NOM:_____COGNOM:_____

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

Problema 1 (B1-B2)

1. En una empresa determinada, el 50% dels documents estan escrits en anglès, 30% en català i 20% en castellà. A partir de les dades recollides sabem que el 40% del documents escrits en anglès tenen més de 15 pàgines; el 20% del documents escrits en català tenen més de 15 pàgines; i el 20% dels documents escrits en castellà tenen més de 15 pàgines.

   1. a) Quina és la probabilitat d'escollir a l'atzar un document de més de 15 pàgines? (1 p)

   Tenim que P(anglès)= 0.5; P(català) =0.3 i P(castellà)=0.2

   Anomenem A='Un document escollit a l'atzar té més de 15 pàgines'
   També sabem que P(A| anglès) = 0.4 ; P(A| català) = 0.2; P(A| castellà) = 0.2

   Aleshores volem calcular P(A)= P(anglès)·P(A| anglès) + P(català)· P(A| català) + P(castellà)· P(A| castellà) = 0.5 · 0.4 + 0.3· 0.2 + 0.2 · 0.2 = 0.3

   Per tant, el 30% dels documents d'aquesta empresa tenen més de 15 pàgines.

   1. b) Hem escollit un document a l'atzar i observem que té més de 15 pàgines, quina és la probabilitat que hagi estat escrit en castellà? (0.5 p)

   $P(\text{castellà} \mid A) = \frac{P(\text{castellà} \cap A)}{P(A)} = \frac{P(A|\text{castellà})\cdot P(\text{castellà})}{P(A)} = \frac{02\cdot 0.2}{0.3} = 0.1333.$

2. Aquesta mateixa empresa també ha estudiat el nombre de fallades de hardware en un sistema informàtic. Per fer-ho estudien el nombre de fallades que ocorren en una setmana. Fan l'estudi durant les 52 setmanes de l'any. A partir de l'estudi troben que no ha ocorregut cap fallada en 8 setmanes de l'any. Aquesta informació i la de la resta de l'estudi la podeu trobar a la següent taula:

Nombre de fallades	0	1	2	3	4	5
Nombre de setmanes	8	13	15	10	5	1

   2a) Indiqueu la funció de probabilitat del nombre de fallades en una setmana: (1 p)

Nombre de fallades	0	1	2	3	4	5
Probabilitat	0.1538	0.25	0.2885	0.1923	0.0962	0.0192

   2b) Indiqueu la funció de distribució del nombre de fallades en una setmana: (0.5 p)

Nombre de fallades	0	1	2	3	4	5
Probabilitat acumulada	0.1538	0.4038	0.6923	0.8846	0.9808	1

   2c) Quina és l'esperança d'X="Nombre de fallades en una setmana" (1 p)

   $E(X)= 0 \cdot 0.1538 + 1\cdot 0.25 + 2\cdot 0.2885 + 3\cdot 0.1923 + 4\cdot 0.0962 + 5\cdot 0.0192 = 1.8847$

   2d) Calcula la variància i la desviació típica d'X. (1 p)

   $E(X^2)= 0^2 \cdot 0.1538 + 1^2\cdot 0.25 + 2^2\cdot 0.2885 + 3^2\cdot 0.1923 + 4^2\cdot 0.0962 + 5^2\cdot 0.0192 = 5.1539$

$Var(X) = E(X^2) - E(X)^2 = 5.1539 - 1.8847^2 = 1.6018$

$\sigma_X = 1.2656$

2e) En la mateixa empresa han estudiat els errors d'una aplicació que corre en el mateix sistema. Denotem per Y el nombre d'errors de l'aplicació. A partir d'un estudi s'ha trobat que el nombre d'errors de l'aplicació està relacionat amb els errors del sistema mitjançant la funció $Y = 5 \cdot X + 3$.

e1) Calcula l'esperança d'Y ((0.5 p)

$E(Y) = E(5X + 3) = 5 \cdot E(X) + 3 = 5 \cdot 1.8847 + 3 = 12.4235$

e2) Calcula la desviació típica d'Y (0.5 p)

$Var(Y) = Var(5X + 3) = 25 \cdot Var(X) = 25 \cdot 1.6018 = 40.045$

$\sigma_Y = 6.3986$

3. En un laboratori mesuren el corrent en un circuit en ampers. A causa de diferents factors aleatoris, la mesura de Z varia. Els estudis realitzats indiquen que el corrent varia seguint la següent funció:

$$f(z) = \begin{cases} 0.025z + b, 1 < z < 5 \\ 0 \quad\quad\quad , \ altrament \end{cases}$$

3a) Troba b de manera que f(z) sigui una funció densitat i representa –la (2 p)

En primer lloc mirarem que és compleixi que $\int_{-\infty}^{+\infty} f(x)dx = 1$. En aquest cas, tenim que $\int_1^5 (0.025z + b)dz = [\frac{0.025z^2}{2} + bz]_1^5 = (0.3125 + 5b - 0.0125 - b) = 0.3 + 4b$

Com que volem que $0.3 + 4b = 1$, aleshores $b = 0.175$

Notem que la funció $f(z) = 0.025z + 0.175$ assoleix sempre valors positius en (1,5)

[També es podria resoldre per àrea de trapezi]

3b) Calcula la probabilitat que la mesura registrada sigui menor que 3. (1 p)

$P(Z \leq 3) = \int_1^3 (0.025z + 0.175)dz = [\frac{0.025z^2}{2} + 0.175z]_1^5 = 0.45$

3c) Calcula l'esperança de les mesures registrades. (1 p)

$\int_1^5 (0.025z + 0.175) \cdot z \, dz = [\frac{0.025z^3}{3} + \frac{0.175z^2}{2}]_1^5 = 3.1333$

NOM:_____COGNOMS:_____

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

Problema 2 (B3-B4)

En una granja ecològica asseguren que la distribució del pes en grams d'un ou és Normal amb esperança 60 grams i variància 16 grams^2.

Quina és la probabilitat que un ou pesi entre 55 i 65 grams?

P(55<Pes<65) = P(Pes<65) – P(Pes<55) = P(Z<1.25) – P(Z<-1.25) = 2P(Z<1.25) -1 = 2*0.8944 -1 = **0.7888**

Quina és la probabilitat que un ou pesi menys de 55.8 grams?

P(Pes<55.8) = P(Z<-1.05) = 1 – 0.8531 = **0.1469** (**0.15**)

Si considerem una dotzena d'aquests ous, quina és la probabilitat de trobar-ne 6 que pesin menys de 55.8 grams? (definiu prèviament la variable nombre d'ous en una dotzena que pesin menys de 55.8 grams)

Num és Bin(n=12, p=0.15)

P(Num=6) = P(Num<=6) – P(Num<=5) = 0.9993 – 0.9954 = **0.0039**

Quina variable representaria el nombre d'ous fins trobar-ne un de menys de 55.8 grams i quina és la probabilitat que sigui el primer

NumFins és Geom(p=0.15)

P(NumFins=1) = $0.15^1 * 0.85^0$ = **0.15**

Quina variable representaria el pes total d'una dotzena d'ous i quina és la probabilitat que aquest pes total sigui inferior a 720 grams?

PesTotal és N(mu=12*60=720, sigma=4*sqrt(12)=13.86)

P(PesTotal<720) = P(Z<0) = **0.50**

A la granja han decidit recollir dades del pes dels ous per tenir evidències quantitatives dels valors que asseguren que compleixen. Per això anoten el pes d'una dotzena d'ous i obtenen els següents valors:

pes <- c(64, 58, 54, 53, 62, 55, 56, 60, 59, 54, 63, 63)

$\sum_{i=1}^{12} pes_i = 701$ $\sum_{i=1}^{12} pes_i^2 = 41125$

Calculeu una estimació puntual de la mitjana i de la desviació poblacionals del pes dels ous

Estimació esperança (mitjana mostral) 701/12 = **58.42 (=m)**

Estimació desviació (desviació mostral) = sqrt((41125-(701^2/12))/11) = sqrt(15.9) = **3.99 (=s)**

Calculeu i interpreteu un interval amb confiança del 95% de la mitjana poblacional

$m \pm t_{11,0.975}$ s/sqrt(12) = 58.42 ± 2.201 3.99/sqrt(12) = **[55.88 60.95]**

Amb un 95% de confiança (i un error del 5%) el valor de l'esperança del pes està entre 55.88 gr i 60.95 gr

Calculeu i interpreteu un interval amb confiança del 95% de la desviació poblacional

s^2 * 11 / qchisq(0.975,11) = 15.9 * 11 / 21.920 = 7.98

s^2 * 11 / qchisq(0.025,11) = 15.9 * 11 / 3.816 = 45.83

sqrt(7.98) = **2.82**

sqrt(45.83) = **6.77**

Amb un 95% de confiança (i un error del 5%) el valor de la desviació del pes està entre 2.82 gr i 6.77 gr

Contrasteu si la mitjana poblacional del pes dels ous és 60 o no amb un risc del 5%. Indiqueu les hipòtesis, el càlcul de l'estadístic i justifiqueu la conclusió respecte les hipòtesis

H_0: μ =60 (hipòtesis conservadora)

H_1 : μ <>60

Estàndard error = se = s /(sqrt(12)) = 3.99/sqrt(12) = **1.15**

Estadístic = (58.42-60)/se = **-1.37**

Punts crítics: $t_{11,0.975}$ = 2.201 i $t_{11,0.025}$ = -2.201

Com que l'estadístic està entre els valors dels punts crítics indica que està a la zona d'acceptació. Per tant no hi ha evidència per rebutjar la hipòtesis nul·la, i és raonable creure que μ és 60 gr

Relacioneu els resultats de l'interval de confiança i de la prova d'hipòtesis anteriors

En l'interval de confiança hem vist que amb un 95% de confiança (i un error del 5%) el valor de l'esperança del pes està entre 55.88 gr i 60.95 gr. Per tant 60 gr és un valor dins aquest interval. En la prova d'hipòtesis hem vist que no hi ha evidència per rebutjar la hipòtesis nul·la, i és raonable creure que μ és 60 gr

La relació és que les dues han de coincidir en quant a la conclusió perquè es basen en el mateix estadístic i amb el mateix nivell de confiança (i/o error)

Si el valor a prova pertany a l'interval de confiança, sempre de la prova d'hipòtesis corresponent es conclou que és raonable acceptar el valor a prova

SOLUCIÓ B5-B6

Un enginyer informàtic es qüestiona quin algorisme pot emprar per trobar un cert node "marcat" en un arbre binari. Disposa de dues opcions, el BFS (Breadth-first search) o el DFS (Depth-first search), i voldria escollir el que en terme mitjà és més ràpid. Per prendre una decisió, utilitza un generador d'arbres binaris i marca un node aleatòriament. Amb aquests arbres, compta el nombre de nodes explorats per l'algorisme fins trobar el node marcat.

1. Si es sospita que, per a arbres d'una alçada determinada, quan el BFS triga molt a trobar el node, el DFS normalment és més ràpid, i viceversa: quin disseny us sembla més adient, el de dues mostres independents o el de mostres aparellades? Raoneu la resposta.

Si la resposta amb BFS augmenta quan baixa a DFS, i viceversa, és que la correlació entre BFS i DFS és negativa, i per tant un disseny aparellat no és adequat perquè la desviació tipus de la diferència entre nombre de nodes seria major que la desviació de qualsevol algorisme. Perquè sigui eficient, la correlació ha de ser positiva (i llavors la diferència disminuirà la variabilitat).
Moltes respostes donen per segur que el disseny aparellat fa disminuir sempre la desviació de la diferència, i això depèn de la relació entre les variables (que en aquest cas s'ha mostrat que és inversa).

2. L'enginyer ha fet una prova amb dues mostres d'alçada 15, cadascuna amb els seus arbres (14 arbres en cada grup). Les mitjanes corresponents han estat 23920 i 31091 nodes, DFS i BFS respectivament. Les desviacions tipus són 17895.36 i 21749.15. Amb aquesta prova es pot confirmar que hi ha diferències entre els dos mètodes? Feu una prova d'hipòtesis formal per justificar la resposta. (*aquesta pregunta no té cap relació amb la resposta de la pregunta anterior*)

S'entén que les mostres són independents (*cada mostra amb "els seus arbres"*). La variància pooled és 396634718 (19915.69^2), la diferència de mitjanes –7171; l'estadístic t = –0.95265 no és suficientment gran per rebutjar la hipòtesi nul·la d'igualtat de mitjanes: amb α=5%, el punt crític és $\pm t_{26,\,0.975} = \pm 2.056$. No es pot confirmar diferència entre mètodes.
No es pot suposar que les desviacions donades són poblacionals, llavors la prova es realitza amb l'estadístic t-Student amb 26 graus de llibertat.

3. És correcte dissenyar la prova com a unilateral perquè hem vist que el BFS ha trigat més en mitjana? Doneu arguments amb la resposta.

No és correcte, no podem dissenyar la prova basant-nos en les dades mostrals. Necessitem evidència prèvia que ens digui que és raonable pensar que, si no són iguals, un determinat és superior.

4. Posteriorment, l'enginyer amplia el seu experiment amb arbres d'alçada variable: decideix generar alçades d'entre 5 i 24 de forma uniforme, i utilitzar un arbre amb aquesta alçada per ambdós algorismes. En aquest disseny, com serà la correlació entre les respostes BFS i DFS? Per què?

Si tenim arbres petits, el nombre de nodes fins a trobar el node marcat serà baix, tant amb BFS com amb DFS, i si el arbre és gran el nombre de nodes serà gran amb els dos algorismes. La correlació serà positiva.

5. Observeu aquestes sortides de R:

```
        Paired t-test

data:  Z$n_DFS and Z$n_BFS
t = -1.365, df = 199, p-value = 0.1738
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -1153782.2   209852.8
sample estimates:
mean of the differences
              -471964.7
```

```
        Paired t-test

data:  log(Z$n_DFS) and log(Z$n_BFS)
t = -0.54263, df = 199, p-value = 0.588
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.245322  0.139445
sample estimates:
mean of the differences
            -0.05293849
```

- Comenteu breument quina prova s'està fent en cada cas.

Esquerra: prova aparellada de comparació de mitjanes del nombre de nodes entre DFS i BFS
Dreta: prova aparellada de comparació de mitjanes del logaritme del nombre de nodes entre DFS i BFS

- Calculeu la desviació tipus de la diferència que s'està emprant en cada cas.

Obtindrem el valor de la desviació tipus prenent el valor de l'estadístic i la mitjana de la diferència:
Esquerra: -1.365 = -471964.7 / (s/sqrt(200)), llavors s = 4889809
Dreta: -0.54263 = -0.05293849/(s/sqrt(200)), llavors s = 1.379694

- Com es pot interpretar l'interval (-0.2453, 0.1394)?

Amb una confiança del 95%, la mitjana del logaritme del rati DFS/BFS (o la mitjana de la diferencia de logaritmes) es troba entre -0.2453 i 0.1394. També es pot interpretar que la mitjana del rati es pot trobar entre 0.782 i 1.150.

- A quin resultat s'arriba en cadascú dels anàlisis? Què es pot dir sobre la comparació dels dos algorismes?

No podem refusar la igualtat de les mitjanes del nombre de nodes explorats pel DFS i pel BFS. Tampoc es pot refusar la igualtat de les mitjanes del logaritme del nombre de nodes explorats pel DFS i pel BFS, sembla que el rati mitjà d'aquests valors és compatible amb 1 i per tant semblen dos mètodes equiparables.

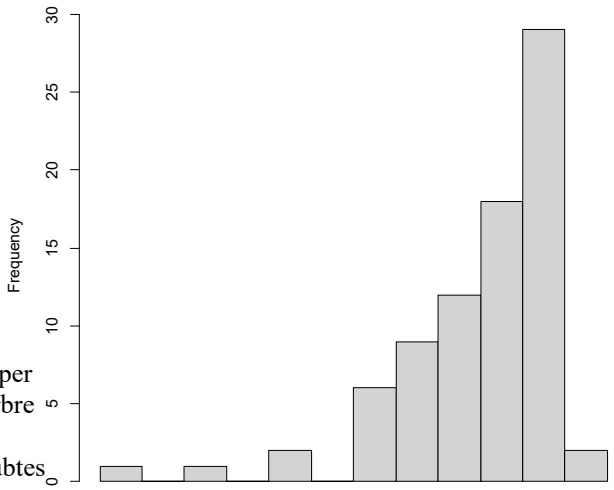
- Dels dos anàlisis de dalt, hi ha alguna opció millor que l'altra? Per què?

La opció de la esquerra és controvertida, perquè la diferència mitjana no és la mateixa si l'arbre és gran o és petit (o en tot cas la variabilitat de la diferència és molt depenent de l'alçada de l'arbre). És a dir, el paràmetre estimat (la mitjana de les diferències) és poc representatiu. En canvi, l'anàlisi del logaritme del rati possiblement no depèn del factor mida de l'arbre i sembla l'opció aconsellable.

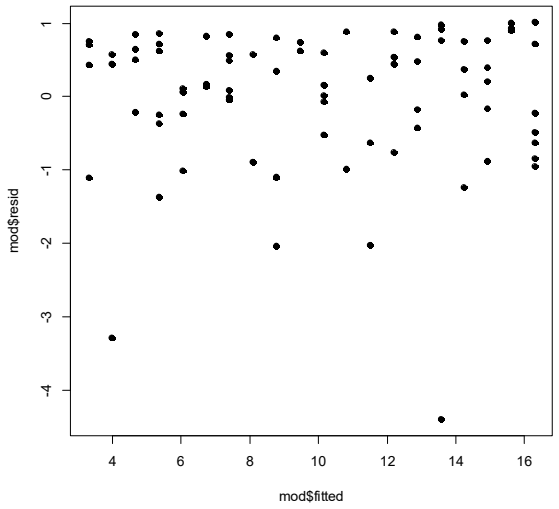
```
Residuals:
    Min      1Q  Median      3Q     Max
-4.4026 -0.4402  0.2295  0.7192  1.0144

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.11551    0.28713  -0.402    0.689
Altura       0.68359    0.01802  37.934   <2e-16 ***

Residual standard error: 0.9713 on 78 degrees of freedom
Multiple R-squared:  0.9486
```



Continuant amb el cas dels algorismes de cerca en arbres, hem investigat el DFS per modelar el logaritme del nombre de nodes recorreguts explicat por l'alçada del arbre (`Altura`) en un subconjunt de 80 casos. Amb aquesta transformació, la relació segueix una tendència lineal indiscutible. De totes formes, encara es plantegen dubtes respecte a les altres premisses del model.



Amb les sortides annexes, discutiu sobre el grau d'acompliment de les premisses amb aquestes dades (les vostres conclusions s'han de justificar amb elements concrets de les sortides).

Expliqueu què representen els resultats següents:

- −0.11551
- 0.28713
- 0.68359
- 0.01802
- 0.9713
- 0.9486

No es vàlid simplement prendre els mots que apareixen a la sortida.

Teòricament, si un arbre té un nivell més, el nombre de nodes es duplica i per tant el cost de trobar un determinat node marcat també. El logaritme de 2 és 0.6931472; comprova fent una prova d'hipòtesis si amb les dades d'aquest experiment la teoria és coherent o no. Interpreta el resultat trobat.

El nostre algorisme DFS ha d'explorar un determinat arbre d'alçada igual a 20. Amb el model anterior, feu una estimació puntual i una per interval de confiança 90% del nombre de nodes que pot recórrer fins a trobar el node marcat.

...grau d'acompliment de les premisses... A partir de l'histograma i també de la descriptiva dels residus veiem que aquests no són gaire Normals (el mínim és -4.4, i el màxim 1; la forma de la distribució és molt asimètrica cap a baix). La premissa de Normalitat no és assumible.

A partir del gràfic dels residus vs valors ajustats veiem que la dispersió és similar a qualsevol nivell de esquerra a dreta. També apreciem com hem dit abans que els valors tendeixen a agrupar-se a dalt i uns pocs prenen valors molt lluny del centre per la part de sota.

Es pot comprovar també amb aquests gràfic que la premissa de Linealitat s'acompleix perquè el núvol és pla. No tenim elements gràfics o numèrics per estudiar si les dades són realment independents.

Expliqueu què representen els resultats següents:

- −0.11551 estimació del terme independent de la recta de regressió (b_0)
- 0.28713 estimació de l'error tipus per al terme independent de la recta (S_{b0})
- 0.68359 estimació del pendent de la recta de regressió: un increment en l'alçada de 1 representa un increment mitjà en el logaritme del nombre de nodes explorats igual a 0.684 (b_1)
- 0.01802 estimació de l'error tipus per al pendent de la recta (S_{b1})
- 0.9713 desviació estimada dels errors o residus; és una desviació de quasi 1 al mesurar el logaritme del nombre de nodes explorats per a una alçada fixada. (S, o S_R)
- 0.9486 coeficient de determinació; el 95% de la variabilitat de la resposta s'atribueix a l'alçada del arbre, i el 5% és aleatori (R^2)

... prova d'hipòtesis...

H_0: $beta_1 = \log(2)$?

Estadístic de la prova: t = (0.68359 - log(2)) / 0.01802 = −0.53. No es pot refusar la hipòtesi nul·la. Es versemblant que la resposta es dupliqui (en mitjana) si l'alçada de l'arbre augmenta en 1.

Compte: no diem que ho hem demostrat, només que la mostra no contradiu un model teòric concret.

... una estimació puntual i una per interval de confiança 90%...

Amb l'equació de la recta estimada:

−0.11551 + 0.68359 × 20 = 13.55629; exp(13.55629) = 771652.8 nodes

Es tracta d'una estimació *individual*, perquè l'arbre és un qualsevol. Per tant, l'expressió és:

$$13.55629 \pm t_{78,0.95} \times 0.9713\sqrt{1 + \frac{1}{80} + \frac{(20-14.75)^2}{79 \times 6.064^2}} = 13.55629 \pm 1.665 \times 0.9713\sqrt{1.022} = (11.92138, 15.19120)$$

Per tant, en nombre de nodes la previsió per a un arbre és: (150449.5, 3957792.9), entre cent cinquanta mil i quatre milions aprox.

Error habitual: deixar el resultat en termes del logaritme (no expressa el nombre de nodes).