1. Un experiment empíric recull dades (x_i) de la quantitat d'energia que capta un panell solar. Concretament, quanta energia produeix un panell en el temps de 60 minuts. Digueu quins dels següents elements corresponen a estadístics i quins no, justificant breument la resposta:
   - E(X), el valor esperat de la energia que el panell produeix en una hora [*wh: watts-hora, unitat d'energia*]No és un estadístic, no depèn de la mostra, en realitat és un paràmetre.
   - *n*, el nombre de mesures que prenem per a l'experiment No és un estadístic, es un valor constant determinat per l'experimentador
   - x_{min}, la quantitat més petita d'energia que s'ha observat durant la recollida de dades Sí és un estadístic, és el valor mínim de la mostra
   - $\sum x_i$, l'energia total produïda Sí és un estadístic, és la suma de tots els valors de la mostra
   - P, la potència nominal declarada pel fabricant [*w: watts, unitat de potència*] No és un estadístic, és una constant, potser coincideix con la potència esperada si està ben calibrat.

   Consideracions generals: confusió respecte a què és un estadístic. Sovint considerat com quelcom que pot servir per a alguna cosa ("l'energia total serveix per a trobar la mitjana", però "el valor mínim no és útil per a res"). Potser es confon amb "estimador".

2. Expliqueu el concepte d'error tipus (*standard error*) basant-se en dos casos concrets: l'error tipus d'una mitjana i l'error tipus d'una proporció. Utilitzeu un exemple per a cada cas amb nombres concrets inventats per vosaltres.

   L'error tipus d'una mitjana $\bar{x}$ val $\sigma/\sqrt{n}$ i significa la desviació "habitual" del valor que pren la mitjana mostral respecte la verdadera mitjana μ, per tant és la desviació tipus que pot presentar una mesura de la mitjana. Normalment ve estimada utilitzant la desviació tipus mostral: $s/\sqrt{n}$. Per exemple, si μ val 100 i σ val 10, amb una mostra de mida 25 la mitjana mostral té un error tipus de 10/5=2 unitats.

   L'error tipus d'una proporció π també significa la desviació "habitual" del valor que pren la proporció mostral P respecte π. Val $\sqrt{\frac{\pi(1-\pi)}{n}}$ i també es sol estimar substituint π per P. Per exemple, si π fos 0.25, una mostra de 100 observacions donaria una proporció mostral P amb error tipus (exacte) igual a 0.0433: la proporció mostral oscil·la normalment al voltant de 25% amb un error tipus de 4.33%.

3. El resultat d'una enquesta pre-electoral ens diu que la intenció de vot per a determinat candidat està entre el 9.3% i el 12.9%. L'enquesta es basa en 1099 entrevistes telefòniques, però no es diu enlloc quin és el grau de confiança exacte d'aquest interval. Sospitant que s'ha utilitzat una aproximació a la Normal per l'elevat nombre d'observacions, trobeu justificadament el grau de confiança emprat.

   P val el centre de l'interval = 0.111, i la distància del centre a un extrem és 0.018, que equival a $z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$

   Trobem el valor de l'error tipus (0.009475748), i d'aquí es pot calcular el valor de $z_{1-\alpha/2} = \frac{0.018}{0.009475748}$=1.90

   De les taules de la Normal veiem que P(Z<1.90) = 0.9713, de manera que 1-α/2 = 0.9713, i α=0.0574.

   Per tant, es tractava d'un interval de confiança 94.26%.

4. A partir dels resultats de l'enquesta s'escolten les següents afirmacions a un programa de TV. Comenteu la validesa d'aquestes afirmacions:

   TERTULIÀ A: "aquesta mostra no és fiable, hi ha milions d'electors, i només s'ha preguntat a poc més de mil"

   La relació entre la mida de la població i la de la mostra no és rellevant, si la mostra és representativa es pot trobar una bona estimació amb un error petit, i una mostra de 1099 individus dona una precisió que sembla suficient per al que es pretén.

TERTULIÀ B: “de cap manera, aquesta mostra és molt fiable, la forquilla té només 3.6 punts d’imprecisió” D’acord amb que l’ample de la forquilla podria ser apropiat per als objectius de l’enquesta, estimar quin vot pot obtenir un candidat, però en tot cas la fiabilitat no depèn només d’això: la mostra ha de ser representativa, idealment una mostra aleatòria simple de la població, i la pregunta ha d’estar relacionada amb l’objectiu.

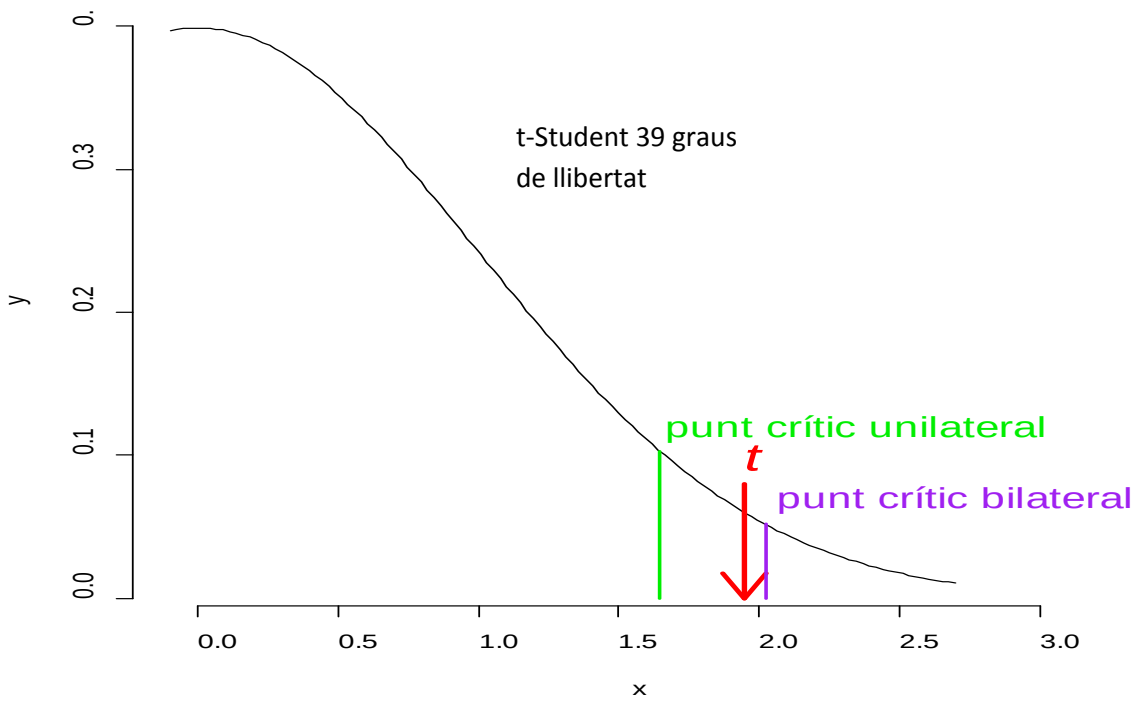
TERTULIÀ C: “no sabeu què dieu, la mostra no és vàlida perquè (posa aquí que) *les entrevistes s’han fet només a telèfons fixes, de 9 del matí a 5 de la tarda*” En aquest punt té raó, perquè pot afectar a la representativitat de la mostra. Els individus que no es troben a casa pel matí o que ja no utilitzen telèfon fixe no es troben representats a la mostra, per tant, si l’opinió d’aquest grup està molt vinculada a les preferències de vot, el resultat pot estar sub o sobrerepresentat.

5. Després de la realització de l’experiment del panell solar obtenim les següents dades: 40 mesures, amb una mitjana de 200 wh i una desviació tipus igual a 65 wh. Si la potència nominal del panell declara 180W,
   - tenim alguna evidència per a dir que la potència declarada no és la correcta?
   - tenim alguna evidència per a afirmar que el panell produeix més energia que la establerta?

   Justifiqueu les dues qüestions amb les proves d’hipòtesis corresponents, i calculeu/aproximeu els p-valors de cada prova, junt amb un diagrama en el que representareu les solucions trobades.

H_0: μ=180 H_1: μ≠180	H_0: μ=180 H_1: μ>180
Estadístic de la prova: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ suposant variable energia mesurada ~ Normal, mostra aleatòria simple	
Rebutjarem H_0 si $\lvert t\rvert > 2.022$	Rebutjarem H_0 si $t > 1.648$
Valor de l’estadístic: $t = \frac{200-180}{65/\sqrt{40}} = 1.946$	
Per tant, no és pot rebutjar que la diferència trobada sigui producte de l’atzar P valor = P(|t| > 1.946) (aproximem per Normal) P valor = P(|Z| > 1.946) = 0.0518	Per tant, si sembla haver proves de que el panel produeix més energia que la que el fabricant ha declarat P valor = P(t > 1.946) (aproximem per Normal) P valor = P(Z > 1.946) = 0.0259

Un error habitual és pensar que 65 és la desviació poblacional (“sigma”) enlloc de la mostral (“s”). Penseu que és una dada que ve de les mesures. Llavors, no és correcte assumir que l’estadístic segueix un model Normal, ni els valors crítics d'aquesta distribució, ni els p-valors que se’n derivarien.

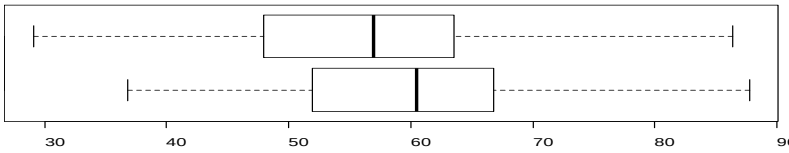


PLANTILLA **(*Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs*)**

Problema 2 (B5)

Volem estudiar si la mida dels videojocs de pagament és més gran que la dels gratuïts i no tenim cap raó teòrica per explicar que els gratuïts puguin tenir mides més grans. Hem agafat 60 jocs de cada tipus, amb el següents resultats:

	mitjana	Desviació tipus
Gratuït	56.218	11.178
Pagament	59.993	12.1




```
…… p-value = 0.04163
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:  0.189       Inf
```

1.- Si volem comparar les seves mitjanes, quin és el disseny (independent/aparellat) emprat? (0.5 punts)

Són mostres independents perquè són diferents arxius els assignats a cada grup, el mateix nombre (60) però diferents casos

- Expresseu i justifiqueu les dues hipòtesis de la prova (0.5 punts)

H0: μA-μB = 0 H1: μA-μB > 0

L'enunciat ha dit que, si no són iguals, no hi ha raó teòrica per justificar que els gratuïts puguin ser més grans.

- Valoreu com podríem argumentar amb els resultats de l'enunciat que es compleixen les premisses de normalitat i igualtat de variàncies (1 punt)

La normalitat es comprovaria amb plots de normalitat però els boxplots mostren força simetria que pot correspondre a DN..

La igualtat de variàncies es comprovaria amb PH de Fisher, però els boxplots indiquen una amplada prou semblant.

- Sota la hipòtesi d'igualtat, quin seria l'error tipus estimat per a la diferència de mitjanes mostrals? (1 punt)

Error_tipus = s_pooled × sqrt(1/60+1/60) = 11.65 sqrt(2/60) = **2.13**

(s_pooled = sqrt((59 12.1^2 + 59 11.178^2) / (60+60-2))= 11.65)

- Indiqueu quin és l'estadístic de la prova i calculeu-lo (1 punt)

t = (59.993 – 56.218) / Error_tipus = **1.7469**

- Si no hi hagués cap diferència en la mida mitjana dels dos tipus de videojocs, com es distribuiria l'estadístic de la prova? Amb un risc α=5%, feu un gràfic per il·lustrar el o els punts crítics i situar les àrees d'acceptació i de rebuig de la hipòtesi nul·la (1 punt)

t_{118} punt critic: $t_{118,\,0.95}$ (taules aprox $t_{120,\,0.95}$ = 1.658)

(zona d'acceptació de – infinit a 1.658 ; zona de rebuig de 1.658 a infinit)

- A quina conclusió arribeu? Incorporeu l'interval de confiança del 95% a la discussió (1 punt)

1.74 > 1.658 -> evidència per rebutjar la igualtat de mitjanes (també per p_value=0.04163 < 0.05 indicat a R en l'enunciat)

Segons IC unilateral dels resultats de R de l'enunciat: `95 percent confidence interval:  0.189  Inf`

Els de pagament són, com a mínim, 0.19 MB més grans en mitjana

2.- També volen estudiar la proporció de mesures per sota del llindar de 50. A la mostra n'hi ha 11 per A i 23 per B. Per comparar si els dos fabricants tenen igual proporcions (o no), indiqueu:

- Quines són les proporcions estimades de videojocs per sota de 50? (0.5 punts)

P_A = 11/60 P_B = 23/60

- Expresseu les hipòtesis de la nova prova (0.5 punts)

H0: πA-πB = 0 H1: πA-πB <> 0

- Indiqueu quin és l'estadístic de la prova i calculeu-lo (1 punt)

($P = (60*P_A + 60*P_B)/120$ = **0.283**

z = (11/60 – 23/60) / sqrt (0.283 0.717)/60 + 0.283 0.717)/60) = **– 2.43**

- Com es distribueix l'estadístic sota la hipòtesi nul·la? Feu un gràfic indicant el o els punts crítics i la zona d'acceptació i de rebuig (1 punt)

N(0,1) Punts crítics -1.96 i 1.96 (zona acceptació entre –1.96 i +1.96)

- Interpreteu el resultat, i feu una conclusió global (1 punt)

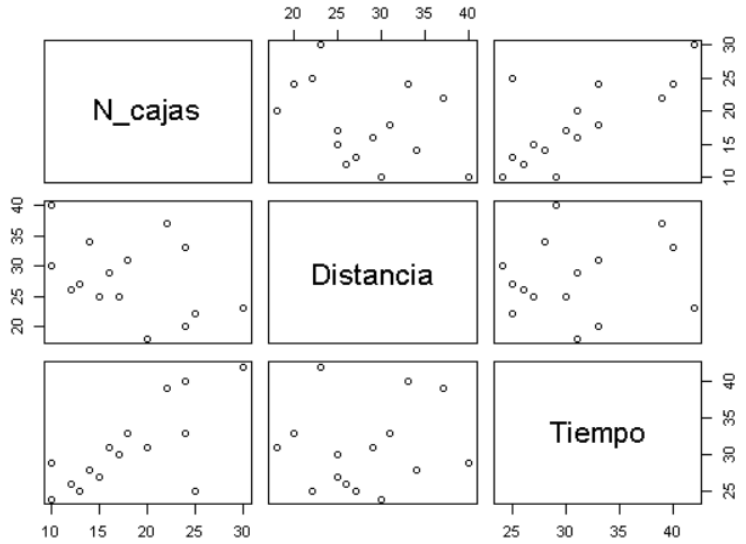
–2.43 < –1.96 - > Evidència per rebutjar la igualtat de proporcions poblacionals. La probabilitat de que un videojoc midi menys de 50 és diferent entre gratuïts i de pagament.

NOM:_____COGNOM:_____

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

Problema 3 (B6)

Disposem de les següents dades de nombre de caixes repartides, el temps invertit i la distancia recorreguda i el model predictiu que vol estudiar quins són els factors importants per predir el temps d'enviament: l'exercici es base en els resultats de: http://www.diegocalvo.es/analisis-de-regresion-lineal-multiple-en-r/



a) Amb les dades que disposes, comenta de forma descriptiva les variables fent èmfasi amb la relació que hi ha amb la variable resposta(1p):

Clarament el nombre de caixes està més relacionat amb el temps d'enviament, per tant, es lògic que el coeficient associat a aquesta variable sigui més alt.

S'ha realitzat un model lineal i s'han obtingut els resultats que veus a continuació:

```
Call:
lm(formula = datos$Tiempo ~ datos$N_cajas, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6583  -1.6018  -0.1821   2.5262   5.3952

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.5452     3.4142   5.432 0.000115 ***
datos$N_cajas   0.6845     0.1805   3.791 0.002244 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.053 on 13 degrees of freedom
Multiple R-squared:  0.5251,    Adjusted R-squared:  0.4886
F-statistic: 14.37 on 1 and 13 DF,  p-value: 0.002244
```

b) ¿És adient el model predictiu en termes dels coeficients estimats? Defineix quins són els contrastos de significació per als paràmetres del model (2p)

El model es correcte perquè els coeficients a les variables explicatives són significatius (p valors més petits a 0.05). podem afirmar que els termes són necessaris.

S'ha d'explicitar quin es el contrast de significació per als coeficients.

c) Calcula un IC per al terme β1 (coeficient associat al nombre de caixes) del model amb un 95% de confiança (1p):

IC(β1;95%) = 0,6845 ± 2,179 * 0,1805 = (0,291; 1,077)

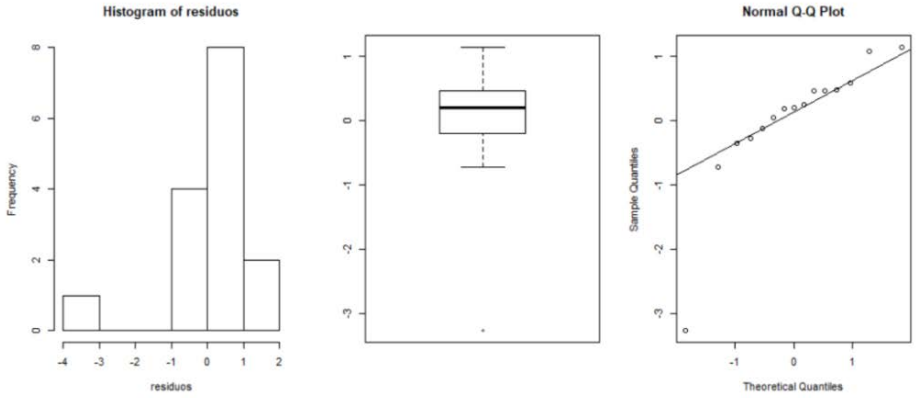
d) El temps està més relacionat amb el nombre de caixes o amb la distancia? (1p)

Clarament amb el nombre de caixes pel que es veu en el gràfic.

e) Comenta i valora la capacitat predictiva del model (1p):

El R2 es del 52% i podem afirmar que el model no serà massa predictiu ja que no explica massa bé les dades.

Un cop analitzem els residus del model, veiem els següents gràfics amb els principals resultats:



f) Són els residus adients? Disposem de tota la informació? (1p)

Hi ha una observació que distorsiona clarament les conclusions però si no la considerem, els residus són clarament normals. No tenim informació de la heterocedesticitat dels residus, faltaria tenir en compte aquest resultat abans de concloure.

g) Com es pot millorar la consistència dels resultats? Penses que seria interessant tenir més variables a l'estudi? Quines? (1p)

Clarament tenim molt poques observacions (14) i per extreure conclusions, seria millor tenir una mostra molt més nombrosa. A mes, hi ha clarament una observació molt diferent que està distorsionant el model. El procediment correcte seria treure aquesta observació del model abans de re calibrar-lo.

- Diferenciar entre zones rurals i urbanes
- Diferenciar en franges horàries (matí, tarda, nit)
- Tenir en compte el mètode d'enviament o el tipus de transport

Considerar el nombre d'empleats que participen en l'enviament

h) Quines consideracions hem fet a les variables del model? En cas que les hipòtesis no es compleixin, qué podem fer? (1p)

Principalment han de seguir una distribució Normal. Si no ho fossin, s'haurien de transformar les dades. Generalment utilitzarem una transformació logarítmica.

i) Quin seria el temps estimat per a 10 caixes? I per interval amb un 95% de confiança per al temps mitjà? (1p)

Temps = 2,3112 + 0,8772(10) + 0,4559 (20) = 20,2012 minuts.