les calculades en l'apartat anterior.



Els valors de la normal són per la distribució normal estandarditzada Z(0,1). Aquests valors els podeu necessitar per als blocs C i D.

qt(0.975,46)= 2.0129	qt(0.975,47)=2.0117	qt(0.975,48)=2.0106	qt(0.975,49)=2.0096	qt(0.975,50)=2.0086
qt(0.975,56)=2.0032	qt(0.975,57)=2.0025	qt(0.975,58)=2.0017	qt(0.975,59)=2.001	qt(0.975,60)=2.0003
qchisq(0.975,46)= 66.6165	qchisq(0.975,47)= 67.8206	qchisq(0.975,48)= 69.0226	qchisq(0.975,49)=70.2224	qchisq(0.975,50)= 71.4202
qchisq(0.025,46)= 29.1601	qchisq(0.025,47)= 29.9562	qchisq(0.025,48)= 30.7545	qchisq(0.025,49)=31.5549	qchisq(0.975,50)= 32.3574
qnorm(0.9)=1.282	qnorm(0.95)=1.645	qnorm(0.975)=1.96	qnorm(0.99)=2.326	qnorm(0.995)=2.576

Nom:

Problema 2 (Bloc D)

Un grup d'estudiants de l'assignatura de PE ha realitzat un experiment per analitzar la relació entre la mida d'un fitxer (X) i el temps de compressió (Y) utilitzant el programari *ZipWins*. Han generat 100 fitxers amb mides entre 1 i 100 MB i han registrat els temps de compressió en segons.

El panell superior de la Figura 1 mostra els gràfics de dispersió de Y vs. X, Y vs. $\log(X)$, $\log(Y)$ vs. X i $\log(Y)$ vs. $\log(X)$. A més, els estudiants han ajustat models de regressió simple i, amb els residus, han generat els gràfics del segon panell de la Figura 1. Els resultats dels models corresponents a Y vs. $\log(X)$ i $\log(Y)$ vs. $\log(X)$ es mostren a la Figura 2.

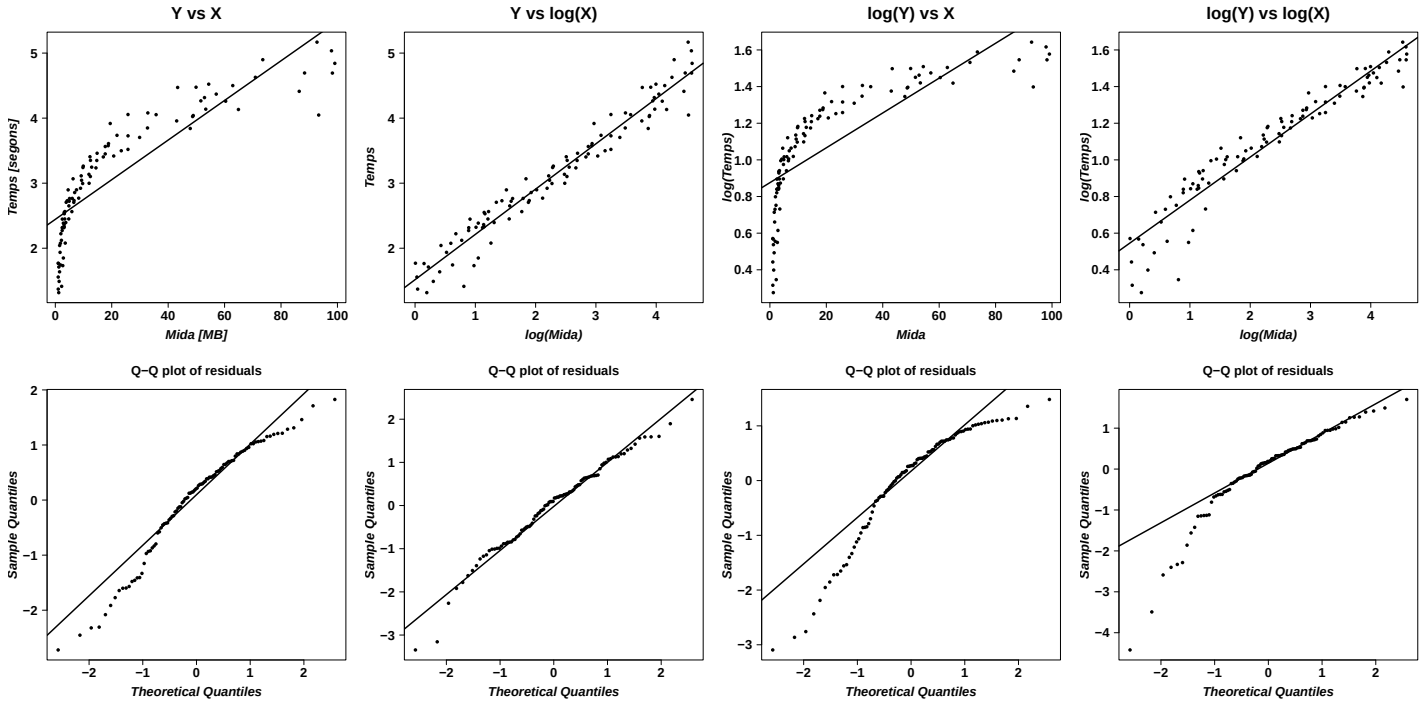


Figura 1: Gràfics de dispersió entre temps de compressió (Y) i mida del fitxer (X).

Model 1: `lm(y ~ log(x))`

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.51633    0.04227   35.87   <2e-16
log(x)       0.69627    0.01578   44.13   <2e-16

Residual standard error: 0.2117 on XXX degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9516
```

Model 2: `lm(log(y) ~ log(x))`

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.545090   0.019387   28.12   <2e-16
log(x)      0.235026   0.007236   32.48   <2e-16

Residual standard error: 0.0971 on XXX degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.9141
```

Figura 2: Models de regressió lineal simple de Temps vs. log(Mida) i log(Temps) vs. log(Mida).

(a) Basant-te en els gràfics dels dos panells de la Figura 1, en quin dels quatre parells de variables sembla que es compleixen les condicions d'un model de regressió simple? Raona la teva resposta. **(1 punt)**

(b) A la Figura 2, quin és el valor d'**XXX** dels dos models? **(0,5 punts)**

(c) Dona una interpretació dels valors `0.69627` (**Model 1**) i `0.235026` (**Model 2**). **(1 punt)**

(d) Quin és l'inconvenient d'ajustar models de regressió amb transformacions logarítmiques? **(0,5 punts)**

(e) Si s'hagués fet l'experiment amb només 25 fitxers d'entre 1 i 100 MB, quins haurien estat els canvis més notables en els models? Òbviament, no podeu dir quins resultats haurien obtingut, però sí descriure la magnitud dels canvis més importants. **(1 punt)**

En una segona fase de l'experiment, els estudiants generen 100 fitxers més amb mides entre 1 i 100 MB, els comprimeixen amb el programari *RareWins* i registren els temps de compressió (en segons). A continuació, ajusten el següent model de regressió lineal, on Z pren el valor 1 en el cas del programari *RareWins* i 0 en el cas de *ZipWins*:

$$Y = \beta_0 + \beta_1 \log(X) + \beta_2 Z + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma).$$

R torna el següent resultat:

```
lm(formula = Y ~ log(X) + Z)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49647    0.03165   47.28   <2e-16
log(X)       0.70467    0.01053   66.90   <2e-16
ZRareWins    0.38913    0.02846   13.67   <2e-16

Residual standard error: 0.2012 on 197 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9591
```

(f) Es tracta de dades emparellades o independents? Raona la teva resposta. **(0,5 punts)**

(g) Dona una interpretació dels valors de $\hat{\beta}_0 = 1.49647$ i $\hat{\beta}_2 = 0.38913$. **(1 punt)**

(h) Calcula l'interval de confiança del 99% per a β_2. Què hi pots concloure? **(1,5 punts)**

(i) Com s'interpreta el valor d'$R^2 = 0.9596$? **(0,5 punts)**

(j) Com canviaria el valor d'R^2 si incloguéssim més variables al model? Raona la teva resposta. **(1 punt)**

(k) Segons el model, quin és el valor esperat del temps de compressió amb *ZipWins* d'un fitxer de 10 MB? **(0,5 punts)**

(l) Quin canvi s'esperaria en el temps de compressió si es duplica la mida d'un fitxer? **(1 punt)**

`qt(0.995, 197)`	2.601	`qt(0.99, 197)`	2.345	`qt(0.99, 196)`	2.346	`qnorm(0.995)`	2.576