NOM:_____ COGNOMS:_____

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

Problema 1 (Bloc C)

Part 1. (4 pts) Izan és administrador del sistema informàtic d´un gimnàs (i també tècnic), amb accés als registres dels usuaris. Per veure si hi veu diferències notables, escull a l'atzar alguns dels seus propis usuaris, i altres del seu col·lega Eidan, que treballa al mateix gimnàs amb usuaris diferents. En aquesta taula veiem el consum calòric (Kcal) diari d'aquests usuaris, deduïble per la seva activitat a les màquines d'exercicis:

Izan (1)	Izan (2)	Izan (3)	Izan (4)	Izan (5)	Eidan (1)	Eidan (2)	Eidan (3)	Eidan (4)	Eidan (5)
880	770	530	770	850	670	890	770	470	730

Comproveu que:
$\sum x_i = 7330$
$\sum x_i^2 = 5551300$

a. Descriviu pas a pas una anàlisi per determinar la diferència mitjana de consum calòric entre els usuaris dels dos tècnics. Per això obtingueu un interval de confiança al 90%.
b. Escriviu la interpretació del resultat anterior.
c. Comenteu sobre les premisses que heu hagut d'assumir per a un ús correcte de l'anàlisi realitzada.
d. Clarament, les mostres són petites. Proposeu amb un criteri concret (per exemple, l'amplitud de l'IC no hauria de ser més gran que 100), i amb aproximacions justificades, una mida de mostra raonable per a un nou estudi.

a) Suma Izan: 3800. Suma Eidan : 3530. Mitjanes: 760 i 706 [Kcal]

Suma quadrats: 2963600 i 2587700. Variàncies: 18900 i 23880. Desv . tipus : 137.4773 i 154.5316 [Kcal]

Variància pooled : 21390 (146.25^2). Error tipus: $146.25\sqrt{\frac{1}{5}+\frac{1}{5}} = 92.5$ [Kcal]

IC($\mu_{Izan} - \mu_{Eidan}$, 90%) = 760 - 706 ± $t_{8,\,0.95}$ 92.5 = 54 ± 1.8595 92.5 = (-118, 226) [Kcal]

b) Amb les dades disponibles, tenim evidències que la diferència entre consums calòrics mitjans dels dos grups és entre 118 Kcal menys i 226 Kcal més per als usuaris de Izan, amb confiança del 90%. Això no implica que les dues mitjanes siguin iguals, simplement que no esperem que es diferenciïn més enllà d'aquests dos límits.

c) En primer lloc, que siguin dues mostres independents (sembla que sí, perquè els grups són diferents). En segon lloc, que cadascuna sigui una m.a.s. (ens han dit que els ha escollit a l'atzar entre tots, i si és administrador tindrà accés a tota la informació, per la qual cosa sembla possible). En tercer lloc, que les variàncies siguin la mateixa en els dos grups (hi ha una petita diferència a les mostrals, però l'atzar podria explicar-ho). En quart lloc, la variable hauria de ser Normal (molt poques dades per verificar-ho).

d) Amb una *n* més gran, el quantil que intervé a l'IC al 90% no serà aquest (1.8595) sinó una mica menor, possiblement proper al quantil d'una N(0,1), que és 1.65. Suposem que la desviació pooled no serà gaire diferent de l'anterior (en prendrem 150, per exemple). Si l'amplitud de l'IC és 100, vol dir que la meitat: 50 = S*$t_{2n-2,\,0.95}\sqrt{2/n}$= 150*1.65 $\sqrt{2/n}$, segons les nostres aproximacions. Aïllant *n* d'aquesta equació obtenim *n* =49.005. Per tant, cal prendre unes 50 observacions de cada grup.

Part 2. (3.5 pts) Una enquesta que es realitza regularment planteja una qüestió relativa a la preocupació del ciutadà per una pujada immediata dels preus a la cistella de la compra (focalitzada en la proporció dels que *sí estan preocupats*). La taula presenta resultats de les dues darreres enquestes :

mes	número participants	núm . resp. afirmatives	IC(π, 95%)	error tipus P	IC($\pi_{abril} - \pi_{gener}$, 95%)
gener			0.42 – 0.56		
abril			0.54 – 0.65		

a) completeu la taula, justificant els passos necessaris.

mes	número participants	núm . resp. afirmatives	IC(π, 95%)	error tipus P	IC($\pi_{abril} - \pi_{gener}$, 95%)
gener	196	96	0.42 – 0.56	0.0357	0.016 – 0.194
abril	306	182	0.54 – 0.65	0.0281	

Proporcions puntuals (els centres dels ICs): 0.49 i 0.595; Radis dels Ics = Extr. Sup. IC - P = $Z_{0.975} \cdot ET_P$:
Gener: 0.56 – 0.49 = 0.07 = 1.96 ET -> ET = 0.07/1.96 = 0.0357
Abril: 0.65 – 0.595 = 0.055 = 1.96 ET -> ET = 0.055/1.96 = 0.0281

$ET_P = \sqrt{\frac{P(1-P)}{N}}$, i d'aquesta expressió obtenim el nombre de participants de cada grup (196 i 306), i el numerador de la proporció (96 i 182)

Finalment, l'error tipus de la diferencia de dues proporcions de mostres independents: el seu quadrat és la suma dels dos errors tipus al quadrat, perquè operem amb les variàncies.

$ET_{diff}^2 = ET_g^2 + ET_a^2$; 0.0357^2 + 0.0281^2 = 0.0454^2

IC = $P_a - P_g \pm 1.96$ 0.0454 = 0.105 ± 0.089 = (0.016, 0.194)

b) comenteu raonadament (justifiqueu les opcions triades):

	D'acord	En desacord	Parcialment
només cal triar a l'atzar els enquestats del gener		cal triar a l'atzar els de gener i els d'abril, en dos processos aleatoris no relacionats perquè siguin dues mostres independents	
és important comprovar que la variabilitat de la resposta és la mateixa en els dos mesos		No té sentit, la resposta és dicotòmica, no hi ha un paràmetre de variància, la seva variabilitat depèn directament de la probabilitat duna resposta positiva.	
hi ha motius per creure que la preocupació del ciutadà ha augmentat	sí, perquè l'IC per a la variació en la proporció de "preocupats" està per sobre de 0, així que amb una confiança del 95% sí que podem creure en un augment.		

Part 3. (2.5 pts) Suposem que es mesura una certa variable resposta I en dues mostres independents de mida n=21 cadascuna, per a dos grups A i B respectivament. Amb les variàncies mostrals obtingudes es construeix un interval de confiança 90% per al quocient de variàncies poblacionals A sobre B. Però els nostres resultats s'han perdut, i només conservem un record incert. Si recordéssim que l'extrem esquerre de l'interval era 1.2:

a) deduir l'extrem superior d'aquest IC(90%);

b) hem recuperat la desviació mostral al grup B, que era 30: determinar la desviació al grup A;

c) si l'extrem esquerre de l'interval estigués just a 1, justificar com es trobaria el grau de confiança corresponent a tal IC (no cal calcular-ho, només desenvolupar el procediment per resoldre la qüestió; es poden fer servir funcions de R en el procediment, sempre que s'especifiquin tots els paràmetres). *En aquest apartat no aplica l'apartat anterior*.

a. Veiem a la taula que el quantil que necessitem és $F_{20,\,20,\,0.95}$ (serveix per als dos extrems): $F_{20,\,20,\,0.95}$ = 2.124, $F_{20,\,20,\,0.05}$ = 1/2.124. Sigui X = s^2_A / s^2_B, aleshores X/2.124 = 1.2, d'on X = 1.2*2.124 = 2.5488. L'extrem superior és X/$F_{20,\,20,\,0.05}$ = X*$F_{20,\,20,\,0.95}$ = 2.5488*2.124 = **5.41**

b. $s_A = s_B \sqrt{X}$ = 30*1.596 = **47.89**

c. Sabem que [s^2_A/s^2_B] / $F_{20,20,1-\alpha/2}$ = 1 , és a dir, que el quantil de nivell 1- ∝/2 buscat val X = 2.5488. Podem trobar ∝ mitjançant:

- `a = 2*(1- pf( 2.5488, 20, 20))` que seria 0.0423 (tot i que això no cal dir-ho); i el grau de confiança llavors seria `1-a` (95.77%).
- O podem expressar 1- ∝ = 1 - 2 (1- P(F < 2.5488)) = 1 - 2 P(F > 2.5488), on F ~ $F_{20,20}$

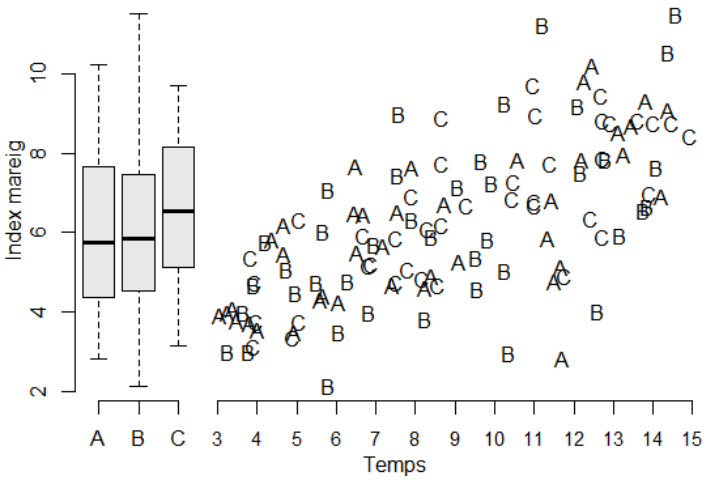
NOM:_____ COGNOMS:_____ *Parcial 2 PE* 2/6/26

Bloc D.

Les dades que mostrem són d'un estudi sobre el mareig que produeix l'ús d'ulleres de realitat virtual. Els participants fan una activitat determinada durant un temps, i al finalitzar es mesura el grau de mareig que presenten. Els participants utilitzen un de tres tipus d'ulleres diferents (A, B i C), que els hi han assignat a l'atzar, 40 a cada grup.

```
Coefficients:
            Estimate Std. Error t value
(Intercept)  6.02682    0.31915  18.884
xB           0.08515    0.45135   0.189
xC           0.53541    0.45135   1.186

Residual standard error: 2.018 on 117 deg of fr
Multiple R-squared:  0.0137,
F-statistic: 0.8127 on 2 and 117 DF,
```



El quadre superior resumeix un model on només s'utilitza la variable resposta i el tipus d'ulleres.

1. Quants paràmetres té el model estadístic, i quins són? Com s'expressa matemàticament el model en aquest cas?

$$Y_i = \mu + \theta_j + \varepsilon_i$$

μ és la mitjana global de la resposta *Y*, el grau de mareig; ϑ_j indica el canvi entre μ i la mitjana en el grup (*j*) al que pertany el participant *i*. Com a R no tenim un paràmetre per al primer grup, el de referència, μ és la mitjana per aquest grup de referència. Finalment tenim V(ε), la variància del terme aleatori o residual. ☠ Poca gent recorda mencionar V(ε).

2. Expliqueu què representa el "Estimate" de la fila "Intercept".

Estimem puntualment un grau de mareig mitjà igual a 6.03 punts per als individus que han utilitzat ulleres del tipus A.

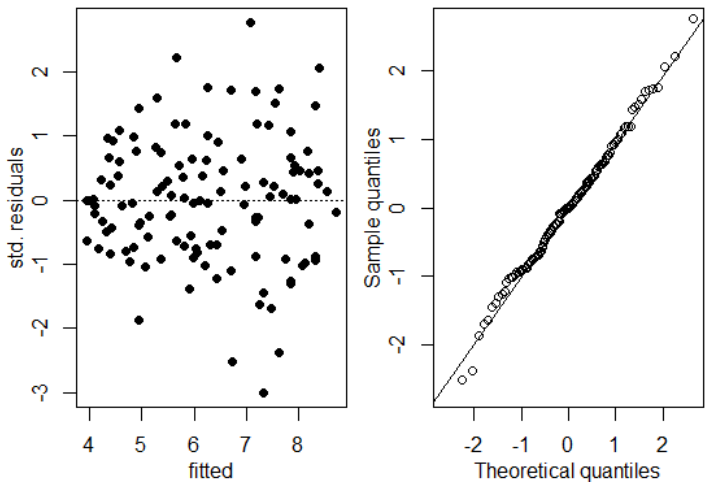
3. Feu el mateix pels valors de les files "xB" i "xC".

Estimem puntualment que els individus assignats a ulleres tipus B experimenten 0.085 punts més de mareig que els A, i els C 0.535 punts més que els A.

4. Els valors de la columna "Std. Error" són les desviacions tipus dels valors que veiem als boxplots de la figura? Per què sí o per què no?

No, els "Std. Error" són els error tipus de les estimacions anteriors, no les variacions individuals en el grau de mareig que tenen amplituds clarament majors.

5. Les premisses del model es podrien acceptar? Quines són, què volen dir, i justifiqueu la resposta.
   - Grups independents: sí, perquè el tipus d'ulleres s'ha assignat a l'atzar
   - Cada grup és una m.a.s.: no se sap com s'han inclòs els participants en l'estudi (podríem esperar que són independents entre sí)
   - Homoscedasticitat: acceptable, els tres boxplots tenen variabilitats semblants
   - Normalitat: no ho podem valorar però la simetria dels boxplots és un bon indicador



Les dades d'aquests dos gràfics provenen dels residus del segon model

☠ És un error freqüent mencionar també *Linealitat* com a premissa. Aquest no és un model lineal, perquè no hi ha variables explicatives numèriques. Tampoc hi ha termes que es puguin interpretar com a *pendents*.

Un altre model incorpora el temps de l'activitat (z), mesurat en minuts. A la dreta teniu un resum de la sortida en R.

```
Coefficients:
            Estimate Std. Error t value
(Intercept)  2.72652    0.41444   6.579
xB          -0.06178    0.33677  -0.183
xC           0.13009    0.33900   0.384
z            0.39345    0.04046   9.725

Residual standard error: 1.505 on 116 deg of fr
Multiple R-squared:  0.4567,
F-statistic:  32.5 on 3 and 116 DF,
```

6. Digueu amb precisió com s'interpreta el valor 2.72652.

És el valor mitjà del grau de mareig que esperem per a un participant del grup de referència (A), amb un temps 0 d'activitat.

7. Feu el mateix amb el valor 0.39345, i calculeu un interval de confiança al 95% del paràmetre associat.

És el pendent comú per a qualsevol tipus d'ulleres: esperem un increment de 0.39 punts en l'escala de mareig per cada minut addicional d'activitat. L'IC(β, 95%) serà: $0.39345 \pm 1.98 \cdot 0.04046 = (0.313, 0.474)$

8. Comenteu el canvi que s'observa al `Multiple R-squared` entre els dos models.

R^2 (model 1): 1.37%; R^2 (model 2): 45.7%. El primer és insignificant, representa que només el tipus d'ulleres no explica pràcticament res la variabilitat de la resposta, per això no s'observen diferències rellevants en les mitjanes. El segon ja és important, és a dir, la combinació del tipus d'ulleres i el temps d'activitat sí expliquen gairebé la meitat de la variabilitat de la resposta, i segurament aquesta contribució es exclusivament del temps. Podríem dir que el temps ens ajuda a predir amb més precisió el grau de mareig del participant.

9. Repetiu amb aquest resultat la valoració de les premisses.

Grups independents i m.a.s.: igual que en l'estudi anterior;

linealitat: correcta, al núvol original o a la gràfica de l'esquerra, a sota, no es veu cap signe de convexitat;

homocedasticitat: ara no ho podem assegurar, ja que el gràfic *std. residuals vs fitted* mostra un patró en el que la dispersió dels residus a l'esquerra és menor que a la dreta (sembla que va creixent: amb més temps i més mareig, més variabilitat);
normalitat: al qqplot, els residus segueixen molt be la referència recta, per tant no es pot posar en dubte.

10. El protocol d'actuació dels autors de l'estudi podria haver estat: a) el temps de l'activitat es decideix prèviament, independentment de qualsevol altre factor; b) el temps de l'activitat es recull quan el participant termina per qualsevol raó. Exposeu breument les conseqüències que es desprenen de seguir un o altre protocol.

Amb un protocol com a), la variable Temps seria una variable experimental, decidida pels investigadors per tal d'estimar l'efecte d'un temps d'activitat en la resposta (o d'estudiar la *relació causa-efecte*). Amb el protocol b), el Temps seria una resposta més, que en aquest model se incorporaria com a covariable per a mesurar el *grau d'associació* que pot tenir amb el grau de mareig del participant, per tant si s'obté abans que la resposta pot servir per a predir-la.