NOM:_____COGNOM:_____

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

Problema 1 (Bloc C)

Nota: el separador decimal en tot l'exercici és el punt (".")

El *Cyber Monday* és un dilluns on les empreses que tenen servei de venta online fan descomptes dels seus productes. S'han recollit dades de preus de productes de 2 grans tendes ("Amazones" i "Market") abans i durant el *Cyber Monday* (A i C respectivament). Algunes dades de 2 productes (mòbils i ordinadors portàtils) de gama intermèdia són:

			(A) Preus abans Cyber Monday		(C) Preus Cyber Monday	
Tenda	**Producte**	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$
Market	mòbil	**12**	**5395**	**2622825**	**5214**	**2455166**
Amazones	mòbil	**12**	**4631**	**1963139**	**4568**	**1929330**
Market	ordinador	**12**	**5925**	**3174325**	**5782**	**3085308**
Amazones	ordinador	**12**	**6273**	**3534247**	**5979**	**3266463**

		(A) Preus abans Cyber Monday		(C) Preus Cyber Monday	
Tenda	N	$\sum A_i$	$\sum(A_i^2)$	$\sum C_i$	$\sum(C_i^2)$
Market	**24**	**11320**	**5797150**	**10996**	**5540474**
Amazones	**24**	**10904**	**5497386**	**10547**	**5195793**

1.- (0.5 punts) Indiqueu els valors a la taula de la dreta per a les dues tendes (sense distingir per tipus de productes)

2.- (1.5 punts) Doneu una estimació puntual de la mitjana de **preus al Cyber Monday** a la tenda ***Market*** (indistintament que siguin mòbils o ordindors). Quin és l'error estàndard d'aquesta estimació?

$$\bar{C}_M = \frac{10996}{24} = 458.17 \ €$$

$$s_M^2 = \frac{(5540474) - \frac{(10996)^2}{24}}{23} = 147.8^2 \rightarrow s_M = 147.8 \ €$$

$$SE_M = \frac{147.8}{\sqrt{24}} = 30.2 \ €$$

3.- (1 punt) Calculeu un interval de confiança (IC) del 99% per a la mitjana de **preus al Cyber Monday** a la tenda ***Market*** (indistintament que siguin mòbils o portàtils) suposant una desviació poblacional de 150 euros

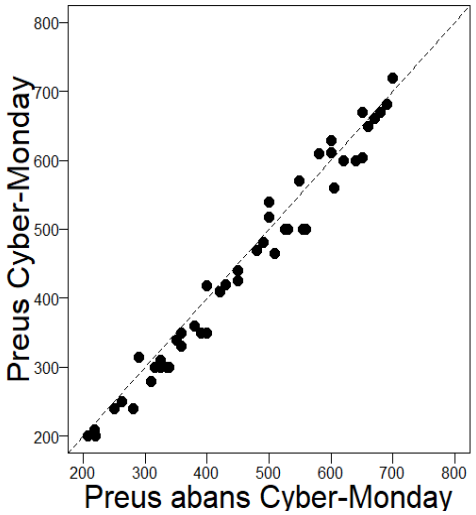
$$IC(\mu_M, 99\%) = \bar{C}_M \pm z_{0.995} \cdot \frac{150}{\sqrt{24}} = 458.17 \pm 2.576 \cdot \frac{150}{\sqrt{24}} = [\mathbf{379.3}, \mathbf{537.04}]$$

4.- (1 punt) L'IC del 95% de la mitjana dels **preus de mòbils al Cyber Monday** amb σ desconeguda a la tenda ***Amazones*** és [297, 464] amb els 12 productes recollits (per tant l'amplada és 167). Amb una mostra de 24 productes de la mateixa tenda i mateixos indicadors, quin dels següents 4 intervals de confiança del 95% seria més plausible obtenir? Per què?

a) [310, 451] b) [340, 421] c) [322, 439] d) [332, 449]

L'amplada de l'interval que ens donen és 167. Si doblem la mostra i assumint que la desviació roman més o menys constant, esperem que l'interval de confiança tingui una amplada de $167/\sqrt{2} \approx 118$.
Les amplades dels intervals són 141, 83, 117 i 117 (aquest últim amb el centre molt diferent).
Per tant, l'interval més plausible és el c)

5.- (1.5 punts) El següent gràfic representa els preus al Cyber Monday versus els preus abans del Cyber Monday pels 48 productes. La línia puntejada representa la recta bisectriu ($y = x$). Observant el gràfic, indiqueu la proporció de productes que pugen de preu el Cyber Monday i calculeu i interpreteu un IC95% d'aquesta proporció



Com que hi ha **10 punts per sobre de la bisectriu**, 10 productes pugen de preu

$$\hat{\pi} = p = \frac{10}{48} = \mathbf{0.21}$$

$$IC(\pi, 95\%) = p \pm z_{0.975} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0.21 \pm 1.96 \cdot \sqrt{\frac{0.21 \cdot 0.79}{48}} =$$

$$= 0.21 \pm 1.96 \cdot 0.06 = 0.21 \pm 0.12 = [\mathbf{0.09}, \mathbf{0.33}]$$

Amb 95% de confiança es pot esperar entre un 9% i un 33% de productes de la categoria estudiada, que pugin de preu pel Cyber Monday

6.- (1 punt) Argumenteu quines característiques han de tenir les dades per tal de calcular un interval de confiança per a la diferència de mitjanes de preus entre les dues tendes (*Amazones* versus *Market*) com a mostres aparellades o independents

Només tindria sentit tractar-les com a aparellades si els 24 productes mostrejats a cada tenda fossin els mateixos. Llavors els podríem aparellar per model i calcular l'interval de confiança de la diferència aparellada (en aquest cas la diferència cal que sigui normal).

En cas contrari, no es podria fer i s'hauria de calcular un interval de confiança de la diferència de mitjanes per mostres independents (en aquest cas les dues mostres han de ser normals).

7.- (1 punt) A partir només de la següent sortida de R, calculeu un interval de confiança del 90% per a la diferència de preus del Cyber-Monday menys els preus abans del Cyber Monday.

```
Paired t-test
data:  preus_cyber(C) and preus_abans(A)

t = -4.1278, df = 47, p-value = 0.0001487
alternative hypothesis: true mean difference
is not equal to 0

95 percent confidence interval:
-21.102015  -7.272985
sample estimates:
mean difference  -14.1875
```

Hem de deduir $\bar{d}$ i s_d de la sortida:

$\bar{d} = -14.1875$

Emprant l'amplada de l'interval es pot trobar s_d i se:

(21.102015-7.272985)/2 -> 6.91 $6.91 = t_{47,0.975} \cdot \frac{s_d}{\sqrt{48}} = 2.012 \cdot \frac{s_d}{\sqrt{48}}$

$s_d = \frac{\sqrt{48} \cdot 6.91}{2.012} = 23.8$ se = $\frac{23.8}{\sqrt{48}} = 3.44$

O bé amb l'estadístic calcular se: -4.1278 = -14.1875/se se = 3.44

$$IC(\mu_d, 90\%) = -14.1875 \pm t_{47,0.95} \cdot se$$
$$= -14.1875 \pm 1.678 \cdot 3.44 = [\mathbf{-19.96}, \mathbf{-8.42}]$$

8.- (1.5 punts) Calculeu un interval de confiança del 95% per a la diferència de mitjanes de **preus abans del Cyber Monday entre mòbils (m) i ordinadors portàtils (o)** a la tenda ***Amazones*** assumint variàncies iguals. Interpreteu l'interval, argumenteu si podem concloure que els mòbils i els portàtils tenen diferent preu (quantificant la diferència si en tenen)

$$\bar{m}_A = \frac{4631}{12} = 385.92 \qquad s_m^2 = \frac{1963139 - \frac{(4631)^2}{12}}{11} = 126.48^2 \quad \rightarrow \quad s_m = 126.48$$

$$\bar{o}_A = \frac{6273}{12} = 522.75 \qquad s_O^2 = \frac{3534247 - \frac{(6273)^2}{12}}{11} = 152.27^2 \quad \rightarrow \quad s_O = 152.27$$

$$S_{pooled}^2 = \frac{11 \cdot 126.48^2 + 11 \cdot 152.27^2}{22} = 19591.6 \quad \rightarrow \quad S_{pooled} = 139.97 \quad \rightarrow \quad se = 139.97 \cdot \sqrt{\frac{1}{12} + \frac{1}{12}} = 57.14$$

$$IC(\mu_m - \mu_O, 95\%) = (\bar{m}_A - \bar{o}_A) \pm t_{22,0.975} \cdot s_{pooled} \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_O}} = (385.92 - 522.75) \pm 2.074 \cdot 139.97 \cdot \sqrt{\frac{1}{12} + \frac{1}{12}} =$$
$$= (-136.83) \pm 57.14 = (-136.83) \pm 118.51 = [\mathbf{-255.34}, \mathbf{-18.33}]$$

Amb un 95% de confiança la diferència de mitjanes de preus està entre 18.33 i 255.34 €.

Com que el 0 no està inclòs a l'interval tenim evidència per dir que els ordinadors portàtils són en mitjana més cars que els mòbils a Amazones abans del Cyber Monday.

Abans del *Cyber Monday,* a Amazones els ordinadors portàtils són en mitjana entre 18.33 i 255.34 € més cars que els mòbils.

9.- (1 punt) Troba la grandària mostral necessària per a que l'anterior interval de confiança del 95% tingui una amplada de 150€ assumint que la σ comuna als 2 grups és 100€ i que volem el mateix nombre de portàtils que de mòbils.

$$Z_{0.975} \cdot \sigma \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_O}} = 75 \xrightarrow[n_M = n_O = n]{} 1.96 \cdot 100 \cdot \sqrt{\frac{2}{n}} = 75 \rightarrow n = \frac{2}{\left(\frac{75}{1.96 \cdot 100}\right)^2} = 13.66 \rightarrow$$

$$n = 14$$

Necessitaríem com a mínim 14 productes de cada tipus enlloc de 12 per assolir aquesta precisió

qnorm(0.900) = 1.282	qnorm(0.975) = 1.960	qt(0.950,22)=1.717	qt(0.975,22)=2.074	qt(0.950,46)=1.679	qt(0.975,46)=2.013
qnorm(0.925) = 1.440	qnorm(0.990) = 2.326	qt(0.950,23)=1.714	qt(0.975,23)=2.069	qt(0.950,47)=1.678	qt(0.975,47)=2.012
qnorm(0.950) = 1.645	qnorm(0.995) = 2.576	qt(0.950,24)=1.711	qt(0.975,24)=2.064	qt(0.950,48)=1.677	qt(0.975,48)=2.011

BD NOM:_____COGNOMS:_____

Model de avaluació i explicació docent. Les interpretacions genèriques ("anumèriques") → ½ p

S'analitza si la diferència en el temps d'execució [ms] entre C i C++ a l'hora de trobar els valors propis d'una matriu depèn de la mida de la matriu (aquí, les dimensions seran n×n on n és 10, 20, ..., 190, 200). Cada operació amb una matriu a l'atzar d'una mida determinada es fa amb cada llenguatge, obtenint-ne els temps d'execució (**C** i **+** respectivament). Per conveniència, treballarem amb el logaritme del temps, i la seva diferència "D = ln(+) − ln(C)" equival al logaritme del rati: "ln(+/C)". A sota veieu les sortides de R per a dos models que hem provat (dades$Y és el logaritme dels temps; dades$L és llenguatge: C o +).

```
Call:
lm(formula = D ~ 1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.41595 -0.34077 -0.05871  0.24805  0.77256
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)  1.23682   0.08229   15.03 5.32e-12
Residual standard error: 0.368 on 19 degrees of
freedom
```

```
Call:
lm(formula = Y ~ L, data = dades)
Residuals:
     Min       1Q   Median       3Q      Max
-0.82690 -0.38031 -0.04163  0.33268  1.30651
Coefficients:
            Estimate Std.Error  t value Pr(>|t|)
(Intercept)   0.5784     0.1169    4.950 1.55e-05
LC           -1.2368     0.1653   -7.484 5.50e-09
Residual standard error: 0.5226 on 38 degrees of freedom.
Multiple R-squared:  0.5958, Adjusted R-squared:  0.5852
F-statistic: 56.01 on 1 and 38 DF,  p-value: 5.496e-09
```

A) Descriviu quins són els dissenys i les premisses que hi ha darrera de cada opció (**1**, esquerra; **2**, dreta).

1) disseny aparellat, tenim per a cada matriu el temps obtingut amb C, el temps obtingut amb C++ i treballem amb la diferència. Les principals premisses són m.a.s; independència entre matrius; i normalitat de la Diferència (no cal per als temps de cada opció)
2) disseny de dues mostres independents. Premisses: les mostres són m.a.s., Y per C és independent de Y per C++ (no s'acompleix), mateixa variància amb C i amb C++, distribució normal amb C i amb C++, per separat.
[Malament si confon model amb disseny: el disseny és com s'obtenen les dades; el model és la construcció que ens agradaria que expliqués les dades.
Per cert: anomenarem *model lineal* als models que tenen predictors quantitatius expressats en forma lineal. Si el predictor és C++ o C no pot intervenir linealment, no podem dir "si s'incrementa el predictor un 1 ...". Si el model no té predictors, difícilment es pot dir *lineal*]

B) Digueu si són dues opcions vàlides; si ho són, justifiqueu la resposta; si no ho són, digueu quina és l'opció apropiada.

Només és vàlida l'opció 1, perquè la resposta no s'ha recollit de forma independent (amb matrius diferents per a C i per a C++) sinó de forma aparellada.

C) Independentment de la resposta anterior, interpreteu (no llegiu) el resultat obtingut amb l'opció 1.

Estima la mitjana del logaritme del rati +/C com 1.24 (error tipus=0.08):
en mitjana, C++ triga 3.46 vegades més que C
[És molt greu no saber interpretar correctament el resultat, i dir que C++ és més ràpid, o més eficient, que C. Tampoc és admissible dir que les unitats de l'estimació 1.23682, o de la seva transformada 3.46, són ms.]

D) El mateix amb l'opció 2 (a aquesta pregunta i a l'anterior, no cal considerar aspectes menors).

La mitjana estimada per a log(temps C++) és 0.5784, error tipus=0.12; aproximadament, 1.78 ms en mitjana. Però quan el llenguatge és C, l'estimació decreix en -1.24 (error tipus=0.17).
Això vol dir que, en mitjana (geomètrica), el temps amb C és exp(-1.24)=0.29 vegades **més petit**.

E) Què representa el "Residual Standard error" en cada cas?

1) S=0.368 és la desviació tipus estimada de les diferències entre log(+) i log(C)
2) S=0.5226 és la desviació tipus estimada en comú (pooled) dels logaritmes del temps per cada llenguatge

La primera és més petita perquè, al ser per una mateixa matriu, les diferències entre matrius ja no hi contribueixen al soroll .

F) Expliqueu de què ens informa el valor de "Multiple R-squared". Perquè una opció inclou el coeficient de determinació i l'altra no?

Informa de la proporció de la variabilitat de la variable resposta que poden explicar els factors predictors.

Perquè el model 2 inclou un predictor (el llenguatge, C++/C), i el model 1 no en té cap.

G) Calculeu un IC al 95% de confiança per al paràmetre del model 1, i explica què ens diu respecte als dos llenguatges.

IC = 1.23682 ± $t_{0.975,\ 19}$ 0.08229 = (1.064, 1.409). Com és escala logarítmica:

IC(rati +/C) = (2.90, 4.09): en mitjana (geomètrica), C++ triga 3.46 (entre 2.90 i 4.09) vegades més que C

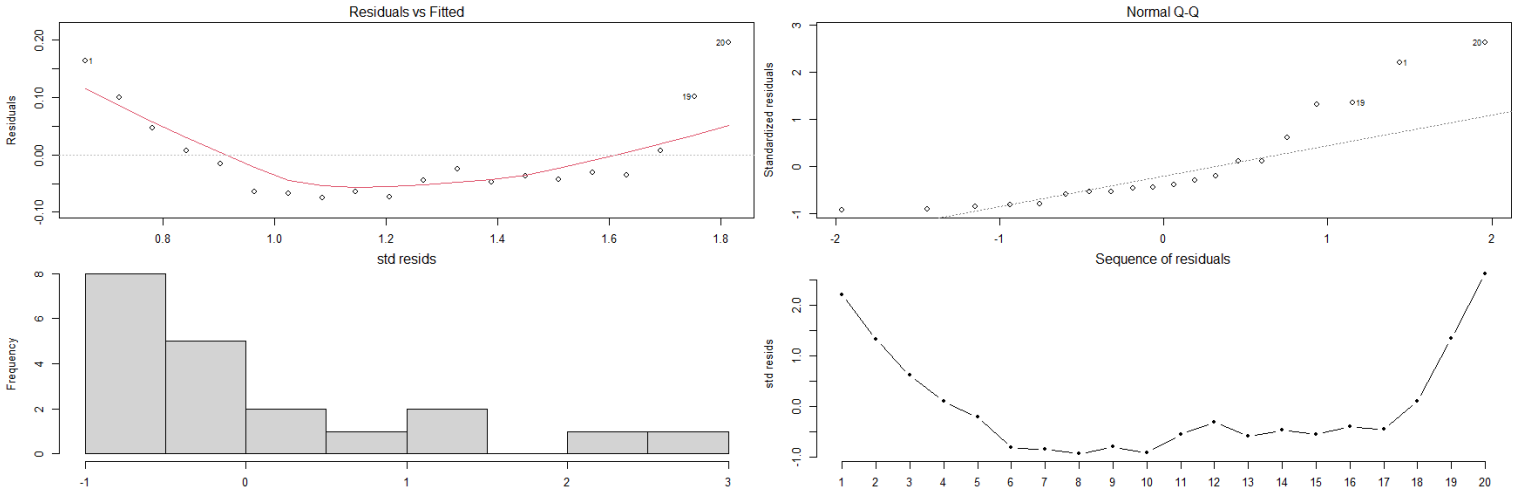
H) Seguidament teniu un fragment de la sortida del model "D en funció de la mida (N)". Comenteu què representen els dos valors de la columna "Estimate", d'acord amb el model estadístic que hem aplicat en aquesta ocasió.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.5993784  0.0383419   15.63 6.45e-12
N           0.0060709  0.0003201   18.97 2.40e-13
Residual standard error: 0.08254 on 18 degrees of freedom
Multiple R-squared:  0.9523,    Adjusted R-squared:  0.9497
```

El model estadístic és D = log(+/C) = 0.60 + 0.006 N. Equival a +/C = $e^{0.60}$ $e^{0.006\ N}$ = $e^{0.60}$ $(e^{0.006})^N$ = 1.822 1.006^N
0.60 és el terme independent. Diríem que *com a base* (per a matrius N=0), C++ triga 1.822 vegades més.
Si la matriu és més gran, la diferència augmenta. Concretament, per a un increment de 1 en N, el rati +/C augmenta un 0.6%.

I) Segons la sortida anterior, utilitzeu un interval de confiança 95% per a dir com afecta la mida de la matriu en la relació entre C++ i C.

0.0060709 ± $t_{0.975,\ 18}$ 0.0003201 = (0.005398395, 0.006743405). L'increment relatiu està entre 0.54% i 0.68% per qualsevol increment en N de 1 (*qualsevol*, perquè suposadament aplica un model lineal: després es veu a l'anàlisi de les premisses que realment no és així).



J) A la vista d'aquests gràfics, com us sembla que ha estat el compliment de les premisses del model emprat? Justifiqueu en detall la resposta.

Clarament, la premissa que no és correcta és la linealitat, perquè els residus dibuixen una forma corba, senyal de que el rati no varia només linealment respecte n. Per tant, el model es millorable.

Com els residus no són únicament variació aleatòria, sinó que depenen no linealment de la mida, és difícil valorar si els residus són homoscedàstics, o segueixen el model Normal (sembla que no) o són independents.