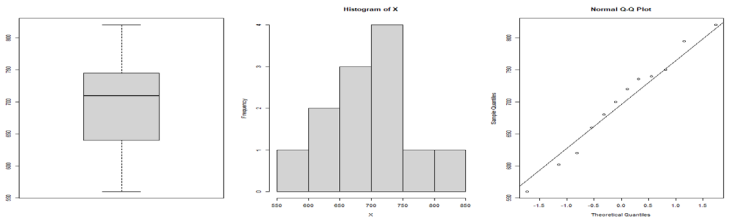
NOM:_____COGNOM:_____

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

Problema 1 (B4)

Una Universitat ha iniciat un pla de recollida de dades monitoritzant la concentració de CO_2 a les aules, en ppm (parts per milió). Obtenim, a l'atzar, 12 mesures de CO_2 en aules amb característiques, horari i ocupacions equivalents:

X <- c(560,750,660,740,620,720,680,700,820,736,602,795) $\sum X_i = 8383$ $\sum X_i^2 = 5923025$



Les mesures de CO_2 presenten una fluctuació natural, amb desviació tipus de 70 ppm- Ens diuen també que 750 ppm és un valor de referència com a llindar superior per considerar òptima la qualitat de l'aire.

1.- Calculeu l'estimació puntual de la mitjana i la desviació de la concentració de CO_2 (1 punt)

mean(X) **698.58** 8383 / 12 = 698.58

sd(X) **77.93** sqrt((5923025-(8383*8383/12))/11) = sqrt(6072.811) = 77.93

2.- Amb els valors anteriors calculats, i els anteriors gràfics descriptius, comenteu la informació que donen sobre la qualitat de l'aire i sobre si es compleixen les premisses per calcular intervals de confiança (1 punt)

La mitjana és inferior al llindar de 750, però la desviació és prou gran (superior a la que es pot esperar com a fluctuació natural)

La premissa de normalitat es compleix amb un boxplot força simètric i el Normal plot força alineat, tot i que l'histograma no presenta les cues ben bé simètriques

3.- Assumint el valor de la fluctuació natural de les mesures de CO_2 com a desviació poblacional, calculeu un interval de confiança al 95% per a la concentració mitjana de CO_2 (1 punt)

658.98 738.19

698.58 – 1.96*(70/sqrt(12)) = 698.58 – 39.61 = 658.97
698.58 + 1.96*(70/sqrt(12)) = 698.58 + 39.61 = 738.19

4.- I calculeu l'interval anterior si no assumim el valor anterior com a poblacional (1 punt)

649.07 748.10

698.58 – 2.201*(77.93/sqrt(12)) = 698.58 – 49.51 = 649.07
698.58 + 2.201*(77.93/sqrt(12)) = 698.58 + 49.51 = 748.10
($t_{11,0.975}$ = 2.201)

5.- Interpreteu i compareu els dos intervals anteriors (2 punt)

Amb un 95% de confiança el valor de la mitjana de CO_2 esperable estarà entre els valors de l'interval (658.98 i 738.19 assumint el valor de sigma, i 649.07 i 748.1 si no l'assumim)

L'interval amb sigma desconeguda és més ample ja que, al no conèixer el valor de sigma, s'aproxima per l'estimador s (que té un valor una mica superior respecte si s'assumeix el valor de 70) i s'usa la distribució t enlloc de la Normal, que porten a menys precisió

Ara ens centrarem en unes dades d'una inspecció un dia i hora concrets en la que es prenen les mesures a 30 aules, i es tenen els dos resultats (A i B) següents:

A:	B:
t = -2.04, df = 29 **alternative hypothesis: true mean is not equal to 750** **95 percent confidence interval:** **677.1729 750.0937** **sample estimates:** **mean of x** **713.6333**	**t = -2.04, df = 29** **alternative hypothesis: true mean is less than 750** **95 percent confidence interval:** **-Inf 743.9237** **sample estimates:** **mean of x** **713.6333**

Una de les dues proves aporta evidència que la mitjana de ppm de les aules és inferior al llindar de 750 amb una confiança del 95%. Indiqueu quina és la prova i indiqueu hipòtesis, conclusió de la prova i interpretació de l'interval de confiança (2 punts)

La prova B pq és unilateral, i permet contrastar un valor respecte valors inferiors

H0: μ=750
H1: μ<750

Punt crític $t_{29,0.05}$ és -1.699 i el valor de l'estadístic (-2.04) està a la zona de rebuig (de -infinit a -1.699). Per tant hi ha evidència per rebutjar la hipòtesis nul·la de 750 ppm, i acceptar que és inferior al llindar de 750 ppm

L'interval indica que amb una confiança del 95% la mitjana de CO_2 en aquestes aules serà inferior a 743.92 ppm

Seguint amb aquestes dades de la inspecció, s'obté que de les 30 aules en 23 no es supera el llindar de 750 ppm. Indiqueu un interval de confiança al 95% pel percentatge d'aules que no superen el llindar (2 punts)

P = 23/30 = 0.767

se = sqrt(0.767*0.233/30) = 0.08 (o bé se = sqrt(0.5*0.5/30) =0.09)

P – 1.96*0.08 = 0.767 – 0.157 = **0.61** (o bé 0.767 – 0.176 = **0.59**)
P + 1.96*0.08 = 0.767 + 0.157 = **0.92** (o bé 0.767 + 0.176 = **0.94**)

➔ (1) [0.61, 0.92] o (2) [0.59, 0.94]

Problema 2 (B5)

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

Volem comparar el temps d'execució de dos algorismes per ordenar vectors amb la mateixa complexitat de temps $O(n^2)$: **B**ubble sort, i **I**nsertion sort. Hem generat a l'atzar 50 vectors amb una mida entre 10^1 i 10^5 i hem calculat el temps que triga cada algoritme. La taula següent proporciona la mitjana i la dispersió (desviació típica o estàndard) per a cada algoritme i per a la seva diferència. Per la resposta 'temps', esquerra; i pel seu logaritme natural, dreta. **Cada pregunta 1 punt.**

Temps en segons			Log(temps)		
Var	Mitjana	Dispersió	Var	Mitjana	Dispersió
B	193'4	175'7	**ln(B)**	4'2	2'4
I	91'6	83'1	**ln(I)**	3'4	2'4
B-I	101'8	98'4	**ln(B)-ln(I)**	0'74	0'03

1.- Indiqueu i justifiqueu si es tracta d'un disseny de dades aparellades o independents

D'acord amb l'enunciat, es tracta de dades aparellades perquè s'ordenen els mateixos 50 vectors amb cada algoritme.

2.- Comenteu què implica cada disseny (independent o aparellat) en quant a la variància de la diferència. Dona això alguna pista sobre el grau d'aparellament (dependència) de les dades?

Si les dades fossin independents, aleshores la variància de la diferència hauria de ser la suma de les variàncies, és a dir
$\text{Var(B-I)}=\text{Var (B)} + \text{Var (I)} = 175'7^2 + 83'1^2 = 30870'5+6905'6 = 37776'1 = 194'4^2$
En ser les dades aparellades, la variància de la diferència segueix la relació $V(B-I) = V(B) + V(I) - 2\cdot Cov(B,I)$:
$Cov(B,I) = [V(B) + V(I) - V(B-I)]/2 = [175'7^2 + 83'1^2 - 98'4^2]/2 = 14046'8$
I $Corr(B,I) = Cov(B,I)/S_B S_I = 14046'8 /175'7\cdot 83'1 = 0'96$
→ B,I tenen una relació directe molt intensa (dades 'molt' aparellades)

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

$$s^2_{pooled} = \frac{(n_S-1)\cdot S_S^2 + (n_H-1)\cdot S_H^2}{(n_S+n_H-2)} = \frac{49\cdot 175'7^2 + 49\cdot 83'1^2}{50+50-2} = 18880'0 \rightarrow S = 137'4$$

$$s.e = \sqrt{\frac{S^2_{pooled}}{n_S} + \frac{S^2_{pooled}}{n_H}} = \sqrt{\frac{18880}{50} + \frac{18880}{50}} \approx 27'5$$

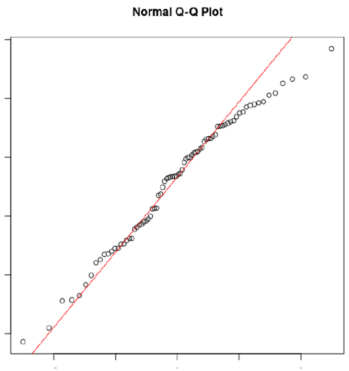
4.- un interval de confiança al 95% de la diferència de mitjanes (podeu utilitzar la convergència a la Normal per 'n' grans)

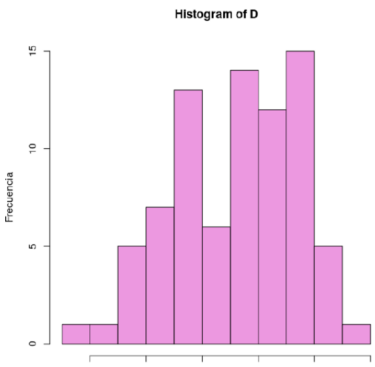
$$IC(95\%, \mu_H - \mu_S) = (\bar{y}_H - \bar{y}_S) \mp z_{0.975}\cdot s.e = (193'4 - 91'6) \mp 1'96\cdot 27'5 \approx 101'8 \mp 53'9 \approx [48'9, 156'7]$$

5.- Opineu sobre les premisses

Sense gràfics de les seves distribucions (histograma i QQPLOT), només puc opinar sobre la igualtat de les variàncies, molt dubtosa, donats els resultats: $175'7^2 = 30870'$ i $83'1^2 = 6905'6$; $VAR(B) = 4'5\cdot VAR(I)$

6.- Es considera ara la diferència dels logaritmes D=ln(B)-ln(I), obtenint aquests dos gràfics. Interpreteu i indiqueu de què ens informen aquests dos gràfics





El quantil-quantil (primer) i l'histograma (segon) 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.

7.- Interpreteu els resultats numèrics descriptius (mitjana i desviació) de la diferència dels logaritmes. Quin triga menys? Quin és més ràpid? Quant més ràpid?

En l'escala logaritme natural les diferències es distribueixen al voltant de 0'74, bastant concentrades, ja que la distància típica a aquesta mitjana val 0.03.
Això indica que B trigarà el doble ($2.09=e^{0'74}$). Per tant, I serà més eficient.

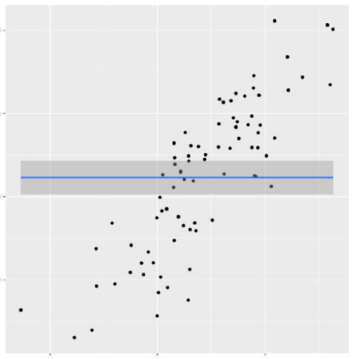
8.- Sigui T una v.a. amb distribució t de Student amb 49 graus de llibertat i P(T<0'68)= 0'75; doneu un interval simètric de confiança per a la diferència (0.5punts) i indiqueu amb quina confiança s'haurà calculat (0.5punts).

$IC(\mu_{D,\,??}) \approx 0'74 \pm 0'68 \cdot 0'03/\sqrt{50} \approx [0'736, 0'744]$
Com P(T<0'68)= 0'75, P(-0'68<T<0'68)= 0'5 → Amb una confiança del 50%

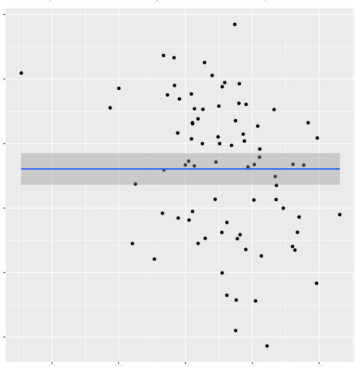
9.- Interpreteu el interval de confiança en la escala del temps (desfeu els logs)

Com ln(B) − ln(I) = ln(B/I), aleshores $B/I = e^D$,
i $IC(B/I\ ,\ 0.50) \approx e^{0'736}, e^{0'744} \approx 2.09, 2'10$
Per vectors de mida entre 10 i 10^5, amb una confiança del 50%, B triga entre 2'09 y 2'10 vegades més.

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



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. No té sentit proporcionar un únic valor de la diferència, una mitjana de 101'8 s, per a tots els vectors entre 10 i 10^5 elements..



En el segon gràfic es mostra que aplicar logaritmes (naturals) ha solucionat el problema: si té sentit proporcionar un únic valor pel rati dels temps.

B6 NOM:_____COGNOMS:_____

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

El temps emprat per a visualitzar una pàgina web a un navegador es pot descompondre en el temps destinat a la connexió amb el servidor remot, i el temps de processament del codi de la pàgina (descàrrega i mostrar a pantalla). Hem dissenyat un petit estudi, les dades del qual es troben a la dreta.

	1	2	3	4	5	6	$\sum v^2$	$\sum v$	Covar
Temps (cs)	84	96	135	112	98	136	75141	661	1365,6
Mida dades (KB)	16	45	125	76	22	190	60266	474	

La *mida de les dades* és l'espai que ocupa el fitxer HTML en kilobytes, i el *temps* és el temps mesurat en centèsimes de segon des de que es llença la petició fins a que la pàgina apareix completa al navegador (per tant, el temps total).

1. [2pts] El primer objectiu de la recerca és el temps de la primera part, el destinat a la connexió, i que no depèn de la mida de les dades. L'anàlisi estadístic definit serà un model lineal amb les variables de la taula. Heu de trobar el valor de les estimacions pels paràmetres del model: 1) terme independent, 2) terme lineal, 3) desviació residual.

   Y=temps; X=mida; estimarem els valors dels coeficients de Y=b_1X + b_0, i s, desviació dels residus.

   $s_X^2 = \frac{60266-6(474/6)^2}{5} = 4564$ $\quad$ $s_Y^2 = \frac{75141-6(661/6)^2}{5} = 464{,}17$

   b_1 = 1356,6/4564 = 0,2992 (terme lineal)

   b_0 = (661/6) – 0,2992·(474/6) = 86,53 (terme independent)

   s^2 = 5(464,17 – 0,2992·1365,6)/4 = 69,45; s = 8,334

2. [1.5pts] Expliqueu el significat de les estimacions anteriors (sigueu curosos amb les unitats corresponents a cada cas).
   1) 86,53 és el punt de tall de la recta amb l'eix Y, és a dir, quan la X=0. El podem interpretar con que hi ha un mínim de 86,53 cs de temps que sempre hi serà només perquè hi ha que fer la connexió.
   2) 0,2992 és el pendent de la recta. Significa que cada vegada que s'incrementa 1 KB la grandària del fitxer descarregat el temps total s'incrementa en 0,3 cs
   3) 8,334 cs és la desviació tipus de l'error de mesura. Significa que el temps té una variabilitat típica d'unes 8,3 cs, encara que la pàgina a descarregar fos la mateixa (o una altra amb la mateixa grandària).

3. [1.5pts] A partir de les estimacions resultants del model anterior, calculeu un interval de confiança al 95% per al temps esperat que es precisa per a connectar amb el servidor remot, i doneu una interpretació per a complementar el resultat.

   El terme independent de la recta correspon precisament a β_0 =E(Y | X=0), és a dir, al temps necessari per fer la connexió. Primer hem de trobar el valor de l'error tipus de l'estimador:

   $s_{b0}^2 = s^2\left(1/n + \bar{x}^2/(n-1)s_X^2\right)$=69,45 $\left(1/6 + 79^2/5 \cdot 4564\right) = 30{,}57$; l'error tipus és l'arrel quadrada: 5,53

   IC(β_0, 95%) = 86,53 ± $t_{4,\,0.975}$ 5,53 = [71,18; 101,88] cs

   Creiem amb una confiança del 95% que la mitjana del temps de connexió es situa entre 71 i 102 centèsimes de segon.

4. [1.5pts] Si es vol trobar un interval més estret per al paràmetre anterior, comenteu sobre l'eficàcia de les següents estratègies, justificant les respostes (preferiblement de manera formal):

   a) Empraria una mostra més gran (per exemple, 12 observacions)
   Si la n augmenta, disminuirà l'error tipus (la n és al denominador) i també el factor de la t-student, però la nova mostra hauria de distribuir-se de manera semblant, sense modificar substancialment la mitjana ni la variància de les X

   b) No augmentaria la mida de mostra, però augmentaria la mida de les pàgines
   La mitjana de les X augmentaria, fent que l'error tipus s'incrementés, per tant no tindríem un IC més estret.

   c) No augmentaria la mida de mostra, però disminuiria la mida de les pàgines
   La mitjana de les X disminuiria, i en principi també l'error tipus. No obstant, es tindria que procurar tindre la major dispersió possible en les X, perquè si la variància de les X fos molt petita l'error tipus creixeria.

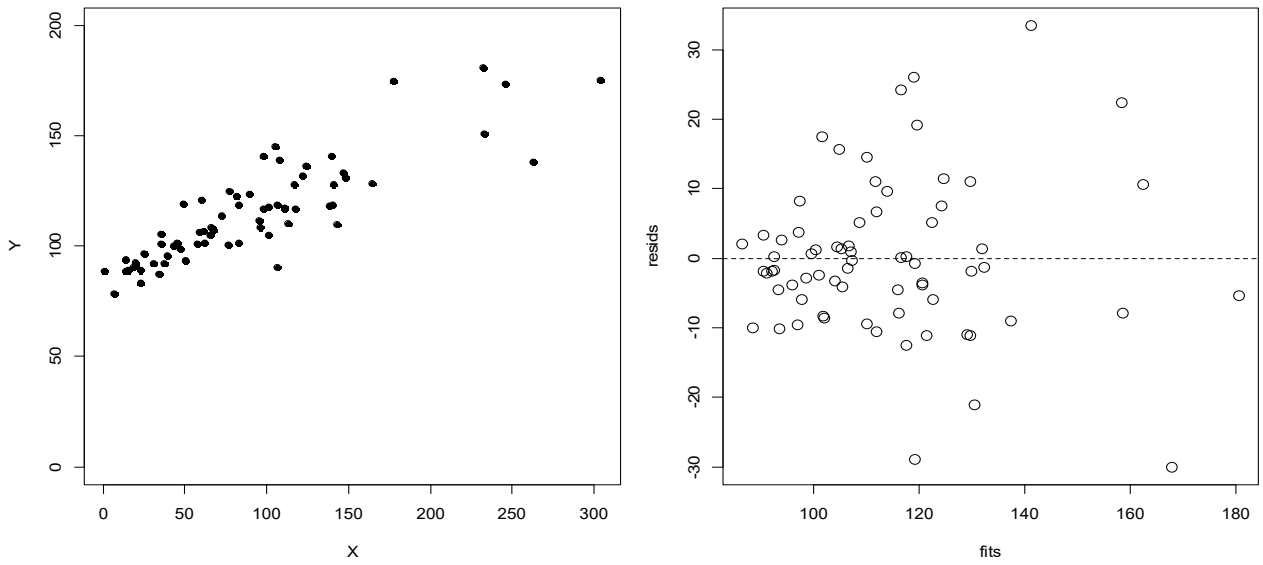
5. [1.5 pts] En base a la mida de la pàgina, quina és la capacitat de predicció del temps total del model anterior? Expliqueu com l'heu deduït. Com es coneix a l'indicador que heu utilitzat i què mesura?

L'indicador apropiat per mesurar la capacitat de predicció d'una variable per una altra és el coeficient de determinació R^2, que mesura la proporció de variabilitat de la variable resposta que es pot explicar per el model lineal que les relaciona. El coeficient R^2 es pot calcular elevant la correlació de les variables al quadrat. Aquí:

$r_{XY} = S_{X,Y} / (s_X s_Y) = \frac{1365,6}{\sqrt{4564 \cdot 464,17}} = 0,9382$

$R^2 = 0,9382^2 = 0,8803$

El 88% de la variabilitat observada als temps mesurats es pot associar a la mida de les pàgines descarregades. El 12% restant correspondria a variabilitat d'origen desconegut (possiblement, soroll aleatori).



Hem replicat l'estudi amb moltes més dades, els resultats del qual es mostren a les figures superiors.

6. [1 pt] Expliqueu breument cada un dels dos gràfics.

A l'esquerra tenim el diagrama de les variables: a l'eix X la mida de les pàgines, a l'eix Y el temps mesurat (ho sabem perquè veiem que a valors de la X propers a 0 tenim valors de l Y propers a 100, tal com hem vist als apartats anteriors). També ens mostra una relació positiva bastant forta (compatible amb la correlació trobada).

A la dreta tenim el diagrama dels residus (resids) front als valors ajustats (fits), que són els valors predits amb la recta de regressió.

7. [1 pt] Amb l'ajut dels mateixos, valideu el model lineal aplicat a aquest cas: quines premisses es podrien valorar? Quines semblen admissibles, i perquè (o perquè no)? Si veieu alguna que no admetríeu, comenteu les possibles causes i solucions.

Linealitat: es pot acceptar sense problemes una relació lineal de les variables. A l'esquerra el núvol és recte i a la dreta no es veuen distorsions, sinó que es simètric respecte la línia horitzontal.

Homoscedasticitat: no està clar que la variància residual sigui constant, més aviat sembla evident que no ho és. En els dos gràfics es veu que els punts del costat esquerre estan més concentrats que els del altre costat. Això té una explicació senzilla, i es que el temps de descàrrega pot ser més variable a mesura que la pàgina és més gran i, per tant, la pertorbació aleatòria que afecta a les observacions no és independent de la mida, en el sentit que la variància del soroll augmenta quan la mida augmenta. La solució no és simple, si no hi hagués un terme independent gran es podria transformar les dades amb logaritmes, però en aquest cas no funcionaria be.

Normalitat: es podria admetre, no tenim histograma ni QQ-plot però al gràfic dels residus es veuen els punts dispersos de forma simètrica.

Independència: aquests gràfics no ens poden donar cap informació.