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.

P1= 53/80 = 0.6625
n1= 80

OPCIÓ 1: se= $\sqrt{\frac{P1\cdot(1-P1)}{n1}} = \sqrt{\frac{0.6625\cdot 0.3375}{80}}$ = 0.0529

IC(π_1 , 95%) = (P1 – $z_{1-\alpha/2}$ · se , P1 + $z_{1-\alpha/2}$ · se) = (0.6625 – 1.96· 0.0529, 0.6625 + 1.96· 0.0529) = (0.5588, 0.7662)

OPCIÓ 2: se= $\sqrt{\frac{0.5\cdot 0.5}{n1}} =$ = 0.0559

IC(π_1 , 95%) = (P1 – $z_{1-\alpha/2}$ · se , P1 + $z_{1-\alpha/2}$ · se) = (0.6625 – 1.96· 0.0559, 0.6625 + 1.96· 0.0559) = (0.5529, 0.7721)

Amb el 95% de confiança, un estudiant de la seu A s'ha instal·lat l'aplicació en els primers quinze dies amb una probabilitat entre el 0.5588 i 0.7662 (o opció 2: entre 0.5529 i 0.7721).

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.

P1= 53/80 = 0.6625
n1= 80

P2= 78/n2
n2

IC($\pi_1 - \pi_2$, 95%) = ((P1-P2) – $z_{1-\alpha/2}$ · se , (P1-P2) + $z_{1-\alpha/2}$ · se) = (-0.3294, -0.0790)

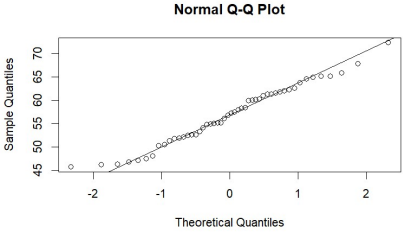
Per tant, el seu punt mig és P1 – P2. El punt mig de l'interval (-0.3294, -0.0790) és (-0.3294 -0.0790)/ 2 = -0.2042

0.6625 – P2 = -0.2042 ; P2= 0.8667 i aleshores n2= 90

Amb el 95% de confiança l'alumnat de la seu B s'ha instal·lat l'aplicació en els primers quinze dies amb una probabilitat superior a la de la seu A, amb increment entre 0.08 i 0.33.

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

A partir de la gràfica (QQplot), no es veu cap problema en validar la premissa de normalitat ja que els quantils de la mostra observats s'ajusten bé amb els quantils teòrics de la normal.

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.

$$\bar{T} = \frac{2842.6781}{50} = 56.8536 \quad \text{i} \quad s_T^2 = \frac{163598.6751 - \frac{2842.6781^2}{50}}{49} = 40.4551, \; s_T = 6.3604$$

I per tant, se = 6.3604/ $\sqrt{50}$ = 0.8995

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. .

$IC(\mu_T, 95\%) = (\bar{x} - t_{n-1,1-\frac{\alpha}{2}} \cdot se \, , \bar{x} + t_{n-1,1-\frac{\alpha}{2}} \cdot se \,) =$

En el nostre cas es té que $t_{n-1,1-\frac{\alpha}{2}} = t_{49,0'975} = 2.0096$

$= (56.8536 - 2.0096 \cdot 0.8995 \, , 56.8536 + 2.0096 \cdot 0.8995 \,) = (55.0460, 58.6612)$

A partir de l'IC calculat amb un 95% és versemblant considerar que la mitjana del temps de connexió a l'aplicació per part de l'alumnat es troba entre els 55 i els 59 minuts, lleugerament per sota d'una hora.

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.

$$IC\,(\partial_T^2, 95\%) = \left(\frac{s_T^2 \cdot (n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{s_T^2 \cdot (n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}} \right) = \left(\frac{40.4551 \cdot 49}{\chi^2_{49,0.975}}, \frac{40.4551 \cdot 49}{\chi^2_{49,0.025}} \right) =$$

(28.2288, 62.8206)

$IC\,(\partial_T\,, 95\%)$ = (5.3131, 7.9259)

La desviació poblacional del temps de connexió de l'alumnat estarà amb un 95% de confiança entre els 5.3 i els 8 minuts.

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.

El paràmetre de l'estudi seria l'esperança. En aquest cas faríem us d'un estudi amb dades independents ja que no es pot realitzar amb dades aparellades perquè cada alumne només ho és d'una seu.

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

Les premisses que cal tenir en compte són:
a) Mostra aleatòria simple: les dades han de provenir d'una mostra aleatòria simple, és a dir, s'ha de garantir que s'han escollit a l'atzar.
b) Les variables aleatòries de les valoracions de la mostra A (VA) i les valoracions de la mostra B (VB) han de seguir un model normal.
c) Les variàncies poblacionals les considerem desconegudes però iguals, és a dir, les variàncies de VA i VB han de complir la premissa d'homoscedasticitat.

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.

Alumnat de la seu A: $\overline{VA} = \frac{207.0133}{30} = 6.9004$ i $s_{VA}^2 = \frac{1447.5261 - \frac{207.0133^2}{30}}{29} = 0.6566$, $s_{VA} = 0.8103$

Alumnat de la seu B: $\overline{VB} = \frac{240.6079}{30} = 8.0203$ i $s_{VB}^2 = \frac{1967.0462 - \frac{240.6079^2}{30}}{29} = 1.2865$, $s_{VB} = 1.1342$

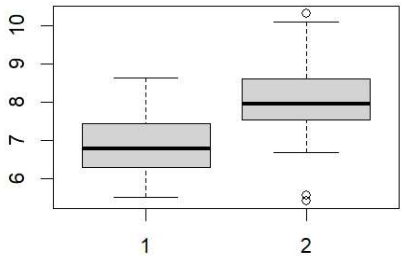
$s^2 = \frac{(n1-1)\cdot sVA^2 + (n2-1)\cdot sVB^2}{(n1+n2-2)} = \frac{29\cdot 0.6566 + 29\cdot 1.2865}{58} = 0.97156$ i $s = 0.9857$

$se = 0.9857 \cdot \sqrt{\frac{1}{30} + \frac{1}{30}} = 0.2545$

I per tant, $IC(\mu_{VB} - \mu_{VA}, 95\%) = (\overline{VA} - \overline{VB} - t_{n1+n2-2,1-\frac{\alpha}{2}} \cdot se, \overline{VA} - \overline{VB} + t_{n1+n2-2,1-\frac{\alpha}{2}} \cdot se) = (0.6105, 1.6293)$

Amb un 95% de confiança, la mitjana de les valoracions dels estudiants de la seu B són superiors a la dels estudiants de la seu A entre 0.6 i 1.6 punts.

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.



En el gràfic es poden observar els boxplots de les valoracions dels estudiants de la seu A (dades 1) i de la seu B (dades B). Això ho podem saber perquè la mitjana de les valoracions dels estudiants de la seu A és 6.9 i la de la seu B, 8.02.

Amb els boxplot es podria tenir una primera informació sobre la normalitat de les dues variables (aquí en ambdós casos, per la bona simetria els boxplots són força coherents amb un model normal), així com l'homocedasticitat de les variàncies (amplitud similar).

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.

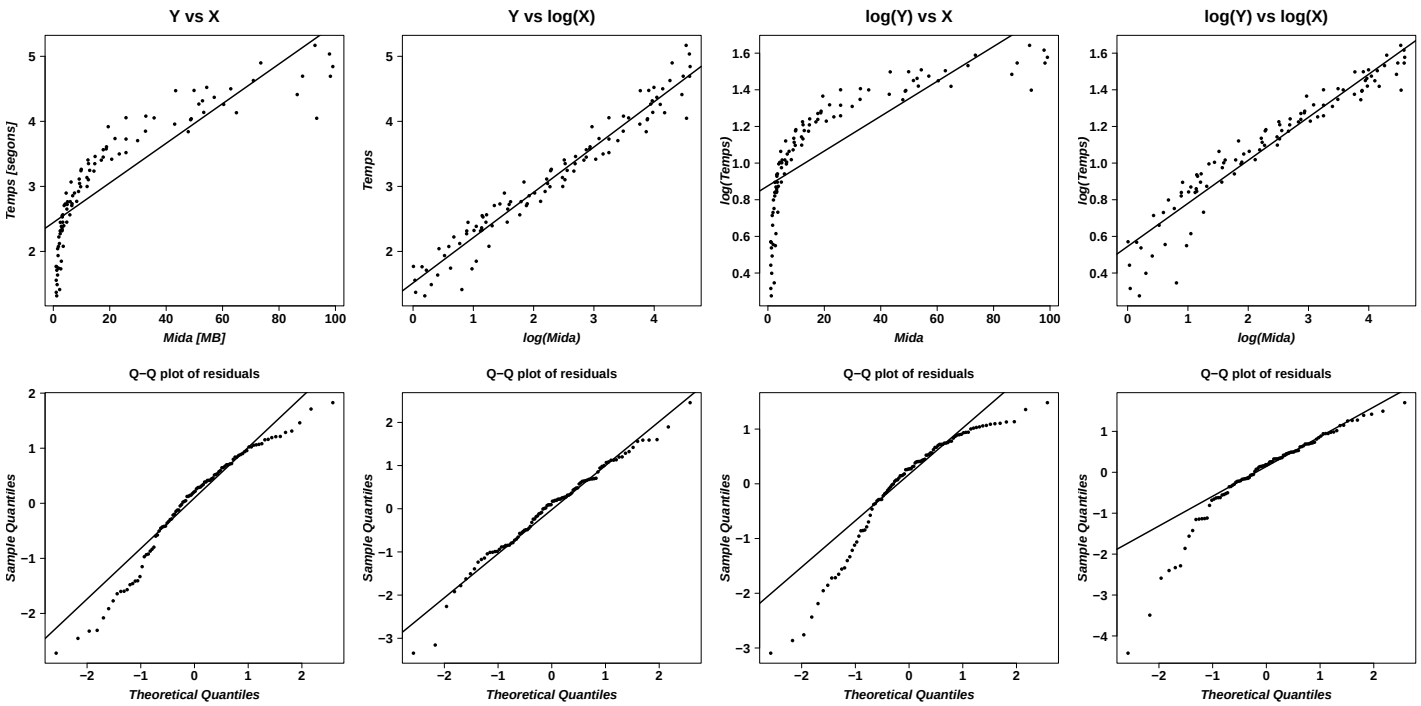


Figura 1: Gràfics de dispersió entre temps de compressió (Y) i mida del fitxer (X).

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)**

Solució

Descartem els models d'Y vs. X i $\log(Y)$ vs. X, perquè les relacions entre les dues parells de variables són clarament no lineal. Sembla més adient el model de regressió lineal simple per a Y vs. $\log(X)$, perquè els residus semblen seguir una distribució normal. A més a més, es pot intuir que es compleix la homocedasticitat al model d'Y vs. $\log(X)$.

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

Solució

Els graus de llibertat són iguals a $n - 2 = 98$.

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

Solució

- **Model 1**: Augmentant el logaritme de la mida del fitxer per 1 (el que equival a multiplicar la mida per $\exp(1) = 2.72$), l'augment esperat del temps de compressió es 0.696 segons.
- **Model 2**: Augmentant el logaritme de la mida del fitxer per 1, l'augment esperat del logaritme del temps de compressió es 0.235. Això equival a un augment **relatiu** pel factor $\exp(0.235) = 1.26$.

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

Solució

La interpretació dels coeficients dels models no és gaire intuitiva. Per obtenir interpretacions intuïtives, cal aplicar transformacions matemàtiques.

(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)**

Solució

Les estimacions puntuals dels paràmetres i l'error residual no hauria de canviar gaire, tampoc l'R^2, però sí els errors estàndards de les estimacions dels paràmetres. Aquests serien més grans.

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)**

Solució

Dades independents, perquè els 100 fitxers comprimits amb *RareWins* son diferents dels 100 fitxers comprimits amb *ZipWins*

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

Solució

Estimacions de:

- ($\hat{\beta}_0 = 1.49647$) Temps (en segons) de compressió amb *ZipWins* esperat en cas d'un fitxer d'1 MB ($\log(1) = 0$).
- ($\hat{\beta}_2 = 0.38913$) Valor esperat de la diferència de temps (en segons) de compressió amb *RareWins* versus *ZipWins* en fitxers de **la mateixa mida**. El valor positiu implica temps més grans (i per tant més lents) amb *RareWins*.

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

Solució

$$\text{IC}(\beta_2; 0.99) = 0.38913 \mp 2.601 \cdot 0.02846 = [0.315,\ 0.463].$$

Podem concloure, amb un nivell de confiança del 99%, que *RareWins* és, en mitjana, més lent que *ZipWins*, perquè tots els valors de l'interval són positius. La diferencia (mitjana) està amb un nivell de confiança del 99% entre 0.315 i 0.462 segons a favor de *ZipWins*.

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

Solució

El 95,96% de la variabilitat del temps de compressió s'explica per la seva relació amb les variables mida de fitxer i programari de compressió.

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

Solució

Afegint més variables al model el valor d'R^2 no baixaria en cap cas i augmentaria si les variables estiguessin relacionats amb la variable resposta, perquè llavors la capacitat predictiva del model milloraria.

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

Solució

$$\widehat{\text{E}}(Y | X = 10, Z = \textit{ZipWins}) = 1.49647 + 0.70467 \cdot \log(10) = 3.12 \text{ (segons)}.$$

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

Solució

$\log(X \cdot 2) = \log(X) + \log(2) = \log(X) + 0.693 \Longrightarrow$ s'esperaria que el temps de compressió augmentés $0.693 \cdot 0.70467 = 0.488$ segons.