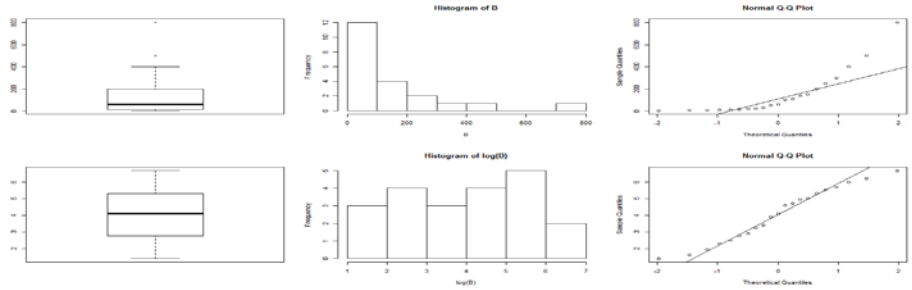
NOM:_____COGNOM:_____

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

Problema 1 (B4)

Els últims anys s'han popularitzat els serveis d'emmagatzematge al núvol. Hem recollit el temps de pujada a Dropbox(B) de 21 arxius de text de diverses mides, amb els següents resultats del temps de pujada (B) i dels logaritmes d'aquests temps (log_B):

	mitjana	Desviació tipus
B	151.76	203.91
log_B	4.03	1.60



Indiqueu com argumentar amb els resultats de l'enunciat si es compleix o no la premisa de normalitat en B i log_B (1 punt)

La normalitat es comprova amb el plot de normalitat (el de més a la dreta quan els punts queden alineats).
Quan el boxplot i l'histograma mostren força simetria, no es descarta la opció d'una distribució normal.

Per B es descartaria la normalitat
Per log_B no es descartaria la normalitat

A) Primer suposarem coneguda la desviació poblacional del logaritme d'aquests temps (log_B), igual a 1.5 i plantejarem la prova d'hipòtesis sobre si el valor esperat de la mitjana de log_B és 5 o no amb un risc del 5%. Per això indiqueu:
- les hipòtesis (1 punt)

H_0: $\mu_{log\_B} = 5$

H_1: $\mu_{log\_B} <> 5$

- l'estadístic amb la seva distribució sota la hipòtesis nul·la, i calculeu-ne el valor (1 punt)

z = (mean(log(B)) – 5) / sigma/sqrt(21)

TCL ens assegura que z és N(0,1)

z = (4.03-5) / (1.5/sqrt(21)) = -0.97 / 0.327 = -2.966

- calculeu el p_valor (1 punt)

2* (1-taules(2.97)) = 2*(1-0.9985)) = 0.003

- representeu gràficament l'estadístic amb els punts crítics, el p_valor i les zones d'acceptació i rebuig (1 punt)

$Z_{0.025}$ = -1.96
$Z_{0.975}$ = 1.96

- en funció dels apartats anteriors, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats (1 punt)

Hi ha evidència per rebutjar H_0 de mitjana igual a 5
(valor estadístic fora punts crítics, és a dir en zona de rebuig)
(p-valor inferior al risc del 5%)

No és raonable creure que la mitjana poblacional del logaritme del temps (log_B) sigui 5
(assumint desviació poblacional de 1.5)

B) Ara no suposarem coneguda la desviació poblacional del logaritme d'aquests temps i tornem a plantejar la prova d'hipòtesis sobre si el valor esperat de la mitjana del logaritme del temps és 5 o no amb un risc del 5%. Per això indiqueu:
- l'estadístic amb la seva distribució sota la hipòtesis nul·la, i calculeu-ne el valor (1 punt)

t = (mean(log(B)) – 5) / s/sqrt(21)

(t és t_{20})

(4.03-5) / (1.60/sqrt(21)) = -0.97 / 0.35 = -2.77

- l'error estàndard de l'estimador de la mitjana poblacional (1 punt)

(1.60/sqrt(21)) = 0.35

- en funció dels apartats anteriors, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats (1 punt)

Punts crítics: $t_{20,0.025}$ = -2.086
$t_{20,0.975}$ = 2.086

Hi ha evidència per rebutjar H0 de mitjana igual a 5
(valor estadístic fora punts crítics, és a dir en zona de rebuig)

No és raonable creure que la mitjana poblacional del logaritme del temps (log_B) sigui 5
(sense assumir desviació poblacional coneguda)

- proporcioneu un interval de confiança del 95% per a la μ i interpreteu-lo (1 punt)

$4.03 \pm t_{20,0.975}$ 1.60 / sqrt(21) = 4.03 ± 2.086 1.60 / sqrt(21) = 4.03 ± 2.086 0.35 = 4.03 ± 0.73 = [3.3, 4.76]

Amb una confiança del 95% el valor poblacional de la mitjana del logaritme del temps està entre 3.3 i 4.76

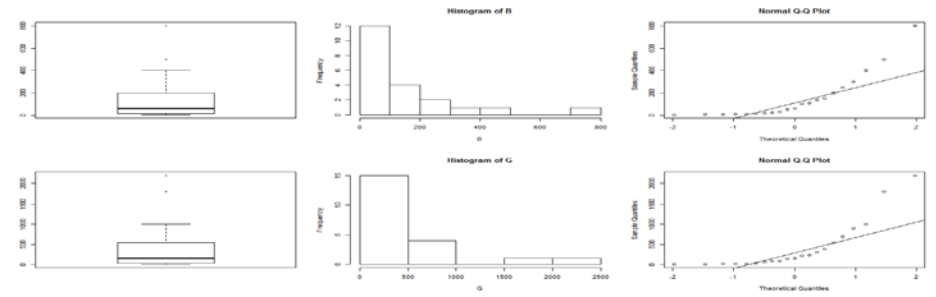
NOM:_____COGNOM:_____

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

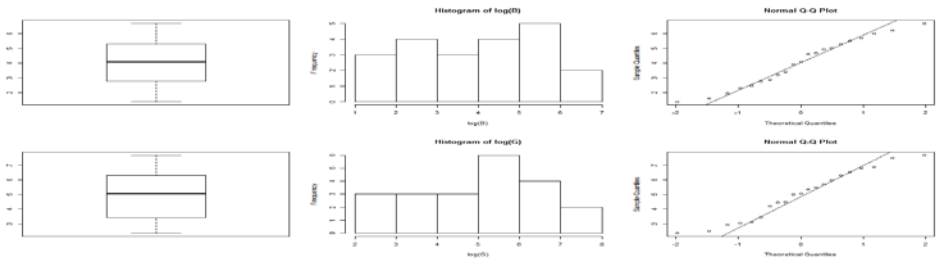
Problema 2 (B5)

Continuant el problema 1, ara volem avaluar la diferència de temps de pujada entre Dropbox (B) i Google Drive (G). Contractem dos comptes amb espai il·limitat i pugem 21 arxius de diverses mides a Dropbox i uns altres 21 arxius de mides similars a Google Drive. Els arxius són pujats en ordre aleatori a cadascun dels 2 sistemes. Els resultats del temps de pujada mesurats (B i G) i dels logaritmes d'aquests temps (log_B i log_G) són:

	mitjana	Desviació tipus
B	151.76	203.91
G	424.94	602.05



	mitjana	Desviació tipus
log_B	4.03	1.60
log_G	4.99	1.65



- Indiqueu i justifiqueu d'acord amb l'enunciat si el disseny emprat és independent o aparellat (1 punt)

Són mostres independents perquè són diferents arxius

- Expresseu les hipòtesis de la prova de igualtat de mitjanes (bilateral) dels logaritmes dels temps (1 punt)

H_0: $\mu_{log\_B}-\mu_{log\_G}=0$

H_1: $\mu_{log\_B}-\mu_{log\_G}<>0$

- Valoreu com podríem argumentar igualtat de variàncies amb els gràfics de l'enunciat (1 punt)

La igualtat de variàncies es comprovaria amb la prova d'hipòtesi corresponent però en els boxplots es pot observar un repartiment i amplada prou semblants (caixa i bigotis)

- 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(1/21+1/21) = 1.63 sqrt(2/21) = 0.50

($s\_pooled^2 = (20\ \ 1.60^2 + 20\ \ 1.65^2) / (21+21-2) = (51.2+54.45)/40 = 105.65/40 = 2.64$

(s_pooled = sqrt(2.64) = 1.63)

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

t = (mean(log_B) – mean(log_G)) / Error_tipus (és t_{40})

t = (4.03 – 4.99) / Error_tipus = -0.96/0.50 = -1.91

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

t_{40}

punts critics: $\pm$ $t_{40,\ 0.975}$ (taules $t_{40,\ 0.975}$ = 2.021)
Zona d'acceptació de – 2.021 a 2.021
Zona de rebuig fins a -2.021 i a partir de 2.021

- Doneu la conclusió de la prova i interpreteu els resultats. (1 punt)

No hi ha evidència per rebutjar H_0 d'esperances iguals
(valor estadístic dins punts crítics, és a dir en zona d'acceptació)

No és raonable no creure que les mitjanes poblacionals dels logaritmes del temps (log_B i log_G) siguin iguals

- Calculeu l'interval de confiança del 95% per a la diferència de esperances ($\mu_{log\_B} - \mu_{log\_G}$) (1 punt)

(4.03 – 4.99) $\pm$ $t_{40,0.975}$ Error_tipus = -0.96 $\pm$ 2.021 0.50 = -0.96 $\pm$ 1.01 = [-1.97 , 0.05]

- Interpreteu el IC anterior i comenteu què aporta a la conclusió de la prova. (1 punt)

Amb una confiança del 95% la diferència dels valors poblacionals de les mitjanes dels logaritmes del temps (log_B i log_G) està entre -1.97 i 0.05

0 està en IC, per tant concorda amb no haver-hi evidència per rebutjar H_0 d'esperances iguals

- Si els 21 arxius pujats a Dropbox i Google Drive fossin els mateixos comenteu què canviaria en el disseny emprat (1 punt)

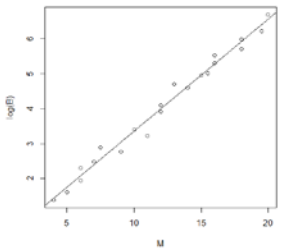
Mostres aparellades
Variabilitat més controlada

NOM:_____COGNOM:_____

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

Problema 3 (B6)

Finalment volem explicar el logaritme del temps de pujada a Dropbox (log_B) en funció de la mida M (entre i 20 GB) dels 21 arxius. Els resultats de la descriptiva de log_B i M, de la covariància entre log_B i M, i de la regressió lineal corresponents són:



	Mitjana	Desviació tipus
Log_B	4.03	1.60
M	12.1	4.97

cov(log_B,M) = $S_{M\ log\_B}$ = 7.90

summary(lm(log_B ~ M)):

Coefficients:	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	----	0.122702	1.284	0.214
M	----	0.009399	34.020	<2e-16

Residual standard error: 0.209 on 19 degrees of freedom

Multiple R-squared: ----

- Calculeu la correlació entre M i log_B i el coeficient de determinació (R^2) (1 punt)

Corr(log_B,M) = 7.90 / (1.60 4.97) = 0.993

R^2 = sqr(0.993) = 0.986

- Interpreteu els valors anteriors (1 punt)

Corr positiva i molt propera a 1 → forta realció lineal positiva (mida més gran implica més temps de pujada)

R^2 proper al 100% El 98% de la variabilitat de log_B és explicable per M

- Calculeu els coeficients de la recta de regressió del logaritme del temps (log_B) en funció de la mida M (1 punt)

$b_1 = 7.90 / 4.97^2 = 0.32$ (o bé 34.020 * 0.009399)

$b_0 = 4.03 - 0.32\ \ 12.1 = 0.16$ (o bé 1.284 * 0.122702)

- Interpreteu els coeficients de la recta de regressió del logaritme del temps (log_B) en funció de la mida M (1 punt)

Cada increment unitari de 1 Gb implica multiplicar el temps de pujada per 1.4 [=exp(0.32)].

1.2 [=exp(0.16)] representaria el temps de pujada fixe

- Contrasteu si la recta passa per l'origen (risc del 5%). Indiqueu l'estadístic i la conclusió i interpretació de la prova (1 punt)

t value = b_0/S_{b0} = 1.284 (p_valor=0.214) No evidència per rebutjar que passa per l'origen

- Contrasteu si la recta de regressió és plana (risc del 5%). Indiqueu l'estadístic i la conclusió i interpretació de la prova (1 punt)

t value = b_1/S_{b1} 34.02 p_valor= <2e-16 Evidència per rebutjar recta plana

- Calculeu un IC95% del pendent de la recta. (1 punt)

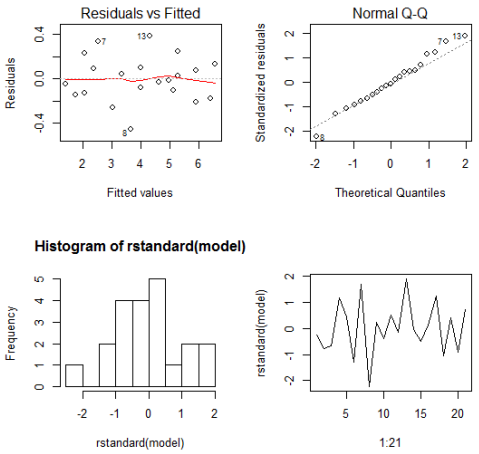
$IC_{95\%}(\beta_1) = b_1 \pm t_{19,0.975}\, Sb_1 = 0.32 \pm 2.093 \cdot 0.009399 \approx 0.32 \pm 0.02 \approx [0.30; 0.34]$

- Quin és el valor de la variabilitat residual (o del terme d'error) i quina informació aporta ? (1 punt)

Residual standard error: 0.209

És un estimador de la variabilitat dels punts al voltant de la recta, de la variabilitat dels residus

- 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: molt clara al plot entre log(B) i M de l'enunciat

Homoscedasticitat: raonable no hi ha patró en el plot [1,1], no hi ha zones amb més i menys variabilitat

Normalitat: l'ajust a una normal és força correcte en el QQplot [1,2] (recta) i l'histograma [2,1] (campana Gauss)

Independència: molt raonable, ja que no hi ha patró que indiqui dependència en el plot [2,2]