NOM:____

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

Problema 1 (B1-B2)

La taula adjunta mostra les proporcions (poblacionals) corresponents als estudiants que obtenen simultàniament una qualificació numèrica X a certa assignatura i una qualificació literal a una determinada competència transversal CT:

	3	5	7	9	10
A	0.006	0.012	0.067	0.04	0.005
B	0.009	0.021	0.096	0.033	0.018
C	0.072	0.144	0.169	0.088	0.044
D	0.055	0.075	0.036	0.01	0

1. (2 pt) Trobeu la funció de probabilitats de la variable aleatòria X, i calculeu el valor esperat i la desviació tipus. Interpreteu el resultat.

x_i	$P_X()$
3	0.142
5	0.252
7	0.368
9	0.171
10	0.067

E(X) = 3 0.142 + 5 0.252 + 7 0.368 + 9 0.171 + 10 0.067 = 6.471

$V(X) = \sum x_i^2 P_X(x_i) - E(X)^2 = 46.161 - 6.471^2 = 4.287$; $\sigma = 2.071$

En aquesta població d'estudiants, un indicador de tendència central de la qualificació és 6.471 punts, i una mesura de dispersió al voltant d'aquesta nota de referència és 2.071 punts.

2. (2 pt) Trobeu la distribució de la variable X condicionada a haver obtingut una D a la CT. Quina és la qualificació esperada entre aquests?

$P_{X\,|\,CT=D}(x) = P(X=x \cap CT=D)/P(CT=D)$

x_i	$P_{X\,|\,CT=D}()$
3	0.3125
5	0.4261
7	0.2045
9	0.0568

$E(X\,|\,CT=D) = \sum x_i \; P_{X\,|\,CT=D}(x_i) = 5.0114$

Es veu que quan la qualificació a la CT és baixa, la tendència és que la qualificació numèrica també sigui menor.

3. (1 pt) Quina és la proporció de B o C, en general? I entre els que han tret un 9? Expresseu formalment aquestes probabilitats abans de contestar.

$P(B \cup C) = P(B) + P(C) = 0.177 + 0.517 = 0.694$

$P(B \cup C\,|\,X=9) = P((B \cup C) \cap X=9) / P(X=9) = (0.033 + 0.088) / 0.171 = 0.7076$

4. (1 pt) Els estudiants poden optar a Matrícula d'Honor (MH) si obtenen un 10, o si obtenen un 9 i una A. Expliqueu la diferència existent entre trobar la probabilitat que un alumne amb MH hagi tret un 9, o que un alumne amb un 9 hagi tret MH. Calculeu-les.

1er: d'entre els que obtenen la matrícula, quants tenen 9 de qualificació? P(X=9 | MH) = P(X=9 ∩ CT=A) / P(MH)

2n: d'entre els que han tret un 9, quants obtenen la matrícula: P(MH | X=9) = P(X=9 ∩ CT=A) / P(X=9)

La diferència consisteix en establir quin és el conjunt sobre el que ens fixem en els casos favorables (que anirà al denominador).

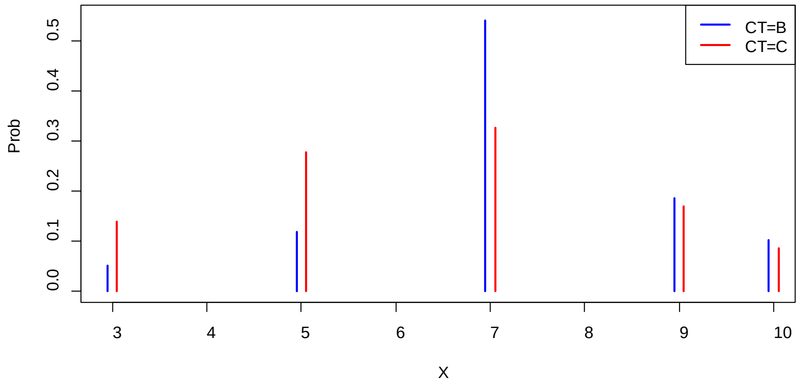
Les probabilitats són: P(X=9 | MH) = 0.04/(0.04+0.067) = 0.3738 i P(MH | X=9) = 0.04/0.171 = 0.2339

5. (1 pt) Si un estudiant escollit a l'atzar resulta que ha obtingut una qualificació alta a la CT (alguna entre A o B, però no sabem quina), quina és la probabilitat que la seva nota sigui superior a 5?

$P(X > 5\,|\,CT=A \cup CT=B) = 0.259 / 0.307 = 0.8436$

6. (2 pt) Són independents les qualificacions X i CT? Trobeu i representeu gràficament les funcions $P_{X\,|\,CT=B}(x_i)$ i $P_{X\,|\,CT=C}(x_i)$. Interpreteu el resultat, vinculant la resposta a la pregunta inicial.

x_i	$P_{X\,\|\,CT=B}()$	$P_{X\,\|\,CT=C}()$
3	0.0508	0.1393
5	0.1186	0.2785
7	0.5424	0.3269
9	0.1864	0.1702
10	0.1017	0.0851



No són independents, si la nota de la CT és millor (B), les probabilitats de que X sigui més alta són una mica majors que si la nota de CT és C.

7. (1 pt) Podem calcular la correlació entre X i CT?

No, no es pot perquè CT no és una variable numèrica, no podem calcular-li un valor esperat o una desviació tipus. Encara que sigui ordinal (A > B > C > D), no seria correcte fer una assignació de valors arbitrària per permetre el càlcul de la correlació. Existeix una associació entre les qualificacions X i CT, però en aquesta ocasió la correlació no és un indicador adequat.

8. (1 pt / •) Pensem ara en la nota X com una variable contínua amb la següent funció de distribució:

$$F_X(x) = \frac{ax+b}{x+1}, 0 \le x \le 10, 0 \text{ si } x < 0, 1 \text{ si } x > 10$$

- És cert que aquesta funció podria ser una funció de distribució? Què hauria de complir?
  Com és constant als extrems, només hi ha que imposar tres condicions:
  a) $F_X(0)=0$; b) $F_X(10)=1$; c) $F_X(x)$ és creixent entre 0 i 10

- Determineu els valors de les constants a i b per tal que sigui funció de distribució;
  $F_X(0)=0$, implica que $b=0$
  $F_X(10)=1$, implica que $10a/11 = 1$, llavors $a=11/10$
  [com la funció també pot escriure's: $F_X(x) = \frac{11}{10}\left(1-\frac{1}{x+1}\right)$, és fàcil comprovar que és estrictament creixent per $x > -1$]

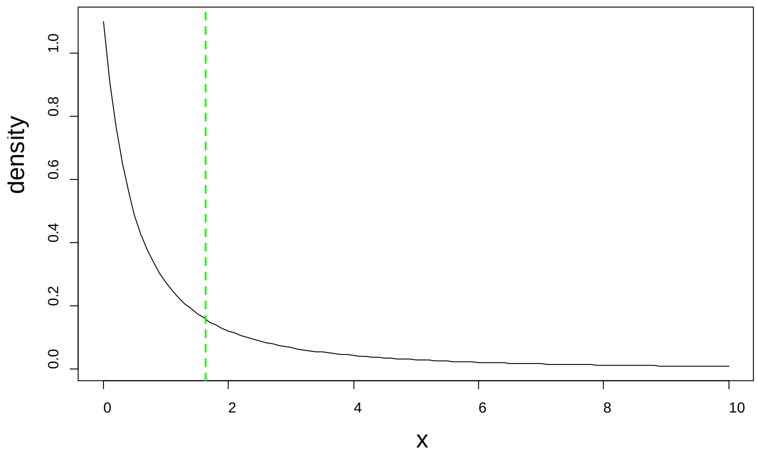
- Calculeu la probabilitat que la nota estigui entre 5 i 7, i la probabilitat que sigui major que 9;

  $P(5 < X < 7) = F_X(7) - F_X(5) = 0.9625 - 0.9167 = 0.0458$
  $P(X > 9) = 1 - F_X(9) = 1 - 11/10\ 9/10 = 1/100$

- Trobeu i dibuixeu la funció de densitat de X.

  $f_X(x) = \frac{d}{dx}F_X(x) = \frac{11}{10}\frac{1}{(x+1)^2}$

  Seria realment molt difícil arribar a una qualificació mitjana, perquè la majoria de valors es troben a prop de 0.



- Com es calcularia l'esperança de X? (no cal que ho resoleu). Marqueu a la gràfica el lloc aproximat on es situaria el valor esperat.

  $\int_0^{10} x\frac{11}{10}\frac{1}{(x+1)^2}\,dx$ (= 1.6377). Efectivament, la nota esperada seria molt baixa.

NOM:____

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

Problema 2 (B3-B4).

Suposem que fem una cerca amb Google. La probabilitat de *clicar* en un enllaç patrocinat que apareix a la primera posició de la primera pàgina al realitzar una cerca és 0.3. Una empresa de reparació d'equipaments informàtics té una plana web i s'està plantejant contractar aquest enllaç patrocinat per a que aparegui al fer una cerca amb els termes "reparació" i "ordinadors".

1. A l'àmbit d'actuació de l'empresa es realitzen 20 cerques <u>en una hora</u> amb la combinació d'aquests 2 termes.
   a. Quina és la distribució de la variable aleatòria "*Nombre de clicks fets en aquesta hora a l'enllaç patrocinat*"? Quina és la seva esperança i desviació estàndard?

$X = \text{"Nombre de clicks en una hora amb 20 cerques"} \sim Bin(n = 20, p = 0.3)$

$E(X) = n \cdot p = 20 \cdot 0.3 = 6$

$V(X) = n \cdot p \cdot (1 - p) = 20 \cdot 0.3 \cdot 0.7 = 4.2 \rightarrow \sigma = \sqrt{4.2} = 2.05$

   b. Quina és la probabilitat que el nombre de *clicks* en aquesta hora sigui igual o superior a 10?

$P(X \geq 10) = 1 - P(X \leq 9) = (Taules) = 1 - 0.9520 = 0.0480$

2. En un dia concret, durant les 8 hores laborables que treballen els tècnics de reparació s'han fet 15 cerques per hora d'aquests termes. La probabilitat que un visitant de la plana web contracti un servei de reparació és 0.15. Considereu la variable aleatòria "Nombre de serveis contractats en 8 hores"
   a. Quina és la distribució d'aquesta variable aleatòria? Quina és la seva esperança?

$Y = \text{"Nombre de serveis contractats en 8 hores"}$

$Y \sim Bin(n = 15 \cdot 8 = 120, p = 0.3 \cdot 0.15 = 0.045) \sim P(\lambda = 120 \cdot 0.045 = 5.4)$

$E(Y) = \lambda = n \cdot p = 5.4\ serveis\ contractats/dia$

   b. El nombre màxim de reparacions diàries que es poden fer per dia són 8. Quina és la probabilitat de no poder realitzar alguna reparació per sobrepassar aquest número de demandes en aquest dia?

$P(Y > 8) = 1 - P(Y \leq 8) = (Taules) = 1 - 0.903 = 0.097$

   c. Quin és el llindar de temps transcorregut entre dues demandes que es sobrepassarà amb una probabilitat de 0.8.

$T = \text{"}Temps\ entre\ dues\ demandes\text{"} \sim Exp(\lambda = 5.4)$

$P(T > t) = 0.8 \rightarrow P(T \leq t) = 0.2 \rightarrow 1 - e^{-5.4t} = 0.2 \rightarrow t = \frac{\ln(0.8)}{-5.4} = 0.298\ dies = 143.1\ minuts$

3. Al entrar en una plana web venint de *Google*, hi ha dos tipus d'usuari, els que surten rebotats de seguida perquè veuen que no els interessa la pàgina web (Usuaris A) i els que naveguen una estona buscant més informació (Usuaris B). El temps de la visita dels usuaris A a la pàgina web es distribueix amb una exponencial de λ=2/minuts, mentre que la distribució dels temps dels usuaris B es distribueix com una Normal amb μ=6 minuts i σ=2 minuts. També sabem que hi ha un 50% d'usuaris de cada tipus.
   a. Quin serà la durada esperada del global de les visites?

$DA = \text{"Durada en minuts de les visites dels usuaris A"} \sim Exp(\lambda = 2)$

$DB = \text{"Durada en minuts de les visites dels usuaris B"} \sim N(\mu = 6, \sigma = 2)$

$D = \text{"Durada en minuts de les visites"}$

$E(D) = E\left(\frac{DA+DB}{2}\right) = \frac{1}{2} \cdot [E(DA) + E(DB)] = \frac{1}{2} \cdot \left[\frac{1}{2} + 6\right] = 3.25\ minuts$

   b. Quina serà la durada que sobrepassaran els usuaris B amb una probabilitat de 0.99?

$P(DB > db) = 0.99 \rightarrow P\left(Z > \frac{db-6}{2}\right) = 0.99 \rightarrow (Taules) \rightarrow \frac{db-6}{2} = -2.33 \rightarrow db = 1.34\ minuts$

c. Dóna la probabilitat que un usuari que sobrepassa la durada de 1.5 minuts sigui un usuari de tipus B?

$D = \text{"}Una\ visita\ d'un\ usuari\ dura\ més\ de\ 1.5\ minuts\text{"}$

$A = \text{"}La\ visita\ la\ fa\ un\ usuari\ A\text{"}$

$B = \text{"}La\ visita\ la\ fa\ un\ usuari\ B\text{"}$

$P(D|A) = P(DA > 1.5) = 0.050$

$P(D|B) = P(DB > 1.5) = P\left(Z > \frac{1.5-6}{2}\right) = P(Z > -2.25) = 0.988$

$P(B|D) = \frac{P(D|B)\cdot P(B)}{P(D|A)\cdot P(A)+P(D|B)\cdot P(B)} = \frac{0.988\cdot 0.5}{0.050\cdot 0.5+0.99\cdot 0.5} = \frac{0.988}{0.050+0.988} = 0.952$

4. Finalment, l'empresa opta per no contractar un enllaç patrocinat de *Google* i decideix contractar-lo dins del cercador *Bing*. En un mes, la descriptiva de la durada en minuts de les visites provinents de Bing és la següent:

```
> length(durada)
[1] 120
> summary(durada)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0020  0.3655  2.8520  3.4370  6.4540  9.4800
> sd(durada)
[1] 3.146428
```

Fes la prova d'hipòtesi de si la durada mitjana de les visites provinents de *Bing* és de 2.5 minuts o no ho és amb un α=0.1.

a. Dóna l'expressió de l'estadístic, la seva distribució sota H_0 i les premisses adients

$\hat{Z} = \frac{\bar{y}-\mu_0}{\sqrt{\frac{S^2}{n}}} \sim N(0,1) \qquad Premisses: m.a.s\ \ i\ \ n \geq 100$

b. Dóna el valor de l'estadístic i del punt crític.

$Estadístic:\ \hat{Z} = \frac{3.44-2.5}{3.14\cdot\sqrt{\frac{1}{120}}} = 3.26 \qquad Punt\ crític: Z_{0.95} = 1.64$

c. Dóna el p-valor i la conclusió de la prova d'hipòtesi

$p - valor = P(|Z| > 3.26) = (Taules) = 2\cdot(1-0.9994) = 0.0011$

$Conclusió: com\ que\ |\hat{Z}| < Z_{0.95}\ i\ p - valor < 0.1\ rebutgem\ que\ la\ durada\ mitjana\ de$ $les\ visites\ provinents\ de\ Bing\ sigui\ igual\ a\ 2.5\ minuts.$

d. Dóna el IC90% per la durada mitjana de les visites provinents de *Bing* i fes la seva interpretació

$IC(\mu, 90\%) = \bar{x} \mp Z_{0.95}\cdot S\cdot\sqrt{\frac{1}{n}} = 3.44 \mp 1.645\cdot 3.14\cdot\sqrt{\frac{1}{120}} = [2.87\ , 4.01]\ \ minuts$

La mitjana de les visites provinents de *Bing* està entre 2.87 i 4.01 minuts amb una confiança del 90%

NOM:____

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

Problema 3 (B5-B6).

Es pren una mostra de 15 estudiants de l'assignatura de PE a la FIB i s'analitza si el nombre d'hores (**H**) que han estudiat per a un examen afecta a les puntuacions que han obtingut (**P1**). A més, tenim les puntuacions d'un altre grup (**P2**).

Les dades són: **Hores (H)**: 2, 3, 3, 4, 4, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8

Puntuació (P1): 5, 5, 6, 5, 7, 7, 8, 6, 9, 8, 7, 9, 10, 8, 9

Puntuació (P2): 4.4, 4.5, 5.3, 3.7, 7.3, 6.7, 6.8, 5.8, 8.2, 8, 7, 7.8, 8, 8.7, 8

i els estadístics obtinguts a partir d'aquestes tres variables són:

Mitjana de H: mean(H)=5.4 **Desviació tipus de H**: s_H=1.88

Mitjana de P1: mean(P1)=7. 27 **Desviació tipus de P1**: s_{P1}=1.62

Mitjana de P2: mean(P2)=6.68 **Desviació tipus de P2**: s_{P2}=1.58

Correlació entre H i P1: cor(H,P1)=0.78

1.- (2 punts) A partir de les desviacions tipus d'aquestes dades mostrals per a P1 i P2, creieu que les desviacions tipus poblacionals d'aquests grups són similars? Plantegeu i resoleu la prova d'hipòtesi que ens permeti rebutjar o no aquesta afirmació.

$\boldsymbol{H_0}: \sigma^2_{P1} = \sigma^2_{P2}$

$\boldsymbol{H_1}: \sigma^2_{P1} \neq \sigma^2_{P2}$

$\hat{F}_{n_{1-1},n_{2-1}} = \frac{S^2_{p1}}{S^2_{p2}} = \frac{1.62^2}{1.58^2} = \frac{2.62}{2.50} = 1.05$ amb $n_1 - 1 = 14$ i $n_2 - 1 = 14$ graus de llibertat

Segons el valor de l'estadístic (1.05) i el punt crític $F_{14,14,0.975}$=2.98 (a les taules $F_{15,15,0.975}$=2.86) no es pot rebutjar H_0, i per tant es pot assumir que les dues variàncies són similars.

2.- A partir de les puntuacions P1 i P2 d'aquestes dades mostrals es creu que les mitjanes d'aquestes mostres podrien ser similars. Plantegeu i resoleu la prova d'hipòtesi que permeti rebutjar o no aquesta hipòtesi**.** Les passes a seguir són:

a) (0.5 punt) Definiu la hipòtesi nul.la i l'alternativa per al test relacionat amb les mitjanes de les puntuacions P_1 i P_2

$H_0: \mu_{P1} = \mu_{P2}$

$H_1: \mu_{P1} \neq \mu_{P2}$

b) (1 punt) Calculeu el valor de l'estadístic d'aquesta prova d'hipòtesi al nivell de significació α=0.05 i resoleu-la

Assumint distribució normal i que les variàncies mostrals son equivalents,

$$s^2 = \frac{(15-1)S_1^2+(15-1)S_2^2}{15+15-2} = \frac{36.74+34.95}{28} = \frac{71.69}{28} = 2.56$$

$$\hat{t} = \frac{\overline{y_{p1}}-\overline{y_{p2}}}{S\sqrt{1/n1+1/n2}} = \frac{7.27-6.68}{1.6\sqrt{\frac{1}{15}+\frac{1}{15}}} = \frac{0.59}{1.6*0.37} = \frac{0.59}{0.59} = 1$$

Valor de la t (a taules): $t_{\upsilon=15+15-2,\alpha=0.975}$= 2.048 Com que $\hat{t}$ =1 < 2.048, no es pot rebutjar la H_0 d'igualtat de mitjanes

c) (0.5 punt) Quin és el valor, en termes de valors absoluts, més petit de l'estadístic t pel qual la hipòtesi nul·la pot ser rebutjada?

$t_{\upsilon=28,\alpha=0.975}$= 2.048.

d) (0.75 punt) Indiqueu si les següents afirmacions són verdaderes o falses i justifiqueu la vostra resposta:

i. Si es rebutja la hipòtesi nul.la H_0al nivell de 0.05, també podem rebutjar al nivell 0.1

Verdadera, perquè si hem pogut rebutjar al 5% és perquè el p-valor era inferior a 5% (i també serà inferior a 10%)

ii. El risc d'error de tipus I és la probabilitat de rebutjar la hipòtesi nul·la quan H_1 és certa

Falsa, és la probabilitat de rebutjar la H_0 quan H_0 és certa

iii. Si el p-valor és igual a 0.15, podem rebutjar la hipòtesi nul·la al nivell del 10%

Falsa, ja que el 10% és menor que 0.15 i per tant no seriem a la zona de rebuig de la H_0

3.- (0.5 punt) Expliqueu com s'haurien de recollir les dades per a que fossin un conjunt de dades aparellades i com s'haurien de recollir si es vol treballar amb dos conjunts de dades independents

Quan tenim dues mostres de forma que els valors de cada mostra pertanyen al mateix individu, parlem de dades aparellades, és a dir, mesurem, per exemple, el que passa abans i després d'una experiència, com podria ser les notes obtingudes per un conjunt d'estudiants abans i després d'estudiar 8 hores diàries. Com exemple de dades independents en aquest cas seria comparar les notes dels estudiants d'un grup A amb les notes dels estudiants d'un grup B.

4.-a) (2 punt) Plantegeu un model lineal i estimeu la recta de regressió que permet estimar la puntuació de **P1** a partir de les hores d'estudi **H**

$P1=\beta_0 + \beta_1 H+\varepsilon$

on β_0 i $\beta_1$son els paràmetres desconeguts i ε és el soroll, desconegut

La recta de regressió estimada és

$\widehat{P1}=b_0 + b_1H$ i els paràmetres estimats són

$b_1=r\frac{s_{P1}}{s_H} = 0.78\frac{1.62}{1.88} = 0.67$

$b_0 = \overline{P1} - b_1\overline{H} = 7.27 -0.67*5.4=7.27-3.62 = 3.65$

Per tant, la recta de regressió estimada és:

$\widehat{P1}=3.65 + 0.67H$

b) (0.5 punt) Un alumne determinat ha estudiat 6.5 hores i ha tret un 7.5: calculeu el seu valor residual

$\widehat{P1}=3.65 + 0.67*6.5 = 8.007$

Valor residual= 7.5−8.007 = −0.507

c) (0.5 punt) Calculeu un estimador no esbiaixat per a la variància dels errors d'aquest model

$$s^2=\frac{\sum e_i^2}{n-2} = \frac{(n-1)s_{P1}^2(1-r^2)}{n-2} = \frac{14*1.62^2*(1-0.78^2)}{13} = \frac{14*2.62*0.39}{13} = \frac{14.31}{13} = 1.1 \qquad (s=1.05)$$

d) (0.75 punt) Calculeu i interpreteu el coeficient de determinació (R^2)

En el cas d'una regressió lineal simple, el coeficient de determinació R^2 coincideix amb el coeficient de correlació al quadrat. En aquest cas serà $0.78^2 = 0.61$.

És una mesura de la qualitat de l'ajust en el model lineal. En aquest sentit, quan més a prop d'1 sigui el R^2, millor qualitat de l'ajust, en el sentit que els errors (diferència entre valor observat i el valor obtingut a la recta de regressió) siguin mínims.

e) (1 punt) Obtingueu un interval de confiança, a un nivell de confiança del 95%, pel valor mitjà de la puntuació, sabent que les hores d'estudi han estat 5.5

$\widehat{P1}_h=3.65 + 0.67*5.5 = 7.34$

Interval de confiança al 95%

$$\widehat{P1}_h \pm 2.16*1.05\sqrt{\frac{1}{15}+\frac{0.01}{49.48}} = 7.34 \pm 0.59 = [6.75, 7.93]$$