NOM:_____

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

Problema 1 (B1-B2)

Es volen estudiar característiques d'una eina de resolució de problemes online que s'usa en una assignatura. Els problemes de la col·lecció estan etiquetats com a fàcils (F) mitjans (M) o difícils (D). Al llarg dels quadrimestres s'ha recollit informació del seu funcionament en quant a si les execucions dels problemes un cop corregides obtenen una nota d'aprovat (5 o superior) o no, i del nombre d'execucions dels problemes que es realitzen i del temps que es triga en cada execució.
Per una part anem a estudiar si obtenir un aprovat (A) o no (NA) en l'execució d'un problema té a veure o no amb l'etiqueta de dificultat (hi ha un 50% de problemes fàcils i un 20% de difícils). Sabem que de les execucions de problemes fàcils un 80% obtenen nota aprovada, mentre que dels mitjans només un 50%. I sabem que en global les execucions aprovades són un 60%.

Indiqueu i justifiqueu les probabilitats de F, M i D. I les probabilitats de A i NA (1 pt)

P(F) = 0.5 P(D) = 0.20 P(M) = 1- (0.5+0.2) = 0.3

P(A) = 0.60 P(NA) = 1- 0.6 = 0.40

Calculeu la probabilitat que el problema sigui difícil i aprovat (1 pt)

P(D i A) = P(A) – (P(F i A) + P(M i A)) =
= 0.60 – (0.5 0.8 + 0.3 0.5) =
= 0.60 – (0.40 + 0.15) = 0.60 – 0.55 = **0.05**

Indiqueu l'arbre d'esdeveniments tenint en compte l'etiqueta i l'aprovat o no. Indiqueu el conjunt de resultats amb les seves probabilitats (1 pt)

P(F i A) = P(F) P(A|F) = 0.5 0.8 = **0.40**
P(F i NA) = P(F) P(NA|F) = 0.5 0.2 = **0.10**

P(M i A) = P(M) P(A|M) = 0.3 0.5 = **0.15**
P(M i NA) = P(M) P(NA|M) = 0.3 0.5 = **0.15**

P(D i A) = **0.05** (-> P(A|D) = 0.05 / 0.2 = 0.25)
P(D i NA) = 1 – (0.4+0.1+0.15+0.15+0.05) = **0.15** (-> P(NA|D) = 0.15 / 0.2 = 0.75)

Calculeu la probabilitat de que una execució aprovada sigui de problema fàcil (1 pt)

P(F|A) = P(F i A) / P(A) = 0.40 / 0.60 = 2/3 = 0.6667

Un estudiant afirma que l'etiqueta de dificultat no té res a veure amb la probabilitat d'aprovar-ne o no una execució. Indiqueu i justifiqueu si creieu que té raó o no (1 pt)

No té raó perquè no és independent l'etiqueta de dificultat de la probabilitat d'aprovar, ja que per exemple

P(A|F) = 0.8 és diferent de P(A|M) = 0.5
(en fàcils s'aproven 8 de cada 10 execucions i en canvi en mitjans només la meitat)

Per altra part anem a estudiar, només pels problemes fàcils i difícils, el nombre i el temps de les execucions.
En el cas del nombre d'execucions s'ha recollit (només pels problemes fàcils i pels difícils i pels casos de 1, 2 o 3 execucions) les següents proporcions

	Fàcils	Difícils
1 execució	0.30	0.20
2 execucions	0.10	0.07
3 execucions	0.20	0.13

Calculeu pels problemes fàcils la probabilitat de fer 1 execució. I la de fer-ne 2. I la de fer-ne 3 (1 pt)

P(1|F) = P(1 i F) / P(F) = 0.3 / 0.6 = ½ = 0.50
P(2|F) = P(2 i F) / P(F) = 0.1 / 0.6 = 1/6 = 0. 1667
P(3|F) = P(3 i F) / P(F) = 0.2 / 0.6 = 1/3 = 0.3333

Calculeu l'esperança i la variància del nombre d'execucions pels problemes fàcils (1 pt)

E(Execucions) = 1 ½ + 2 1/6 + 3 1/3 = 11/6 = 1.8333

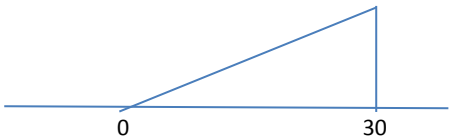
V(Execucions) = $(1-1.8333)^2$ ½ + $(2-1.8333)^2$ 1/6 + $(3-1.8333)^2$ 1/3 = 0.8056

Indiqueu i justifiqueu formalment si és independent o no l'etiqueta fàcil/difícil del fet de fer 1,2 o 3 execucions (1 pt)

Si que és independent perquè totes les probabilitats conjuntes coincideixen amb el producte de les marginals:
P(1 i F)=0.3=0.6 0.5 P(2 i F)=0.1=0.6 0.1667 P(3 i F)=0.2=0.6 0.3333
P(1 i D)=0.2=0.4 0.5 P(2 i D)=0.07=0.4 0.1667 P(3 i D)=0.13=0.4 0.3333

O bé perquè les probabilitats condicionades coincideixen:
P(1|F) = 0.50 P(2|F) = 0. 17 P(3|F) = 0.3
P(1|D)=0.2/0.4=0.5 P(2|D)=0.07/0.4=0. 17 P(3|F)=0.13/0.4=0.3

I finalment pel temps de les execucions de problemes fàcils i el dels difícils s'ha recollit la distribució de els execucions que han trigat entre 0 i 30 minuts, obtenint respectivament:



Indiqueu la funció de distribució de probabilitat i calculeu la probabilitat de trigar menys de 15 minuts en els dos casos (1pt)

(tempsF(x) = 1/30 0<=x<=30) (tempsD(x) = (1/450) x 0<=x<=30)

FtempsF(x) = (1/30) x 0<=x<=30 ($\int_{-\infty}^{\infty} \frac{1}{30} dx = \frac{1}{30} x$) FtempsD(x) = (1/900) x^2 0<=x<=30 ($\int_{-\infty}^{\infty} \frac{1}{450} x dx = \frac{1}{900} x2$)

P(tempsF < 15) = FtmepsF(15) = 15/30 = 0.50
P(tempsD < 15) = FtmepsD(15) = 15^2/900 = 0.25

Interpreteu i comenteu les diferències dels dos casos tenint en compte la distribució i els càlculs anteriors (1 pt)

En el cas dels fàcils la probabilitat és constant entre l'inici i 30 minuts. En canvi en els difícils la probabilitat va creixent, és a dir és més probable trigar cap a 30 minuts.

En el cas dels fàcils al cap de 15 minuts ja hi ha 50% de probabilitat d'haver acabat. En canvi en difícils al cap de 15 minuts hi ha només el 25% de probabilitat d'haver acabat. (A més el temps esperat en fàcils és 15 i en difícils és 20, geomètricament o fent integrals)

NOM:_____COGNOMS:_____

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

Problema 2 (B3-B4)

La tecnologia d'un model concret de discs d'estat sòlid de 512 GB consisteix en 8 sectors de 64 GB. Per la seva part un sector es composa d'un gran nombre de "pistes". Anomenarem X al nombre de pistes que al cap d'un any d'ús s'inutilitzen a un sector: assumiu que els errors són independents uns dels altres i que el nombre mitjà és de 2.25/any/sector.

1. Doneu el model de probabilitat que segueix la variable aleatòria X, justificant la resposta. Quina és la probabilitat que en un any un sector donat no presenti cap pista inutilitzada? (1pt)

   X segueix un model de probabilitat de Poisson: P(2.25) (que una pista estigui malament és un esdeveniment "rar", però hi ha moltes pistes a un sector, i l'error pot caure a una pista amb independència d'altres)

   Probabilitat cap pista malament en un sector: $P(X=0) = \exp(-2.25) = 0.1054$

2. Calculeu la probabilitat que un disc d'un any d'ús tingui més d'un sector amb totes les pistes correctes. (1pt)

   Variable que compta sectors amb totes les pistes correctes: $M \sim B(8, p)$

   És binomial perquè la probabilitat $p = P(X=0) = 0.1054$ és la mateixa a qualsevol sector, i aquests també són independents entre sí (es dedueix).

   La probabilitat demanada: $P(M > 1) = 1 - F_M(1)$. Es poden utilitzar les taules si prenem p = 0.1

   $1 - F_M(1) = 1 - 0.8131 = 0.1869$

3. Amb un error del 5%, quin és el màxim nombre de pistes inutilitzades que pot haver a un disc d'un any d'ús? (1pt)

   Per la variable que compta les pistes inutilitzades a un disc i un any: és la suma de 8 variables com X (compte: no 8 vegades X). Per tant, també segueix una distribució de Poisson.

   $X_D \sim P(8 \cdot 2.25) = P(18)$

   La pregunta és: per a què x $P(X_D > x) = 0.05$? Busquem a les taules a quin punt on $\lambda=18$ la funció de distribució és superior a 0.95. Llavors:

   $x = 25$

4. Si el disc ha funcionat 4 anys, digueu un model de probabilitat aproximat per el nombre de pistes inutilitzades, amb els paràmetres corresponents. Trobeu la probabilitat que un disc d'aquests presenti més de 100 pistes fora d'ús. (1pt)

   Ara la variable serà Poisson també:

   $X_{4D} \sim P(4 \cdot 8 \cdot 2.25) = P(72)$. El valor de la mitjana és tan gran que podem aproximar molt be aquesta distribució per un model Normal, amb esperança i variància 72. És a dir, la desviació tipus ha de ser 8.49 $= \sqrt{72}$.

   $P(X_{4D} > 100) = P(Z > (100 - 72)/8.49) = P(Z > 3.30) = 1\text{-}F_Z(3.3) = 0.0005$ [Taules]

   Aproximadament, 1 de cada 2000 discs.

5. Escolliu la resposta més encertada: "Un disc de cada ____ (a: 100; b: 2000; c: 20) presenta més de 50 pistes inútils al cap de 2 anys", i justifiqueu la resposta. (1pt)

   Encara que el temps és la meitat, i el nombre de pistes inutilitzades també sigui al menys la meitat que abans, la probabilitat no és la mateixa. Per un procés idèntic, arribaríem a que $P(X_{2D} > 50) = P(Z > (50 - 36)/6) = P(Z > 2.333)$

   Per tant, la resposta correcta és a): aproximadament un de cada 100 discs.

6. Amb els discs d'un altre fabricant, s'assumeix en principi que la variància de la variable Y: "nombre de pistes inutilitzades després de sis mesos d'ús" val 9, i es pretén estimar el valor esperat de la variable Y. Calculeu el nombre de discs que caldria observar per obtenir un interval de confiança al 90% amb una amplitud no major de 1. (1pt)

   L'amplitud de l'interval de confiança és 2 $z_{0.95}$ $\sigma/\sqrt{n}$ = 1

   Per tant, $\sqrt{n} = 2 \times 1.645 \times \sigma = 2 \times 1.645 \times 3 = 9.869$

   n = 98

7. Es decideix prendre una mostra de 15 observacions, mantenint per similitud amb el primer fabricant una suposada variància poblacional igual a 9. Entre els 15 discs, seguits durant mig any i completament independents uns dels altres, hi havia 160 pistes inutilitzades. Poseu a prova formalment si la mitjana de la variable Y pot ser igual a 9. Trobeu el p valor de la prova i expliqueu la conclusió. (1.5pt)

   H0: E(Y) = 9
   H1: E(Y) ≠ 9

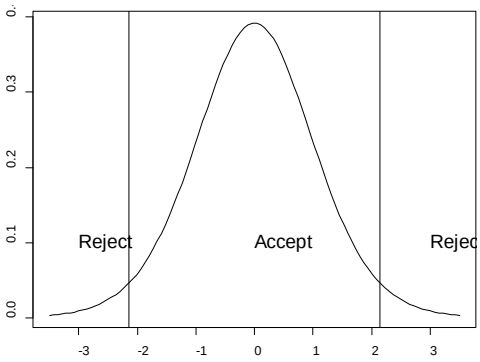
   Sota H0, $z = \frac{\bar{y}-E(Y)}{\sigma/\sqrt{15}} \sim N(0,1)$, admetent correctes les premisses de l'anàlisi.

   La mitjana mostral val $\bar{y} = \frac{160}{15} = 10.67$. L'estadístic val z = 2.15

   P valor = P(|Z| > |z|) = 2 P(Z > 2.15) = 2 (1-F_z(2.15)) = 2 (1-0.9842) = 0.0316

   La conclusió és que és poc versemblant que els discs d'aquest fabricant tinguin la taxa d'errors de 9 pistes cada mig any, ja que el p valor és bastant petit.

8. Repetiu la prova sense adoptar una variància per a la població. La variància mostral valia 9.53. No cal trobar el p valor, però representeu en la distribució de l'estadístic les zones de rebuig i acceptació. Quina és la conclusió? (1.5pt)



   H0: E(Y) = 9
   H1: E(Y) ≠ 9

   Sota H0, $t = \frac{\bar{y}-E(Y)}{s/\sqrt{15}} \sim t_{14}$, admetent correctes les premisses de l'anàlisi.

   Per a una t de Student amb 14 graus de llibertat i un test bilateral, la zona de acceptació es troba entre -2.145 i 2.145, amb un risc del 5%.

   L'estadístic val t = 2.09. Per tant, per poc no podem rebutjar la hipòtesi nul·la: hi ha indicis de que la taxa podria ser diferent de 9, però no són suficientment forts.

9. Per a l'anàlisi anterior, critiqueu les premisses següents: (1pt)
   - Mostra aleatòria simple

No ens donen detalls de com s'han seleccionat els discs a observar d'entre la població. No es pot dir res més.

   - Normalitat de la variable resposta

Dubtes. La variable Y no és Normal, és una variable de Poisson amb taxa al voltant de 9. Per tant s'observen valors enters encara que pot ser la Normal podria ser una aproximació bastant raonable.

   - Mida de la mostra

La mida no és una premissa, però com que la Normalitat està en dubte segurament una mostra major seria preferible

   - Homocedasticitat

L'homocedasticitat (variància constant) no és una premissa de la prova d'una mitjana amb variància desconeguda.

NOM:_____COGNOMS:_____

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

Problema 3 (B5-B6)

En un estudi per comparar les notes de batxillerat (B) i les notes a l'examen de selectivitat (S) es fa una enquesta preguntant a 16 alumnes què han tret en ambdues notes. *Per contestar els següents apartats considereu un risc α=0.05 i una confiança del 95%*.

	Mean	Var
B	7.03	
S	6.19	
D	0.84	2.07

D = B-S

1) Quin tipus de mostres estaríem tractant, aparellades o independents? Justifiqueu la resposta. **[1p]**

Es tracta de mostres aparellades, ja que a cada alumne se'ls hi pregunta les dues notes i comparem la diferència de cada parell de notes. S'ha utilitzat aquest disseny perquè és més eficient que el de mostres independents (la variància de diferència és menor) donat que les diferències entre les unitats (alumnes) desapareixen.

2) Imagineu que tractem les dades aparellades.
   a) Plantegeu les hipòtesis nul·la i alternativa per estudiar si la nota de Batxillerat és superior a la de Selectivitat. Indiqueu si la prova és bilateral o unilateral. **[0.5p]**

$H_0: \mu_B = \mu_S <=> (\mu_B - \mu_S) = \mu_D = 0$
$H_1: \mu_B > \mu_S <=> (\mu_B - \mu_S) = \mu_D > 0$
Unilateral.

   b) Indiqueu la distribució de l'estadístic corresponent sota la hipòtesi nul·la i les premisses. **[1p]**

Estadístic $t \sim t(n-1)$
La variable diferència D = B – S ha de seguir una distribució Normal.

   c) Calculeu l'estadístic segons les dades i raoneu si podem rebutjar la hipòtesi nul·la. **[1p]**

$t = (0.84 - 0) / (1.438/\sqrt{16}) = 2.335$
$t_{n-1,0.95} = t_{15,0.95} = 1.753$.
Donat que l'estadístic és major que el punt crític, rebutgem la Ho. És a dir, tenim evidències suficients per dir que la nota mitjana de B és superior a la nota mitjana de S amb un risc α=0.05.

3) Imagineu que considerem les dades com a mostres independents. Volem posar a prova si les variàncies de les notes de Batxillerat i de Selectivitat són iguals.

   a) Plantegeu el contrast d'hipòtesi. **[0.5p]**

$H_0: \sigma_B^2 = \sigma_S^2$
$H_1: \sigma_B^2 \neq \sigma_S^2$

L'output de R que obtenim és el següent:

```
                F test to compare two variances

data:  data$S and data$B
F = ---- ,      num df = 15, denom df = 15, p-value = 0.5
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.486693    3.986782
```

b) A partir de l'output, calculeu l'estadístic F i interpreteu-ho. Suggeriment, dibuixeu el gràfic de la F **[1p]**

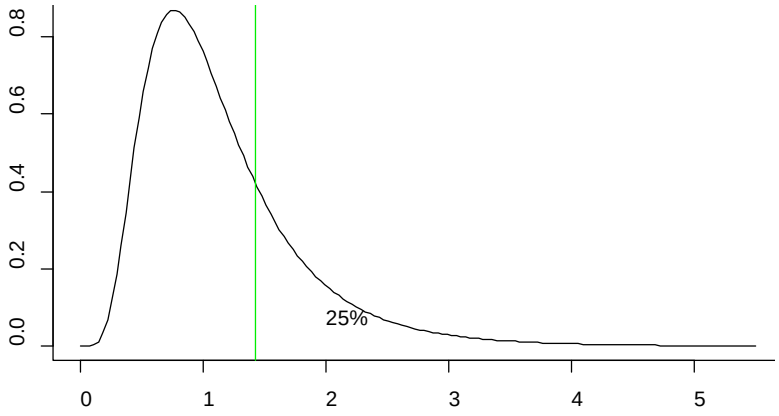
$2*P(F_{15,15} > F) = 0.5$

$P(F_{15,15} > F) = 0.25$

$P(F_{15,15} < F) = 0.75$

F = 1.43 (mirar en taules) [0.70 que és l'invers de 1.43 també seria correcte]

La variància en un grup és un 43% superior a la de l'altre grup.



c) Observant només l'interval de confiança, a quina conclusió arribaríem? Justifiqueu-ho. **[0.5p]**

Com que l'IC conté l'1, no podríem rebutjar la Ho d'igualtat de variàncies.

c) Quin seria el mínim rati de variàncies que donaria lloc a rebutjar la hipòtesi nul·la? **[0.5p]**

Es tracta de trobar el punt crític: $F_{15,15,0.975} = 2.86$.

4) Volem estudiar si existeix relació entre les dues notes. Per fer-ho, hem estimat un model lineal amb R que ens ha donat el següent output:

```
Call:
lm(formula = S ~ B, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8414 -0.9800 -0.1589  0.5532  2.3706

Coefficients:
                  Estimate     Std. Error    t value      Pr(>|t|)
(Intercept)       ---          ---           1.413        0.179
B                 ---          ---           1.606        0.131

Residual standard error: 1.335 on 14 degrees of freedom
Multiple R-squared: 0.1555,     Adjusted R-squared: 0.09519
F-statistic: 2.578 on 1 and 14 DF,  p-value: 0.1307
```

a) Calculeu els estimadors de β_0 i β_1, i escriviu l'equació de la recta sabent que $S_S / S_B = 1.2$. **[1p]**

r = sqrt(0.1555) = 0.39.

```
> b1 <- r * (1.2)
> b1
[1] 0.4732019
> b0 <- 6.19 -  b1*7.03
> b0
[1] 2.863391
```

S = b0 + b1*B.

S = 2.863391 + 0.4732019*B.

b) Podem considerar que la nota de Batxillerat està associada amb la nota de selectivitat(Justifica la resposta amb arguments d'inferència estadística)? Calculeu també l'interval de confiança del pendent. **[1.5p]**

Les notes no estan associades, ja que el p-valor és major de 0.05. És a dir no tenim evidències suficients per dir que el paràmetre β_1 sigui diferent de zero.

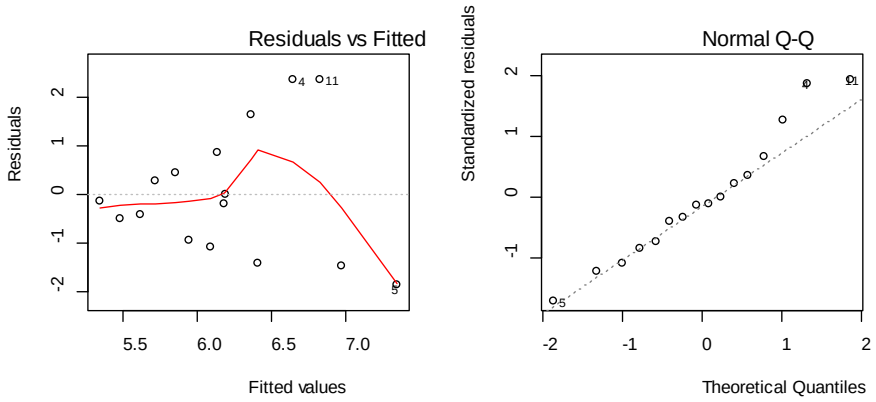
Sb1 = b1/(*t value B*) = 0.4732019/1.606 = 0.2946 (error de l'estimació del pendent b1)

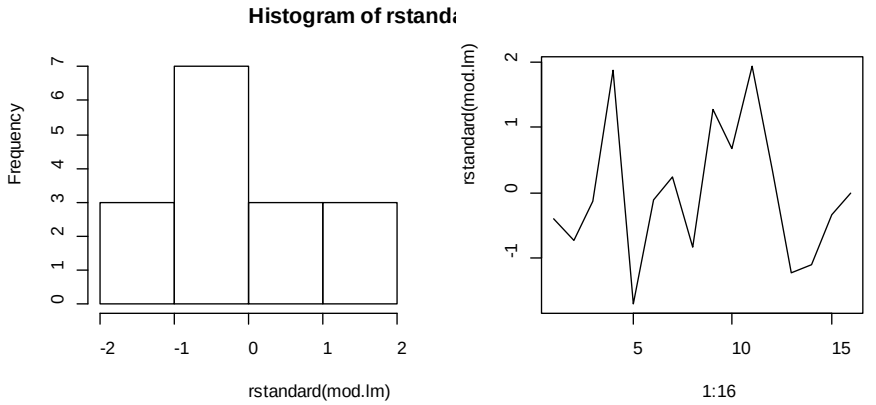
IC (β_1, 95%) = b1 ± $t_{n-2,0.975}$ * Sb1 = 0.4732019 ± 2.145*0.2946 = [-0.16, 1.11].

c) Podem dir que el model fa un bon ajustament de les dades? **[0.5p]**

No, ja que el R^2 = 0.1555 està molt lluny de 1. És a dir, només un 16% de la variabilitat de S s'explica per B.

d) Segons els plots següents, podem validar les premises? **[1p]**





- Linealitat entre S i B: en dubte; és difícil valorar-la amb tants pocs punts i amb tanta heteroscedasticitat. No hi ha evidència per pensar que no es compleix.
- Homoscedasticitat dels residus: NO.
- Normalitat dels residus: OK.
- Independència del residus: OK.