Problema 1 (Bloc C)

Totes valen 1 pt menys 1ª (0.5 pts) i 7ª (1.5 pts)

Un grup de recerca en imatges digitals està considerant un possible índex, una escala de 0 a 1000, per mesurar la nitidesa (S, de "Sharpness") d'una imatge, un tema important per diferenciar imatges reals d'altres generades per IA. Un dels passos inicials en la validació de l'índex consisteix en aplicar-ho a imatges reals per veure si es distribueix de forma coherent. Aquests són els índexs en una primera mostra a l'atzar de 5 elements: 75, 200, 250, 400, 600.

Cada cop confondre paràmetre (e.g, σ) i estadístic (S) **– ½**

1) Heu sentit que les 5 imatges s'han obtingut agafant les 5 primeres imatges que han sortit posant a Google "random images". Critiqueu i/o milloreu aquest procediment d'obtenció de la mostra.

No: el sentit de "**random" per a un buscador no és el d'un element a l'atzar + ¼ ,** i potser ni tan sols independents.

Hauríem de disposar d'una definició operativa (quin tipus d'imatge?) per extraure elements amb un **generador de nombre aleatoris +¼**, potser de R o d'una web, de manera reproduïble per altres investigadors.

*Si parla de 'n' petita i poca 'representativitat, mal - **¼ tant** si estava malament, regular o bé.*

*Si només diu no podem garantir ... les dades siguin **reals**, màxim **+¼***

2) Amb la mostra disponible, calculeu una estimació per interval de confiança al 95% del valor mig de S.

Suma: 1525; suma valors al quadrat: 628125 Mitjana: **305**;

Variància mostral: $(628125 - 1525^2/5)/4 =$ **40750**; desviació tipus: **201.8663**

$IC(\mu,0.95) = 305 \pm 2.776 \times 201.8663/\sqrt{5} =$ **(54.39, 555.61) (1)**

3) Amb la mateixa mostra, calculeu una estimació per interval de confiança al 95% de la desviació tipus de S.

$IC(\sigma^2, 0.9) = (4*40750/11.14, 4*40750/0.484) = (14631.96, 336776.86)$ **+ ½**

$IC(\sigma, 0.9) =$ (120.96, 580.32) **+ ½**

Atenció: els valors d'una distribució χ^2 no s'han d'elevar al quadrat (utilitzar 11.14^2, 0.484^2 indica un error de concepte, i no haver fet gaires exercicis)

Una nova mostra més gran (n=100) ha resultat en un IC al 90% de confiança (365, 415) per a la esperança de S.

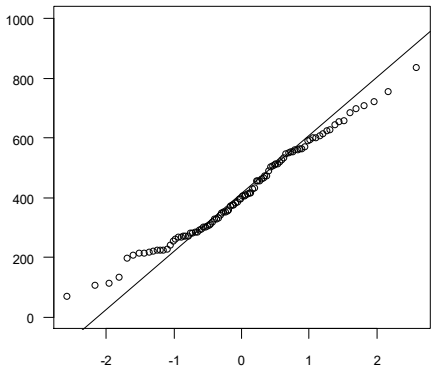
4) Obtingueu d'aquest resultat els valors de la mitjana i la desviació mostrals, i interpreteu l'interval de confiança.

La mitjana mostral és el punt central de l'interval: 390 **+ ¼**

La semi-amplada correspon a $t_{99,0.95} * s/\sqrt{100}$; (aproximem quantil per 1.65).

Igualem a 415-390=25, s = 25 × 10 / 1.65 = 151.5 **+ ¼**

Amb 90% de confiança l'índex **mitjà** es troba entre 365 i 415. **+ ½** (els índexs, mal)



5) Què és la figura de la dreta (obtinguda amb la segona mostra de S), i què ens diu?

El diagrama **Quantil-Quanti**l valora la **proximitat** entre distribució empírica de la mostra i la distribució **Normal** teòrica. **+ ½**

Bona aproximació, encara que els extrems no s'allunyen tant com la Normal (les cues són curtes). **+ ½**

Els investigadors d'aquest grup volen mostrar que les imatges generades per una IA tenen propietats diferents però que no són fàcils de percebre a simple vista, com la nitidesa S. Creuen que les imatges artificials són més nítides i molt més homogènies (menor dispersió). Per provar-ho, seleccionen aleatòriament 100 imatges creades amb una IA i apliquen el mateix índex que a les imatges reals, per analitzar els resultats anteriors amb els de les imatges IA.

6) Es tracta d'un estudi amb dades aparellades o amb dades independents? Raoneu la resposta.

Dades independents, perquè no s'explica cap relació entre les imatges reals i les imatges IA. **(1)**

En qualsevol cas, ara suposeu que es tracta de dues mostres independents. La segona mostra presenta un índex mitjà igual a 489 punts i una desviació tipus de 111 punts (si no heu resolt l'apartat 4, utilitzeu a partir d'ara els valors 400 i 125 com a mitjana i desviació tipus de la primera mostra).

7) Es demana un IC al 95% de confiança per a la diferència de mitjanes de S entre imatges IA i imatges reals. Comenteu sobre les premisses necessàries, i si creieu que són assumibles.

Variància comuna: (99 151.5^2 + 99 111^2)/198 = **17638.93 (S = 132.81)** + ¼ [*118.21 si no heu resolt l'apartat 4*]

Error tipus de la diferència de mitjanes mostrals: S √(1/100 + 1/100) = **18.78 24** + ¼ [*16.717*]

IC: 489 – 390 ± 1.96 × 18.7824 = (**62.186, 135.813**) + ½ [*56.235, 121.765*]

Premisses: 1) **dues** mostres **aleatòries** simples **independents**; 2) mateixa **variància** poblacional; 3) **Normalitat** (no preocupant per a mostres d'aquesta mida) + ½

8) Es pot trobar l'error tipus de $\mu_{IA} - \mu_{real}$? I el de $\hat{\mu}_{IA} - \hat{\mu}_{real}$? Justifiqueu les respostes i expliqueu de què ens informa.

El primer no té error tipus (μ és constant: l'error tipus aplica a estimadors no a paràmetres). + ½

$\hat{\mu}_{IA} - \hat{\mu}_{real}$ sí es refereix a estimadors. + ¼

18.78 [*16.72*] calculat abans estima la fluctuació "habitual" de la diferència de mitjanes mostrals. + ¼

9) Avalueu la suposada major homogeneïtat de les imatges artificials respecte les reals i interpreteu.

Farem un interval de confiança al 95% per al rati entre ambdues variàncies $\sigma^2_{IA}/\sigma^2_{real}$:

$$\left[\frac{s^2_{IA}/s^2_{real}}{F_{0.975,99,99}}, \frac{s^2_{IA}/s^2_{real}}{F_{0.025,99,99}}\right] = \left[\frac{0.5367}{1.4862}, \frac{0.5367}{1/1.4862}\right] = [0.3611, 0.7976] \qquad \left[\frac{0.7885}{1.4862}, \frac{0.7885}{1/1.4862}\right] = [0.531, 1.172]$$

Amb 95% confiança, la variància de imatges artificials és com a molt un 80% de la variància per a imatges reals

[Com la variabilitat és menor les imatges artificials tenen valors més semblants.]

[*I si prenem el valor suposat 125: com el valor 1 hi és a dins de l'interval, la homogeneïtat podria ser la mateixa*] **1**

Si interpreta bé els estimador mostrals S, màxim ½

10) L'índex pot ser emprat per classificar una imatge com a nítida o no nítida, segons algun llindar crític. Es vol determinar <u>quantes imatges reals i quantes artificials</u> s'han d'utilitzar a un estudi per trobar un IC 95% d'amplada 15% per la diferència de proporcions (o diferència de probabilitats) d'imatges nítides entre les dues tipologies. Tingueu en compte que, per determinades raons, en aquest estudi el nombre d'imatges generades per IA <u>no pot ser més de la meitat</u> que el nombre d'imatges reals.

La semi-amplada del IC per a $\pi_{IA} - \pi_{real}$ correspon a $z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_{IA}(1-\hat{\pi}_{IA})}{n_{IA}} + \frac{\hat{\pi}_{real}(1-\hat{\pi}_{real})}{n_{real}}}$, on $n_{IA} = n$ i $n_{real} = 2n$,.

[Com no disposem de més informació, escollim el criteri conservador ($\hat{\pi}_{IA} = \hat{\pi}_{real} = 0.5$) de màxima incertesa.]
Llavors:

$$1.96\sqrt{\frac{0.5(0.5)}{n} + \frac{0.5(0.5)}{2n}} = 1.96\sqrt{\frac{2\,0.5^2}{2n} + \frac{0.5^2}{2n}} = 1.96\,0.5\sqrt{\frac{3}{2n}} = 0.075 \rightarrow \sqrt{n} = 16 \rightarrow n = 256$$

Necessitem 256 imatges generades per IA i 512 reals. **1**

qnorm(0.900) = 1.282	qnorm(0.975) = 1.960	qt(0.95,3)=2.353	qt(0.975,3)=3.182	qchisq(0.025,3)=0.216	qchisq(0.05,3)=0.352
qnorm(0.925) = 1.440	qnorm(0.990) = 2.326	qt(0.95,4)=2.132	qt(0.975,4)=2.776	qchisq(0.025,4)=0.484	qchisq(0.05,4)=0.711
qnorm(0.950) = 1.645	qnorm(0.995) = 2.576	qt(0.95,5)=2.015	qt(0.975,5)=2.571	qchisq(0.025,5)=0.831	qchisq(0.05,5)=1.145

qchisq(0.95,3)=7.815	qchisq(0.975,3)=9.348	qf(0.95,99,99)=1.3941	qf(0.975,99,99)=1.4862	qf(0.95,100,100)=1.3917
qchisq(0.95,4)=9.488	qchisq(0.975,4)=11.14	qf(0.95,99,100)=1.3927	qf(0.975,99,100)=1.4844	qf(0.975,100,100)=1.4833
qchisq(0.95,5)=11.07	qchisq(0.975,5)=12.83	qf(0.95,100,99)=1.3931	qf(0.975,100,99)=1.4850	

Fins a +0.5 si solució bona, clara, fàcil de llegir i corregir (un 5% aproximadament de solucions)

PE T23 Q1 2023/24. Parcial 2. 9 de gener 2024

NOM:_____COGNOM:_____

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

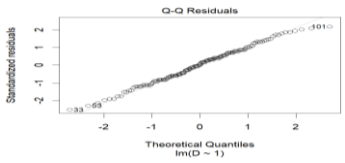
Problema 2 (Bloc D)

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

Per fer un estudi sobre els factors que poden predir el rendiment dels alumnes de PE, s'han recollit les següents variables corresponents a 150 alumnes: hores d'estudi setmanals, si va emprar e-status durant l'assignatura (sí/no) i l'edat a l'inici del curs. A més, s'han recollit les notes dels 2 exàmens parcials (y_P1 i y_P2) que es consideren les variables que mesuren el rendiment. Finalment, es té una darrera variable calculada que és la diferència (D= y_P2 - y_P1) de notes entre els exàmens parcials i la diferència de logaritmes (LD = log(y_P2) – log(y_P1)). S'han ajustat amb R els següents 4 models:

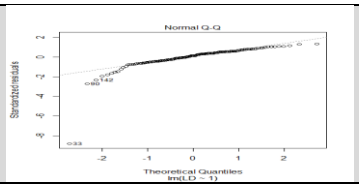
Model 1

```
Call: lm(formula = D ~ 1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2093     0.1067  -11.34   <2e-16 ***
---
Residual standard error: [illegible] on 149 degrees of freedom
```



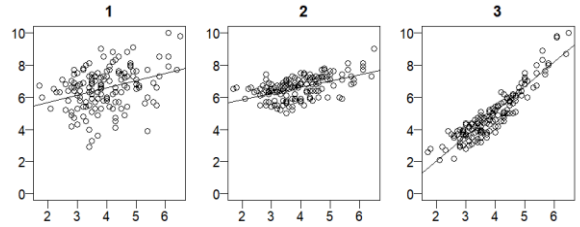
Model 2

```
Call: lm(formula = LD ~ 1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24500    0.03302  -7.419  8.3e-12 ***
---
Residual standard error:0.4044 on 149 degrees of freedom
```



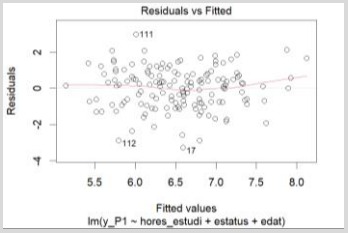
Model 3

```
Call: lm(formula = y_P1 ~ hores_estudi)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.7539     0.4066   11.69  < 2e-16 ***
hores_estudi   0.4554     0.1012    4.50 1.37e-05 ***
---
Residual standard error: 1.185 on 148 degrees of freedom
Multiple R-squared: 0.1204, Adjusted R-squared: 0.1144
F-statistic: 20.25 on 1 and 148 DF,  p-value: 1.365e-05
```



Model 4

```
Call: lm(formula = y_P1 ~ hores_estudi + estatus + edat)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.70745    1.27938   3.679 0.000328 ***
hores_estudi   0.49170    0.09688   5.075 1.16e-06 ***
estatusSi      0.78500    0.19526   4.020 9.28e-05 ***
edat          -0.03132    0.06158  -0.509 0.611848
Residual standard error: 1.129 on 146 degrees of freedom
Multiple R-squared:  0.2114,Adjusted R-squared:  0.1952
F-statistic: 13.05 on 3 and 146 DF,  p-value: 1.344e-07
```



1. Emprant els models disponibles, quines són les estimacions puntual per a la <u>diferència</u> ($\mu_{y\_P2} - \mu_{y\_P1}$) i pel <u>rati</u> (μ_{y_P2} / μ_{y_P1}) de mitjanes del segon parcial respecte el primer? Interpreta-les.

   D= y_P2 - y_P1 (Model 1, càlculs per a D) → $\hat{\mu}_{y\_P2} - \hat{\mu}_{y\_P1} = \hat{\mu}_D = \bar{y}_D = \mathbf{-1.2093}$

   Podem esperar una **diferència mitjana** de -1.2 punts en el parcial 2 respecte el parcial 1

   R= y_P2 / y_P1 (Model 2, càlculs per a LD=log(R)) → $\hat{\mu}_{\log(R)} = -0.245$ → $\hat{\mu}_R = \exp(-0.245) = \mathbf{0.783}$

   Podem esperar en **mitjana un rati de 0.8** entre les nota del 2n parcial respecte el 1r. Les notes del 2n parcial s'esperen que siguin un 80% de les del 1r en mitjana (aproximadament un 20% inferiors)

2. Emprant els models disponibles, calcula un interval de confiança (IC) del 95% per a la diferència de mitjanes de notes del segon parcial respecte el primer.

$$IC(\mu_D, 95\%) = \bar{y}_D \pm t_{149,0.975} \cdot se = \bar{y}_D \pm qt(0.975,149) \cdot se = -1.2093 \pm 1.976 \cdot 0.1067 = [\mathbf{-1.42}, \mathbf{-1.0}]$$

3. Calcula la desviació estàndard residual del model 1.

   És la pròpia desviació de D, que podem obtenir a partir del se

$$se = \frac{s_D}{\sqrt{n}} = 0.1067 \frac{s_D}{\sqrt{150}} \quad \rightarrow \quad s_D = se \cdot \sqrt{n} = 0.1067 \cdot \sqrt{150} = \mathbf{1.307}$$

> qt(0.90,148) = 1.287298	> qt(0.95,148) = 1.655215	> qt(0.975,148) = 1.976122
> qt(0.90,149) = 1.287259	> qt(0.95,149) = 1.655145	> qt(0.975,149) = 1.976013
> qt(0.90,150) = 1.287221	> qt(0.95,150) = 1.655076	> qt(0.975,150) = 1.975905
> qnorm(0.90) = 1.281552	> qnorm(0.95) = 1.644854	> qnorm(0.975) = 1.959964

4. Digues quina o quines premisses s'avaluen en els gràfics associats als models 1 i 2 i, basant-te en aquests gràfics, quin model sembla més idoni.
   Aquests gràfics són Normal QQ plots i avaluen la premissa de **Normalitat** dels residus (en aquest cas, equivalent a la Normalitat de la variable resposta). Es veu com el **primer model compleix aquesta premissa** (s'ajusten bé els quantils teòrics del model normal i els de les dades), mentre el **segon model no la compleix** (principalment pels primers valors de la part esquerra). Per tant, **el model 1 és més idoni**.

5. Segons el model 3, indica quina és la recta estimada pel model i la seva interpretació.

   Y_p1 = $b_0 + b_1$*hores_estudi = **4.7539 + 0.4554*hores_estudi**

   L'ordenada a l'origen és 4.75 punts, que representaria la nota esperada per un alumne amb zero hores d'estudi.

   La pendent val 0.46 i representa que per a cada hora més d'estudi la nota puja 0.46 punts

6. Segons el model 3, calcula un IC90% pel canvi esperat en la nota del primer parcial si incrementem en 2 hores el temps d'estudi setmanal, i interpreta'l.
   (no és IC de la predicció per un increment de 2h, sinó del canvi esperat. És el doble del canvi per increment de 1 (pendent))
   Tenim $IC(\beta_1, 90\%) = \left(\hat{\beta}_1 \pm t_{148,0.95} \cdot se_{\hat{\beta}_1}\right) = \left(b_1 \pm t_{148,0.95} \cdot se_{b_1}\right) = (0.4554 \pm 1.655 \cdot 0.1012) = [0.288, 0.623]$
   llavors per a 2 hores l'IC serà $2 * [0.288, 0.623] = \mathbf{[0.576, 1.246]}$

   o bé $IC\left(\mu_{P1,h} - \mu_{P1,h-2}, 90\%\right) = 2 \cdot \left(\hat{\beta}_1 \pm t_{148,0.95} \cdot se_{\hat{\beta}_1}\right) = 2 \cdot (0.4554 \pm 1.655 \cdot 0.1012) = \mathbf{[0.576, 1.246]}$

   Un augment de 2 hores en el temps setmanal d'estudi s'associa amb un increment de la nota mitjana d'entre 0.576 i 1.246 punts en mitjana.

7. Els 3 gràfics associats al model 3 representen les notes del parcial 1 (eix vertical) en funció de les hores d'estudi (eix horitzontal). Basant-te en la informació del model 3, digues quin del gràfics (1, 2 o 3) representa les dades reals. Argumenta la resposta.
   El **gràfic 3 queda descartat** perquè ni el pendent ni el terme constant es corresponen amb el model: p.ex, la *intercept* del model és 4.75 i en el gràfic 3, es veu que el punt on la recta talla a l'eix x=0 ha de ser menor que 2.
   Per discriminar entre el gràfic 1 i 2, ens hem de fixar en la desviació residual (1.18). Aquest és un valor que representa la desviació estàndard dels residus. S'observa que en el 2n gràfic el 100% (o gairebé el 100%) dels residus estan per sota d'aquest valor. **Per tant, el gràfic correcte és el 1.**

8. Emprant el model 4, fes una predicció puntual de la nota del primer parcial per un estudiant de 20 anys que fa servir e-status i que estudia 5 hores a la setmana.

$$\hat{y}_4 = 4.70745 + 0.49170 * 5 + 0.78500 - 0.03132 * 20 = \mathbf{7.32}$$

9. Interpreta el R^2 del model 4.
   En el model 4, **R^2 és 0.2114**, i és un valor **força baix** representant la **variabilitat de la resposta (nota del parcial 1) explicada** per les variables explicatives del model (hores d'estudi, estatusSi i edat). Per tant la **capacitat predictiva és baixa**
   Les **hores d'estudi**, el fet de saber **si s'empra e-status** i l'**edat** expliquen **només el 21.14% de la variabilitat de la nota** del parcial 1.

10. Digues quina o quines premisses s'avaluen al gràfic associat al model 4 i argumenta segons el que es veu si aquesta o aquestes premisses es compleixen i el perquè.
    En el gràfic del model 4 es poden avaluar les premisses de **linealitat** i **homoscedasticitat**. Ambdues premisses es **compleixen raonablement**: la tendència de la línia vermella és pràcticament horitzontal al voltant del 0 i no posa en qüestió la premissa de linealitat. Per altre banda, no es veu cap tendència creixent o decreixent en la variabilitat dels residus al llarg dels valors predits, per tant, es validaria la premissa de variància constant (homoscedasticitat)