NOM:_____COGNOM:_____

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

Problema 1 (B4)

Un grup d'estudiants de la FIB ha creat una empresa de resolució de problemes informàtics on-line. Mitjançant un complex algoritme de resolució dels problemes més comuns, l'usuari pot arribar a l'arrel del seu problema sense la necessitat de que intervingui un tècnic. El primer pas ha estat crear una plana web que ja porta activa uns mesos. Amb l'ajuda de l'eina de Google Analytics (GA) obtenim estadístiques del rendiment de la pàgina.
La nostra primera preocupació és la durada de les visites. Excloent els usuaris que reboten (aquells que no interactuen amb la pàgina), el temps mitjà de les visites segons GA és de 120 segons. Com que no estan segurs de la fiabilitat d'aquesta dada, han creat un script propi que mesura precisament aquest indicador. Per les 30 primeres visites després d'iniciar el script, han obtingut els seguts resultats pels temps i pel logaritme natural dels temps:

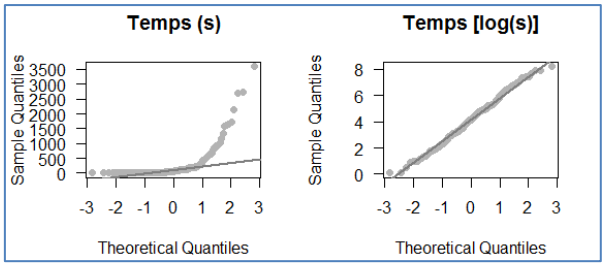
Temps

```
> summary(temps)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   2.386   51.260  221.100  487.400  408.800 3775.000
> sd(temps)  [1] 793.1788
```

```
> summary(ltemps)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8694  3.9370  5.3940  4.9890  6.0130  8.2360
> sd(ltemps) [1] 1.764887
```

Volen testejar a través d'una prova d'hipòtesis si el valor de 120 segons que dóna GA pot ser la mitjana de temps poblacional o no emprant aquestes dades.

- Observant els dos *qqnorms* de les dades, argumenteu si s'han d'emprar les dades transformades o sense transformar (1 punt)



Sense transformar no compleixen normalitat. Amb la transformació logarítmica sí.

- Plantegeu les hipòtesis de la prova (a les dades transformades o no segons la resposta anterior) tot mencionant si ha de ser unilateral o bilateral (1 punt) (ltemps és X)

$\begin{cases} H_0: \mu = \ln(120) = 4.787 \\ H_1: \mu \neq \ln(120) = 4.787 \end{cases}$ (bilateral)

- Indiqueu l'expressió de l'estadístic per resoldre el test d'hipòtesis i la seva distribució sota la hipòtesi nul·la (1 punt)

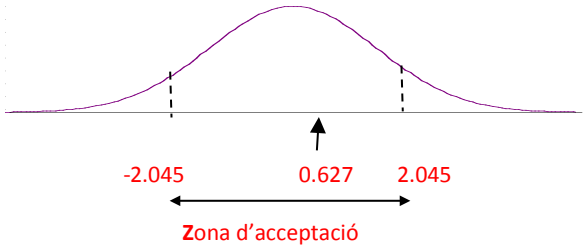
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{29} \qquad Premisses: m.a.s, X \sim N$$

- Calculeu el valor de l'estadístic i el/s punt/s crític/s suposant un α = 0.05 (1 punt)

$$t = \frac{4.989 - \ln(120)}{1.76/\sqrt{30}} = \frac{12.057}{1.76/\sqrt{30}} = 0.627$$

$$Punts\ crítics = \begin{cases} t_{29,0.025} \approx -2.045 \\ t_{29,0.975} \approx 2.045 \end{cases}$$

- Representeu gràficament la distribució de l'estadístic, el seu valor, el/s punt/s crític/s, i les zones d'acceptació i rebuig (1 punt)



- Indiqueu, justifiqueu i interpreteu quina és la conclusió de la prova d'hipòtesis? (1 punt)

No podem rebutjar H_0,
ja que el valor de l'estadístic està dins de la regió compresa entre els punts crítics (està a zona d'acceptació)
Per tant, no tenim prou evidència per dir que la mitjana poblacional de les visites sigui diferent de 120 segons.

-Calculeu un Interval de Confiança del 95% (IC95%) per a la mitjana poblacional (1 punt)

$$IC(\ln(\mu)\,,95\%) = \bar{x} \pm t_{29,0.975} \cdot \frac{s}{\sqrt{n}} = 4.989 \pm 2.045 \cdot \frac{1.76}{\sqrt{30}} = [4.33\,,5.65]$$
$$IC(\mu, 95\%) = [e^{4.49}\,, e^{5.49}] = [76.09\,, 283.19]$$

- Interpreteu el IC anterior i relacioneu-lo amb la conclusió de la prova d'hipòtesis (1 punt)

Amb una confiança del 95% la mitjana poblacional del ln del temps estarà entre 4.33 i 5.65
(amb una confiança del 95% la mitjana poblacional del temps estarà entre 76.09 i 283.19)

El valor a prova (ln(120) o 120) està dins l'IC. Res s'oposa a acceptar H_0

Un altre preocupació que tenen són els visitants que romanen menys d'un *temps mínim* (que tenen estipulat en 15 segons) a la pàgina perquè consideren que són usuaris als que alguna cosa no els hi ha agradat. En les dades obtingudes hi ha un 20% de persones que romanen menys d'aquest *temps mínim*. Calculeu un IC 95% per a aquesta proporció de persones que romanen menys del *temps mínim* a la pàgina i interpreteu-lo. (1 punt)

$$Emprant\ estimació \rightarrow IC(\pi, 95\%) = p \pm z_{0.975} \cdot \sqrt{p \cdot \frac{1-p}{n}} = 0.2 \pm 1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{30}} = 0.2 \pm 0.14 = [0.06\,, 0.34]$$

(o bé $Màxima\ indeterminació \rightarrow IC(\pi, 95\%) = p \pm z_{0.975} \cdot \sqrt{0.5 \cdot \frac{0.5}{n}} = 0.2 \pm 1.96 \cdot \sqrt{\frac{0.5 \cdot 0.5}{30}} = 0.2 \pm 0.18 = [0.02\,, 0.38]$)

Amb una confiança del 95% la proporció poblacional de persones que romanen menys d'un *temps mínim* està entre 6% i 34% (o 2% i 38%)

I calculeu quan hauria de ser la n de la mostra per obtenir l'anterior IC amb la meitat d'amplada (1 punt)

$1.96 \cdot \sqrt{\frac{0.2 \cdot 0.8}{n}} = 0.14/2$ $\quad 1.96\sqrt{\frac{0.2 \cdot 0.8}{n}} = 0.07$ $\quad \sqrt{0.2\ 0.8} = 0.036\sqrt{n}$ $\quad 123.5 = n$ $\quad n > 123$

(o bé $1.96 \cdot \sqrt{\frac{0.5 \cdot 0.5}{n}} = 0.18/2$ $\quad 1.96\sqrt{\frac{0.5 \cdot 0.5}{n}} = 0.09$ $\quad \sqrt{0.5\ 0.5} = 0.046\sqrt{n}$ $\quad 118.1 = n$ $\quad n > 118$)

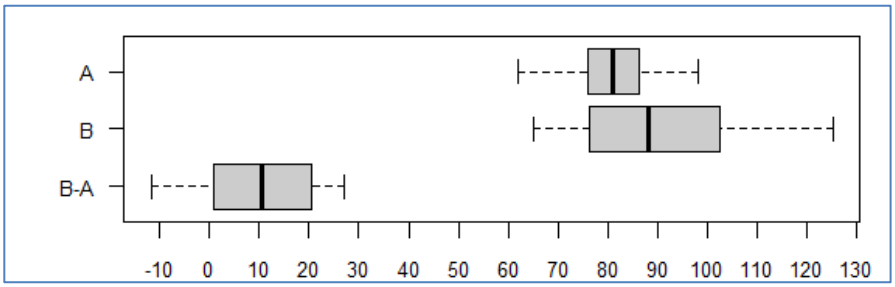
NOM:_____COGNOM:_____

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

Problema 2 (B5)

A Google Analytics (GA), hi ha una eina per assignar a certes pàgines un determinat import econòmic per visita (per exemple, a una pàgina de petició de servei de 20 euros que el contracten 1 de cada 20 persones, es pot assignar el valor de 1 euro). Els estudiants han dissenyat un experiment en que durant 2 mesos (61 dies) recullen els guanys estimats segons dos prototipus de pàgina (A i B). La descriptiva es mostra a continuació:

	A	B	B-A
Mitjana	80.77	89.80	9.04
Desv. Tipus	8.34	16.52	12.09
Mín	61.82	64.93	-11.62
Q1	76.38	76.87	1.37
Mediana	80.95	88.26	10.52
Q3	86.12	100.60	20.00
Màx	98.17	125.10	26.97



- Justifiqueu si es tracta de dades aparellades o independents (1 punt)

Aparellades pel dia (uns mateixos dies es recullen les dades del prototipus A i del B)

- Plantegeu una prova d'hipòtesis per decidir si la mitjana de guanys diaris és la mateixa en ambdós dissenys o si en el cas B són majors. Indiqueu les hipòtesis i si ha de ser unilateral o bilateral (1 punt)

$$\begin{cases} H_0: \mu_d = 0 \\ H_1: \mu_d > 0 \end{cases} \text{ (unilateral)}$$

- Indiqueu l'expressió de l'estadístic, la seva distribució sota H_0 i les premisses assumides (1 punt)

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{60}$$ Premisses: d = B-A ~ N i m.a.s

- Calculeu el valor de l'estadístic i el/s punt/s crític/s suposant un α = 0.05 (1 punt)

$$t = \frac{9.04}{12.09/\sqrt{61}} = 5.84$$

$$Punt\ crític \rightarrow t_{0.95,60} = 1.671$$

- Indiqueu, justifiqueu i interpreteu quina és la conclusió de la prova d'hipòtesis? (1 punt)

Hi ha evidència per rebutjar H_0,
ja que el valor de l'estadístic està més enllà del punt crític
Per tant, hi ha prou evidència per dir que la mitjana dels guanys són superiors en el cas B

Ara suposem que les dades anteriors no s'han recollit els mateixos dies pel cas A i B sinó en diferents mesos, i plantegem la prova d'hipòtesis per decidir si la mitjana de guanys diaris és la mateixa en ambdós prototipus o no.
- Indiqueu les hipòtesis i si ha de ser unilateral o bilateral (1 punt)

H_0: $\mu_B-\mu_A=0$
H_1: $\mu_B-\mu_A<>0$
bilateral

- Indiqueu l'expressió de l'estadístic, la seva distribució sota H_0 i les premisses assumides. Indiqueu si la descriptiva de les dades recolza assumir les premisses (1 punt)

t = (mean(B) – mean(A)) / Error_tipus (és t_{120})
i Error_tipus és s_pooled sqrt(1/61+1/61) amb s_pooled = sqrt((60 s_B^2 + 60 s_A^2)/120)

Premisses:
A i B normals (boxplots força simètrics tot i que el de B menys)
Sigma_B=Sigma_A (s_B força més gran que s_A i també boxplot de B més dispers que el de A. Molt probablement no es compleix)

- Calculeu l'error tipus estimat per a la diferència de mitjanes mostrals i el valor de l'estadístic (1 punt)

Error_tipus = s_pooled sqrt(1/61+1/61) = 13.086 sqrt(2/61) = 2.37
(s_pooled^2 = (60 16.52^2 + 60 8.34^2) / (61+61-2) = (16374.62+4173.34)/120 = 171.235
(s_pooled = sqrt(171.235) = 13.086)

t = (89.80 – 80.77) / Error_tipus = 9.03 / 2.37 = 3.81

- Calculeu i interpreteu l'interval de confiança del 95% per a la diferència d'esperances (1 punt)

(mean(B) – mean(A)) $\pm$ $t_{120,0.975}$ Error_tipus = 9.03 $\pm$ 1.98 2.37 = 9.03 $\pm$ 4.69 = [4.34 , 13.72]

Amb una confiança del 95% la diferència de mitjanes poblacionals (B-A) dels guanys en els prototipus A i B podria estar entre 4.3 i 13.7 [euros].

- Indiqueu, justifiqueu i interpreteu quina és la conclusió de la prova d'hipòtesis? (1 punt)

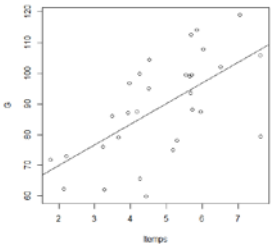
Hi ha evidència per rebutjar H_0,
ja que el valor posat a prova no està al IC (l'estadístic estaria fora dels punts crítics en zona de rebuig)
Per tant, hi ha prou evidència per dir que la mitjana dels guanys d'A i B no són iguals

NOM:_____COGNOM:_____

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

Problema 3 (B6)

Els nous emprenedors volen estudiar si existeix alguna mena de relació lineal entre la durada de les visites i el guany monetari obtingut en visites en que es fa alguna compra. Ho estudien a partir del logaritme del temps (ltemps) i els guanys (G) de 30 visites. Els resultats de la descriptiva, de la covariància entre ltemps i G, i de la regressió lineal corresponents són:



	Mitjana	Desviació tipus	Min	Max
G	88.90	16.37	58.2	119.33
ltemps	4.83	1.51	1.5	7.8

$\text{cov}(G, ltemps) = S_{G,ltemps} = 15.34$

```
summary(lm(G ~ ltemps)):  Coefficients:      Estimate     Std.Error    t value      Pr(>|t|)
                            (Intercept)       ----        8.110        6.957        1.45e-07 ***
                                ltemps        ----        1.606        4.191        0.000251 ***
                          Residual standard error: 13.06 on 28 degrees of freedom
```

- Calculeu els coeficients de la recta de regressió dels guanys G en funció del logaritme del temps. I indiqueu el coeficient de determinació i la desviació residual (2 punts)

$b_1 = 15.34 / 1.51^2 = 6.73$ (o bé 4.191 * 1.606)

$b_0 = 88.90 - 6.73\ 4.83 = 56.4$ (o bé 6.957 * 8.110)

Corr(G,ltemps) = 15.34 / (16.37 1.51) = 0.62 R^2 = sqr(0.62) = 0.385 (coeficient de determinació)

La desviació residual és 13.06

- Expliqueu què ens indiquen el pendent de la recta, el coeficient de determinació, la correlació i la desviació residual (2 punts)

Pendent: cada increment de 1 en el logaritme natural del temps implica augmentar 6.73 els guanys

R^2 : només el 38.5% de la variabilitat de G és explicable per ltemps

Correlació: relació positiva però no molt forta

Desviació residual: la desviació en les prediccions serà de l'ordre de 13 euros

- Calculeu un IC 95% del pendent de la recta i resoleu la prova d'hipòtesis de si la recta és plana o no. (2 punts)

$IC_{95\%}(\beta_1) = b_1 \pm t_{28,0.975}\, Sb_1 = 6.73 \pm 2.048 \cdot 1.606 \approx 6.73 \pm 3.29 \approx [3.44; 10.02]$

H_0: $\beta_1=0$

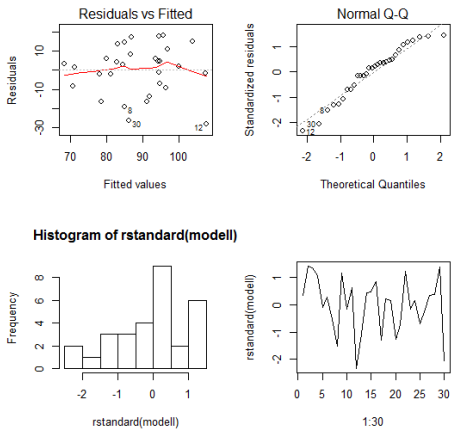
H_1: $\beta_1<>0$

Hi ha evidència per rebutjar H_0,

ja que el valor posat a prova no està al IC (l'estadístic estaria fora dels punts crítics en zona de rebuig)

Per tant, hi ha prou evidència per dir que la recta no és plana

- Enuncieu les premisses o hipòtesis de la regressió lineal i comenteu si es compleixen o no per aquest cas concret. Especifiqueu de quins resultats i/o gràfics es dedueixen els vostres comentaris. (2 punts)



Linealitat: prou clara al plot entre G i ltemps de l'enunciat, encara que amb força dispersió. Al gràfic dalt-esquerra la distribució dels punts és dispersa però entorn de la recta

Homoscedasticitat: raonable no hi ha patró, no hi ha grans zones amb més i menys variabilitat (gràfic dalt-esquerra i baix-dreta)

Normalitat: l'ajust a una normal és força correcte en el NormalQQ, tot i que l'histograma mostra certa asimetria amb cua a l'esquerra

Independència: molt raonable, ja que no hi ha patró que indiqui dependència (gràfic dalt-esquerra i baix-dreta)

- Calculeu una predicció del guany esperat per visites que romanen 120 segons amb un interval de confiança al 95% (1 punt)

ln(120) -> 4.787
G → 56.42 + 6.729 * ln(120) → 88.64

IC → $88.64 \pm t_{28,0.975}$ 13.06 sqrt (1/30 + ($(4.787-4.827)^2$ / s_{ltemps}^2(30-1)))
88.64 ± 2.048 13.06 sqrt (1/30 + ($(4.787-4.827)^2$ / 1.51^2(30-1)))
88.64 ± 2.048 13.06 0.18
88.64 ± 4.88
[83.76 , 93.52]

- Feu una valoració global d'aquest model de regressió en quant a la validació, el coeficient de determinació i la desviació residual (1 punt)

La validació de premisses és prou correcta
Però el coeficient de determinació és força baix, per tant el model lineal no ajusta massa bé
El problema no sembla de falta de linealitat (com ho mostra el plot de l'enunciat i la correlació) sinó de la desviació residual (de l'ordre de 13 euros de desviació en les prediccions, tenint en compte el rang de valors dels guanys entre 58.2 i 119.33) i del poc ajustament del núvol de punts a la recta.