NOM:_____COGNOM:_____

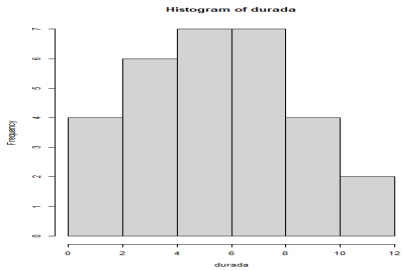
Problema 1 (B4)

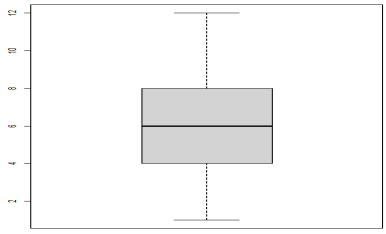
(Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs)

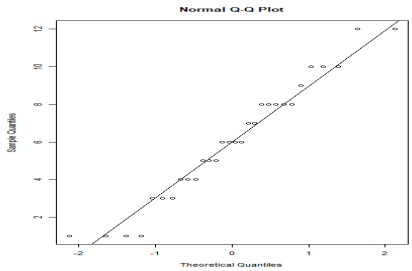
Per avaluar el funcionament d'un aplicatiu de salut, una de les dades que es recullen és la durada en minuts en que els usuaris estan connectats. Just abans de la pandèmia els valors poblacionals dels paràmetres que s'assumien eren de 5 minuts de mitjana i 3 minuts de desviació

Per comprovar si l'ús de l'aplicatiu ha canviat molt, es recull una nova mostra de 30 durades:

$$\sum_{i=1}^{30} \text{durada}_i = \mathbf{181} \qquad \sum_{i=1}^{30} \text{durada}_i^2 = \mathbf{1385}$$







1.- (0.5 punts) Amb els resultats de la mostra, justifiqueu si es pot considerar que segueixen el model normal.

Si, perquè
l'histograma sembla força la campana de Gauss (encara que cues no totalment simètriques),
el boxplot és força simètric i centrat (tant la caixa com els bigotis),
el normal_QQ_plot s'alinea força a una recta, indicant que els quantils de les dades es corresponen al quantils del model normal

2.- (0.5 punts) Calculeu una estimació puntual de l'esperança i de la desviació de la durada

mitjana: 181 / 30 = **6.03 min**

desviació tipus: s = sqrt ((1385 – (181^2/30))/29) = **3.18 min**

3.- (1.5 punt) Calculeu un interval de confiança al 95% de l'esperança de la durada assumint la desviació poblacional de 3 minuts, i interpreteu-lo

$6.03 \pm Z_{0.975} * (\sigma/\text{sqrt}(30))$ = 6.03 ± 1.96 * (3/sqrt(30)) = **[4.96 , 7.11]**

Amb 95% de confiança la mitjana esperada de la durada de les connexions a l'aplicatiu serà d'entre una mica menys de 5 minuts (4.96 minuts) i una mica més de 7 minuts (7.11 minuts). En un 2.5% de casos es pot esperar inferior a 4.96 i en un altre 2.5 % superior a 7.11

4.- (1.5 punts) Calculeu un interval de confiança al 95% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 3

$6.03 \pm t_{29,0.975} * (s/\text{sqrt}(30))$ = 6.03 ± 2.045 * (3.18/sqrt(30)) = **[4.84 , 7.22]**

És un interval més ample (menys precís) pq usem la desviació estàndard de la mostra i no la desviació poblacional, usem informació de la mostra que té més incertesa que no saber el valor teòric). Usant s i no σ implica usar la distribució t enlloc de la Z que és més ample per la mateixa confiança

5.- (1.5 punt) Calculeu un interval de confiança al 99% de l'esperança de la durada sense assumir que la desviació poblacional és coneguda, i compareu-lo amb el calculat a l'apartat 4

$6.03 \pm t_{29,0.995} * (s/\text{sqrt}(30))$ = 6.03 ± 2.756 * (3.18/sqrt(30)) = **[4.43 , 7.63]**

És un interval més ample perquè el volem amb més confiança i assumint menys risc; una zona de confiança més ampla porta a valors més extrems de la distribució

6.- (2 punts) Uns dels responsables de l'anàlisi d'aquestes dades tenien la sospita que la pandèmia havia fet incrementar la mitjana de la durada de la connexió dels usuaris. Per això van calcular aquests resultats:

```
t.test(durada,mu=5,alternative="greater")
t = 1.7807, df = 29, p-value = 0.04272
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:   5.047337      Inf
sample estimates:   mean of x    6.033333
```

Amb aquests resultats plantegeu la prova d'hipòtesis que representa (indiqueu hipòtesis, càlculs i conclusió), i interpreteu l'interval de confiança

H0: μ=5
H1: μ>5 (hipòtesis unilateral)

Estadístic: (6.03-5)/(3.18/sqrt(30)) **1.78**
Punt crític: $t_{29,0.95}$ **1.699**

No és raonable acceptar H0 (mitjana esperada de la durada de 5 minuts) sinó la H1 indicant que la mitjana esperada és superior a 5 minuts, ja que l'estadístic > punt crític, el p-value < risc del 5%, el valor 5 queda fora del IC [5.047,Inf))

IC [5.047,Inf) és unilateral i indica que amb confiança del 95% com a mínim la durada mitjana serà una mica superior a 5 minuts (concretament superior a 5.047 minuts)

7.- (1 punt) Amb aquesta mateixa mostra de 30 durades calculeu un interval de confiança al 95% per a la desviació i interpreteu-lo

IC var [6.41, 18.25] (10.1*29) / 45.722) i (10.1*29) / 16.047)

IC desv: **[2.53, 4.27]**

Amb un 95% de confiança la desviació esperada en la durada de les connexions és d'entre 2.46 minuts i 4.15 minuts (res s'oposaria a acceptar que el valor de 3 minuts fos una opció vàlida per a la desviació esperada)

8.- Per altra part, en un moment de molts intents d'accés, es vol quantificar el percentatge d'èxits accedint a l'aplicatiu. Per això es recull una nova mostra de 100 intents, obtenint que en 68 sí s'ha aconseguit l'accés.
(1.5 punts) Calculeu un interval de confiança per a la proporció d'intents que sí aconsegueixen accedir-hi i interpreteu-lo relacionant-lo amb el fet de que es voldria assegurar un 80% d'èxit d'accés

sqrt(0.68*(1-0.68)/100) és 0.047 (o bé sqrt(0.5*0.5/n) és 0.05)

IC: **[0.59 , 0.77]**
0.68 ± 1.96*0.047

IC: **[0.58 , 0.78]**
0.68 ± 1.96*0.05

Amb un 95% de confiança el percentatge esperable d'èxits en l'accés està entre el 59% i el 77% (o 58% i 78%)

Les dades no mostren evidència que el percentatge d'èxit en l'accés sigui acceptable (no arriba al 80%)

NOM:_____COGNOM:_____

Problema 2 (B5)

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

Per comparar la velocitat de transferència de discos HDD (H) i SSD (S), s'han copiat fitxers, de 1 MB a 4 GB, amb un mateix ordinador. Es mesura la velocitat a MB/s de copiar uns mateixos 80 fitxers amb ambdós tipus de disc i decidint a l'atzar quin s'usarà primer. La taula següent proporciona el nombre de rèpliques, la mitjana i la desviació típica o estàndard, per a cada tecnologia de disc i per a la seva diferència:

		Velocitat en MB/s	
	Nº observacions	Mitjana	Desviació
H	80	28	3
S	80	120	9
S-H	80	92	9

Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents (0.5 punts).

D'acord amb l'enunciat, es tracta de dades aparellades perquè es copien els mateixos 80 fitxers en cada tipus de disc (H i S)

Comenteu en cada cas (o disseny) què implica en quant a la variància de la diferència. (0.5 punts)

Si les dades fossin independents, aleshores la variància de la diferència és la suma de les variàncies, és a dir, Var(S-H)=Var (S) + Var (H) = $3^2 + 9^2 = 90$.
En ser les dades aparellades, la variància de la diferència segueix la relació següent V(S-H) = V(S) + V(H) – 2·Cov(S,H), de la qual no tenim informació directe sobre la covariància, però si de V(S-H)=9^2.

Si es tracten com a mostres independents (assumint normalitat i igualtat de variàncies poblacionals), calculeu:
- la desviació pooled i l'error estàndard de la diferència de mitjanes(1 punt)

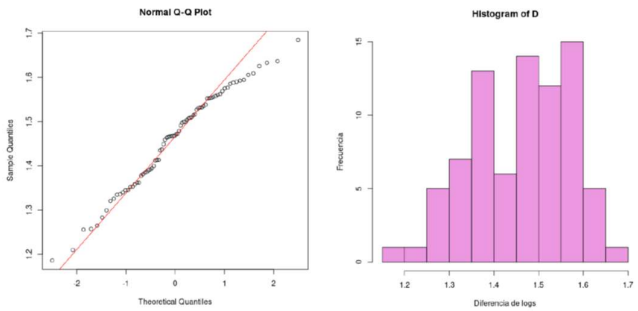
$$s^2_{pooled} = \frac{(n_S - 1)\cdot S_S^2 + (n_H - 1)\cdot S_H^2}{(n_S + n_H - 2)} = \frac{79\cdot 9^2 + 79\cdot 3^2}{80 + 80 - 2} = 45;\ s_{pooled} = 6.708$$

$$s.e = \sqrt{\frac{S^2_{pooled}}{n_S} + \frac{S^2_{pooled}}{n_H}} = \sqrt{\frac{45}{80} + \frac{45}{80}} = \sqrt{\frac{9}{8}} \approx 1.06$$

- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans) (1 punt)

$$IC(95\%, \mu_H - \mu_S) = (\bar{y}_H - \bar{y}_S) \mp z_{0.975}\cdot s.e = (28 - 120) \mp 1.96\cdot 1.06 \approx -92 \mp 2{,}08 \approx [-94, -90]$$

Es considera ara la diferència dels logaritmes ln(S)-ln(H), obtenint aquests dos gràfics:



(1 punt) Interpreteu i indiqueu de què ens informen aquests dos gràfics

El quantil-quantil (esquerre) i l'histograma (dreta) ens informen de la forma de la distribució. L'histograma és més intuïtiu, però la seva forma depèn de l'amplitud dels intervals. El quantil-quantil és més informatiu perquè reflecteix cada punt, sense necessitat de talls arbitraris.

Tots dos apunten a una distribució simètrica amb cues aplanades, que es podria modelar amb la D. Normal de Gauss-Laplace. És assenyat assumir aquesta distribució per a la inferència estadística.

Per la variable "ln(S)-ln(H)" els resultats han estat $\sum_{i=1}^{80} x_i = 116'88$ i $\sum_{i=1}^{80} x_i^2 = 171'72$.
(1 punt) Feu una estimació puntual de l'esperança i de la desviació

$\bar{x}$= 1'461 i s_X= 0'110

(1 punt) Interpreteu les estimacions anteriors

En l'escala logaritme natural les diferències es distribueixen al voltant de 1.46, bastant concentrades, ja que la distància típica a aquesta mitjana val 0.1. Això indica que S serà unes 4.3=$e^{1.46}$ vegades més ràpid.

(2 punts) Sabent que la funció de distribució corresponent a 1'664371 en una t de Student amb 79 graus de llibertat val 0.95, useu aquest valor per donar un interval simètric de confiança per a la diferència i indiqueu amb quina confiança s'haurà calculat

IC($\mu_{D,\ ??}$) ≈ 1.461 ± 1.664371 · 0.110/ √ 80 ≈ [1.436521, 1.485479] ≈ [1.44, 1.49]

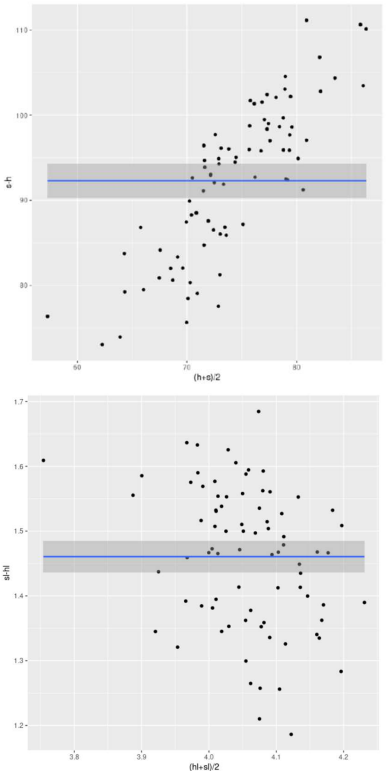
90%

(1 punt) Interpreteu l'interval anterior en relació a la comparació de les velocitats dels discs H i S.

Per poder interpretar a'han de fer exponents,
Com ln(S) − ln(H) = ln(S/H), aleshores S/H = e^D,
i IC(S/H , 0.90) ≈ $e^{1.436521}$, $e^{1.485479}$ ≈ 4.206, 4.417 ≈ 4.2, 4.4

Per fitxers entre 1 MB a 4 GB, amb una confiança del 90%, el disc S es entre 4.2 y 4.4 vegades més ràpid que el H

(1 punt) Els següents dos gràfics mostren les diferències S-H per cada fitxer en ordenades en funció de les mitjanes [(H+S)/2] en abscisses. Primer, el gràfic inicial, sense transformar; i després, el gràfic amb la transformació logarítmica. Sabent que copiar fitxers grans pot resultar en diferències mes grans, interpreti aquests gràfics. Té sentit estimar una diferència única per aplicar a tots els casos amb les dades sense transformar i amb les dades transformades?



En el primer gràfic, el núvol apunta a que la diferència és més gran com més gran és la mitjana: relació directa entre diferències i mitjanes. Aquells fitxers en què es triga més (potser per ser més grans?), la diferència és més gran. Fa dubtar si té sentit proporcionar un únic valor, la mitjana de 92 MB/s, per a tots els fitxers.

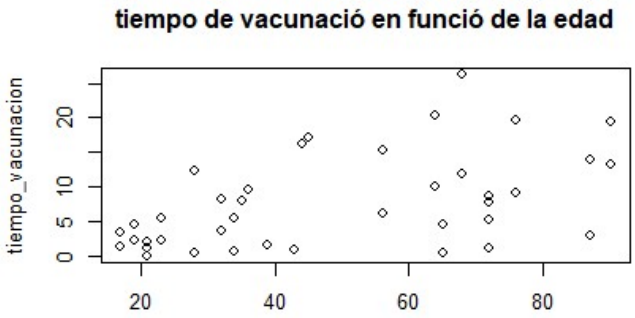
En el segon gràfic es mostra que aplicar logaritmes (naturals) ha solucionat el problema

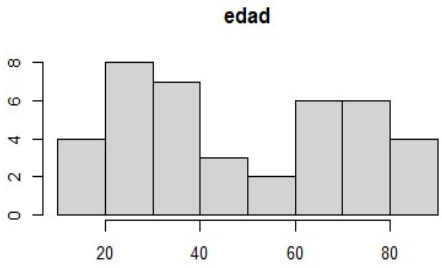
NOM:_____COGNOM:_____

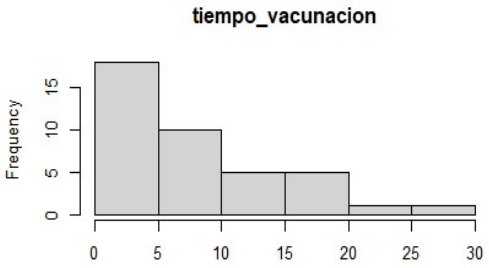
Problema 3 (B6)

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

El temps de vacunació s'entén com els minuts que passen entre la arribada d'una persona al centre sanitari, amb l'objectiu de vacunar-se, i la sortida amb la vacuna administrada. Donada l'actual situació de la COVID19 el personal sanitari es planteja si el temps de vacunació depèn de l'edat de les persones que s'han de vacunar, ja que s'han observat patrons que poden portar a pensar que sí que hi ha relació. A continuació, es proporciona alguna informació sobre les dades recollides en un centre sanitari que ha fet un estudi observacional:







1 (1 punt) Amb la informació prèvia, creieu que seria raonable plantejar un model lineal per explicar el temps de vacunació en funció de l'edat?

En el primer gràfic, aparentment hi ha una correlació positiva encara que feble entre el temps de vacunació i l'edat. La relació sembla monòtona creixent. Destaquen alguns individus amb un temps de vacunació anormalment baix. D'entrada amb la informació disponible no es descartaria un model lineal.

2 (2 punts) Donat la següent sortida de R, comenteu les estimacions de tots els paràmetres del model lineal i què signifiquen:

```
Call:
lm(formula = tiempo_vacunacion ~ edad)

Residuals:
    Min      1Q  Median      3Q     Max
-10.265  -3.495  -1.191   2.466  15.661

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.55417    2.10631   0.263 0.793894
edad         0.14626    0.03917   3.734 0.000615 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.858 on 38 degrees of freedom
Multiple R-squared:  0.2685,    Adjusted R-squared:  0.2492
F-statistic: 13.95 on 1 and 38 DF,  p-value: 0.0006155
```

Intercept o terme constant: és el terme independent de la recta, val 0.55417 (minuts), significa el temps fix que donem a un participant de qualsevol edat.

Pendent de l'edat: és el terme lineal de la recta, val 0.14626 (minuts/any), significa l'increment en el temps de vacunació per cada any d'edat addicional.

Residual Standard error, o desviació residual, val 5.858 (minuts), significa la desviació tipus de l'error aleatori que afecta a cada individu, amb una edat determinada. O l'error típic que es pot esperar a partir de l'estimació de la recta.

3 (1 punt) Valoreu la capacitat del model per explicar la variabilitat de la variable resposta.

A la vista dels resultats obtinguts el model és poc explicatiu ja que el coeficient de determinació R^2 val 0,2686. És a dir, encara que el factor "edat" és estadísticament significatiu, només explica un 27% de la variació que observem en el temps de vacunació.

4 (1 punts) Calculeu el IC del 90% de confiança per al pendent de la recta i interpreteu-lo.

IC(β1, 90%) = 0,14626 ± 1,684*0,03917 = [0,080; 0,212]

S'observa que el 0 no pertany a l'interval. El temps de vacunació s'incrementa entre 0.08 i 0.212 minuts en mitjana per cada any de més que tingui el pacient, amb una confiança del 90%.

5 (2 punts) Es vol analitzar si, donat el model anterior, podem afirmar que el temps de vacunació s'incrementa 1 minut per cada 3 anys del pacient (amb un risc del 5%):

Considerem una prova d'hipòtesi bilateral:

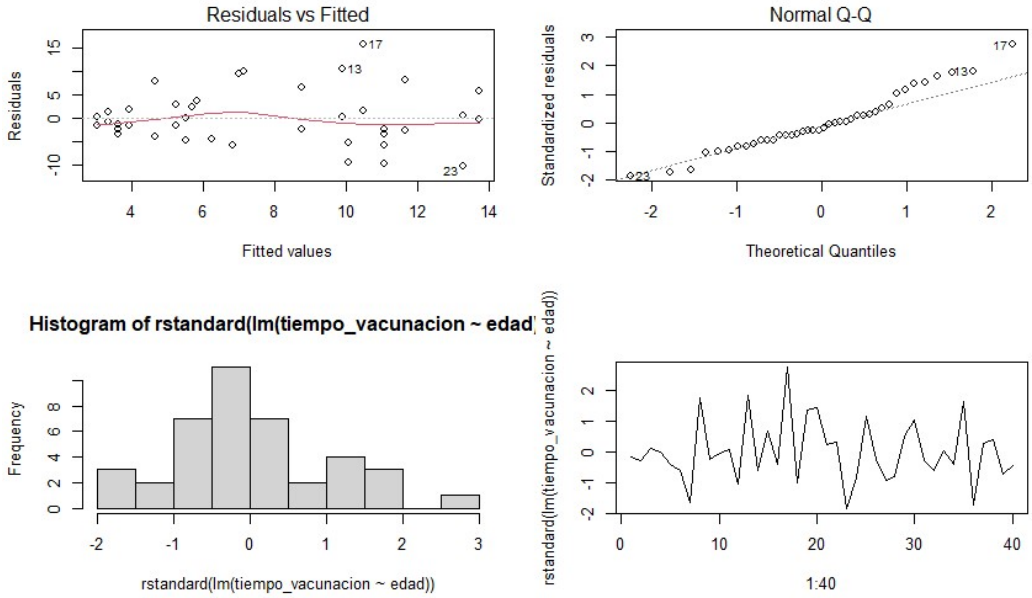
H0: β1 = 1/3

H1: β1 ≠ 1/3

t = (0,146 – 0,333)/0,039 = -4,794

L'estadístic de la prova pren un valor considerablement extrem. Si fos certa la hipòtesi nul·la, es distribuiria com una t-Student amb 38 graus de llibertat, i el resultat -4.794 estaria molt allunyat de la zona d'acceptació (límits a prop de ±2). A la vista dels resultats, clarament podem rebutjar la H0 i afirmar que el temps de vacunació no s'incrementa 1 minut per cada 3 anys d'edat.

6 (2 punts) Donats els següents gràfics de residus, valideu les premisses del model.



Com es veu en el primer gràfic, es pot assumir linealitat, però es presenten clars indicis de heteroscedasticitat en els residus. A banda d'aquest aspecte, la normalitat dels residus és qüestionable donats els resultats del Q-Q Plot, per la part dreta, que s'allunya més del que hauria de ser. La independència dels residus sembla que no està compromesa. Amb tot això, no es podria validar el model.

7 (1 punt) Si haguéssiu de fer alguna transformació sobre les dades, quina seria aparentment una transformació que podria millorar els resultats? Justifiqueu la proposta.

Una transformació logarítmica sobre el temps de vacunació disminuiria la variància en els individus que presenten més variabilitat (els que tenen temps més alts) i podria millorar el compliment de les premisses.