NOM:_____

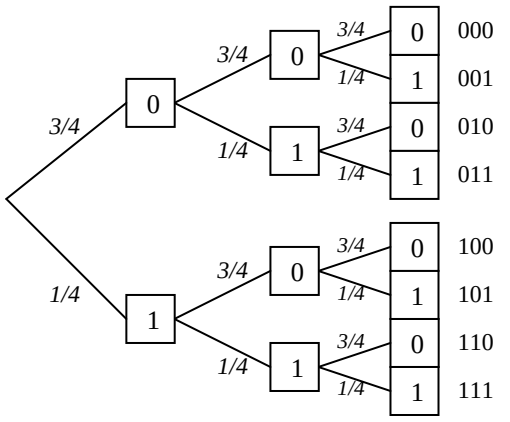
(Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs.)

PROBLEMA 1. Considerem el conjunt de tots els paquets de 3 bits que es poden enviar per una linea de comunicació (000,001,010,011,100,101,110,111) com a resultats d'una experiència aleatòria, i suposem que la probabilitat del 0 és el triple que la del 1.

Definim, també, dues variables aleatòries X i Y. La variable X és la suma dels 3 bits i la variable Y és el número d'alternances en la seqüència de bits de cadascun dels resultats. Per tant, X ∈ {0,1,2,3} i Y ∈ {0,1,2}.

1.- (1 punt) Indiqueu el conjunt de resultats de l'experiència aleatòria i les seves probabilitats associades justificant-les (en representació en arbre si voleu) i associeu a cada resultat un valor per les variables X i Y

$\Omega = \{000,001,010,011,100,101,110,111\}$



Ω	Probabilitats
000	P(000)=3/4*3/4*3/4 = 27/64
001	P(001)=3/4*3/4*1/4 = 9/64
010	P(010)=3/4*1/4*3/4 = 9/64
011	P(011)=3/4*1/4*1/4 = 3/64
100	P(100)=1/4*3/4*3/4 = 9/64
101	P(101)=1/4*3/4*1/4 = 3/64
110	P(110)=1/4*1/4*3/4 = 3/64
111	P(111)=1/4*1/4*1/4 = 1/64

X (suma)	*Y (#alternances)*
0	*0*
1	*1*
1	*2*
2	*1*
1	*1*
2	*2*
2	*1*
3	*0*

2.- (1 punt) Calculeu la probabilitat que la suma dels 3 bits sigui 2. I la probabilitat que la suma dels tres bits sigui 2 si sabem que el primer bit ha estat 1

Resultats amb suma dels 3 bits igual a 2 és X=2, és a dir {011,101,110}
P(X=2) = P(011)+P(101)+P(110) = 3/64 + 3/64 + 3/64 = **9/64 = 0.14**

Resultats que comencen per 1: A={100,101,110,111} P(A)=P(100)+P(101)+P(110)+P(111)=1/4
P(X=2 | A) = P(X=2 i A)/P(A) = (P(101)+P(110)) / P(A) = (6/64)/(1/4) = 6/16 = **3/8 = 0.375**

3.- (2 punts) Indiqueu la taula de probabilitats conjuntes de les variables X i Y, les probabilitats marginals corresponents i la seves esperances

P_{YX}	X=0	X=1	X=2	X=3	
Y=0	27/64	0	0	1/64	28/64
Y=1	0	18/64	6/64	0	24/64
Y=2	0	9/64	3/64	0	12/64
	27/64	27/64	9/64	1/64	

E(X) = 27/64 • 0 + 27/64 • 1 + 9/64 • 2 + 1/64 • 3 = 48/64 = **3/4 = 0.75**
E(Y) = 28/64 • 0 + 24/64 • 1 + 12/64 • 2 = 48/64 = **3/4 = 0.75**

4.- (1 punt) Hi ha independència entre les variables X i Y? Justifiqueu-ho i expliqueu què implica

No, perquè per exemple P(X=0,Y=0) = 27/64 no és igual a P(X=0)•P(Y=0) = 27/64 • 28/64

Per tant: X i Y depenen, no es distribueixen independentment, les probabilitats dels valors d'una de les variables varien depenent dels valors de l'altra.
Per ex: la probabilitat de 0 alternances és 0 si la suma és 1 o 2, però no és 0 si la suma és 0 o 3

S'ha estudiat que en l'enviament de paquets de 3 bits, es pot considerar una nova variable Q que indica en quants d'aquests bits s'ha presentat algun problema (en 0,1,2 o 3 casos), i les probabilitats de la qual són 0.4 0.3 0.2 i 0.1 respectivament

5.- (1 punt) Calculeu l'esperança i la variància de la variable Q

E(Q) = $0.4 \cdot 0 + 0.3 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 3$ **= 1.0**

V(Q) $= E(Q^2) - E(Q)^2 = 2 - (1)^2$ **= 1.0**

($E(Q^2) = 0.4 \cdot 0 + 0.3 \cdot 1 + 0.2 \cdot 4 + 0.1 \cdot 9 = 2$)

o bé $=(0-1)^2 \cdot 0.4 + (1-1)^2 \cdot 0.3 + (2-1)^2 \cdot 0.2 + (3-1)^2 \cdot 0.1 =$

$= 1 \cdot 0.4 + 0 \cdot 0.3 + 1 \cdot 0.2 + 4 \cdot 0.1 = 0.4+0+0.2+0.4$ **= 1.0**

6.- (1 punt) Calculeu la probabilitat de tenir 2 o menys problemes. I la probabilitat de tenir 2 o menys problemes si sabem que almenys 1 n'ha tingut

$P(Q<=2) = 0.4+0.3+0.2$ **= 0.9**

$P(Q<=2 \mid Q>1) = P(Q<=2 \text{ i } Q>1) / P(Q>1) = P(Q=2) / P(Q>1) = 0.2 / (0.2+0.1)$ **= 0.666**

Ara, enlloc de considerar els anteriors paquets de tres bits, considerem enviaments de 8 bits amb probabilitat d'enviar 0 sent el triple que la d'enviar 1.

7.- (1 punt) Definiu la variable nombre de 0's en els 8 enviaments i indiqueu l'esperança i variància d'aquesta variable

X "nombre de 0's en 8 enviaments" és Binomial(n=8, p=0.75)
E(X) = 8 * 0.75 = 6
V(X) = 8 * 0.75 * 0.25 = 1.5

Finalment, estudiem el comportament més centrat en l'enviament per unitat de temps que en el nombre d'enviaments, i suposem que el nombre mitjà de bytes enviats per *unitat de temps* és de 5.

8.- (1 punt) definiu la variable nombre de bytes enviats per *unitat de temps*, indiqueu l'esperança i variància d'aquesta variable i calculeu la probabilitat de que se n'enviïn exactament 5

Y és Poisson(λ=5)
E(Y) = 5
V(Y) = 5

$P(Y=5) = P(Y<=5) - P(Y<=4) = 0.616 - 0.440 =$ **0.176**

9.- (1 punt) definiu la variable temps entre enviaments, indiqueu-ne l'esperança i la variància, i calculeu la probabilitat de que el temps entre enviaments sigui superior a 1 *unitat de temps*

T és Exp(λ=5)
E(T) = 1/5 = 0.2
V(T) = 1/25 = 0.04

$P(T>1) = 1 - P(T<=1) = 1 - (1 - \exp(-5*1)) =$ **0.0067**

NOM:_____

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

PROBLEMA 2. Memòria dels ordinadors

Un servidor de la marca ***Fuig Primer R100*** te 12 xips de memòria de 8Gb cadascun.

a) **(1PUNT).** Si la probabilitat de funcionament de cada xip és 0.92, quina és la probabilitat que funcionin 9 xips?

N=12 xips

P(X=9)=$\binom{12}{9} 0.92^9(1-0.92)^3 = \frac{12!}{9!3!}0.92^9 0.08^3 = 0.05$

b) **(1PUNT).** A continuació ens diuen que han millorat la qualitat d'aquests xips i ara la probabilitat de que funcioni cada xip és del 95%. Trobeu la probabilitat de que funcionin 10 o més xips.

P(X≥10)= $\binom{12}{10} 0.95^{10}(1-0.95)^2 + \binom{12}{11} 0.95^{11}(1-0.95) + \binom{12}{12} 0.95^{12}$=0.98

Si es vol calcular utilitzant les taules:

P(X≥10|p=0.95)= P(X≤3|p=0.05)=0.98

c) **(1PUNT)**. Tenim una granja de servidors que te 20 ordinadors, cadascun d'aquests servidors amb les mateixes característiques que el servidor ***Fuig Primer R100,*** i la probabilitat que falli un xip és del 0.05. Dieu quin serà el valor mitjà i la desviació tipus del nombre de xips que fallin a la granja de servidors.

N=20 servidors * 12 xips = 240 xips

X=#xips

E(X)=n*p=240*0.05=12 xips

V(X)=n*p*q=240*0.05*0.95=11.40

σ(X)=sqrt(11.40)=3.38

d) **(1PUNT)**. S'ha recollit una mostra de n=100 xips, observant que la proporció de xips defectuosos és del 6%. Doneu un interval de confiança al 90% per a la proporció de xips defectuosos.

IC(p_def,90%)=$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.06 \pm 1.65\sqrt{\frac{0.06*0.94}{100}} = (0.02, 0.1)$

e) **(2PUNTS)**. Ara es suposa que en una mostra de 100 xips n'hi ha 88 que funcionen.
Poseu a prova si es pot admetre que la proporció de defectuosos és del 6% o no, amb un risc del 5%.

H_0: $\pi=0.94$
H_1: $\pi\neq0.94$

$$\hat{z} = \frac{0.88-0.94}{\sqrt{\frac{0.94*(1-0.94)}{100}}} = -2.52$$

Prob(z<(-2.52))= 0.006
Per ser una prova d'hipòtesi bilateral, el p_valor valdrà
2*Prob(z<(-2.52))=0.01, i aquesta probabilitat és inferior al risc del 5%, per tant es rebutja la H_0, és a dir, que la proporció de defectuosos no pot ser del 6%.

f) **(2PUNTS)**. Hem detectat que, amb un 95% de confiança, la proporció de xips defectuosos podria estar entre 0.3 i 0.5. Calculeu quina és la grandària de la mostra que permet assumir aquests valors.

IC(π,0.95)= $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$
La meitat de l'amplada d'aquest IC és 0.5*(0.5-0.3)=0.1 i p=0.4
Llavors $1.96\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.4(1-0.4)}{n}} = 0.1$ i
n=$\frac{0.4*(1-04)*1.96^2}{0.1^2} = \frac{0.92}{0.01} = 92$
i per tant, es necessitaran mínim 92 observacions.

g) **(2PUNTS)**. Tenim una mostra de 31 xips que son similars en quant a tecnologia, no obstant els preus són diferents degut a que s'han comprat a diferents proveïdors. La desviació tipus mostral d'aquests preus és 20€. Doneu un interval de confiança al 95% per a la desviació tipus poblacional d'aquests preus.

$$IC(\sigma^2, 95\%) = \left(\frac{s^2(n-1)}{\chi^2_{n-1,1-\alpha/2)}}, \frac{s^2(n-1)}{\chi^2_{n-1,\alpha/2)}}\right) = \left(\frac{20^2*30}{46.979}, \frac{20^2*30}{16.791}\right) = (255.43, 714.67)$$

Per tant l' interval, calculat al 95% per a la desviació tipus poblacional és ($15.98 < \sigma < 26.73$)

Els responsables de la xarxa social "tonti" volen provar dos tipus de perfil amb els seus usuaris. A data 1 de maig, a uns els assignen el perfil *normal*, i a uns altres el perfil *cool*. Escullen aleatòriament uns usuaris de cada tipus i analitzen les següents dades a dos moments diferents:

	perfil *normal* (n=250) $\bar{x}$ (s)	perfil *cool* (n=250) $\bar{x}$ (s)	
#F-mg	85.4 (17.3)	87.8 (19.0)	Nombre de *followers* a 1 de maig
#F-jn	91.2 (18.7)	97.5 (21.2)	Nombre de *followers* a 1 de juny
#com	127 (33.5)	146 (41.4)	Nombre de comentaris enviats al mes de maig

A. Quina d'aquestes afirmacions té més sentit? No facis cap PH, només comenta.
   a. L'assignació del perfil als usuaris a l'inici no s'ha fet bé: els *cools* tenen més *followers*
   b. La diferència de *followers* a l'inici es pot explicar només per l'atzar

Hi ha una diferència de 2.4 entre els grups, i les desviacions tipus són al voltant de 20. Encara que els grups són grans (250 individus), no sembla difícil que pugui ser una diferència casual. Si es vol fer un càlcul més precís, es pot estimar l'error típic de la diferència de dues mitjanes, en aquest cas 1.62, que no és molt més petit.

B. Escull i comenta dos de les afirmacions. Subratlla els aspectes que no són certs, fent la correcció corresponent:
   a. Hauríem de saber la variació abans-després de *followers* per a cada usuari, i analitzar el canvi entre els dos grups, *normal* i *cool*, que són dues mostres independents

Totalment correcte. És el que es fa a l'apartat D.

   b. Hauríem de saber les diferències de *followers* entre usuari *normal* i *cool*, que són mostres aparellades, i veure si aquesta mostra prové d'una mitjana 0.

No, les mostres no són aparellades entre els dos grups, és impossible: un usuari o té el perfil *normal* o el té *cool*. Les mostres són aparellades en el sentit abans-després.

   c. Si les mitjanes de #F-mg fossin les mateixes, compararíem les mitjanes de #F-jn que són dues mostres independents, però de variàncies poblacionals diferents. Però potser els *cools* tenen més *followers* perquè al principi ja tenien més.

És veritat que per comparar només les mitjanes al final hem d'estar segurs que a l'inici els grups eren comparables, però no és tan cert assegurar que no ho eren. Ja hem dit a l'apartat A que la diferència no és important. Tampoc és tan clar dir que les variàncies són diferents, s'hauria de verificar. Per suposat, grups desequilibrats a l'inici implica que l'anàlisi estaria esbiaixat.

   d. Podem comparar els dos perfils utilitzant simplement la informació de la fila de juny, i assumint que poblacionalment els dos grups al maig eren idèntics.

Exactament, si estem segurs que els dos grups s'han format aleatòriament (com sembla que s'ha fet), tenim una garantia de que al maig són la mateixa població, i podem concentrar-nos exclusivament en les dades de juny.

C. Ens fan notar que el nombre d'usuaris que, al juny, tenen més de 200 *followers* és més o menys un 5%. Què representa aquesta informació? Pot comprometre l'anàlisi estadístic que es vol fer?

Si un 5% dels usuaris tenen més de 200 followers, amb mitjanes inferiors a 100 i desviacions al voltant de 20, això representa una proporció molt més gran que el que correspondria a una normal, és a dir, sembla que aquesta variable està bastant desequilibrada per la dreta. No obstant, si anem a comparar mitjanes, la grandària de mostra (250 per grup) és més que suficient per assegurar la premissa principal, que és la normalitat de la diferència de mitjanes mostrals.

D. Per a la diferència de *followers* entre juny i maig hem observat desviacions estàndards de 7.8 i 10.6 als grups *normal* i *cool* respectivament. Plantegeu i resoleu la PH corresponent per contestar la qüestió de si el perfil *cool* fa aconseguir més *followers* que el *normal,* treballant amb aquest variable *diferència*.

Treballarem amb l'increment de *followers* I = #F-jn − #F-mg, i podem conèixer les mitjanes per a cada grup directament: $m(I_N) = 91.2-85.4=5.8$; $m(I_C) = 97.5-87.8=9.7$. Les desviacions no las podíem deduir però ens les han donat a l'enunciat.

H_0: $E(I_N) = E(I_C)$

Són dues mostres independents, i assumirem que les desviacions poblacionals són la mateixa. Sota H0, l'estadístic següent segueix una distribució aproximadament N(0,1) perquè $n_N + n_C = 500$:

$$t = \frac{\overline{y_C} - \overline{y_N}}{S\sqrt{1/n + 1/n}} \approx N(0,1)$$

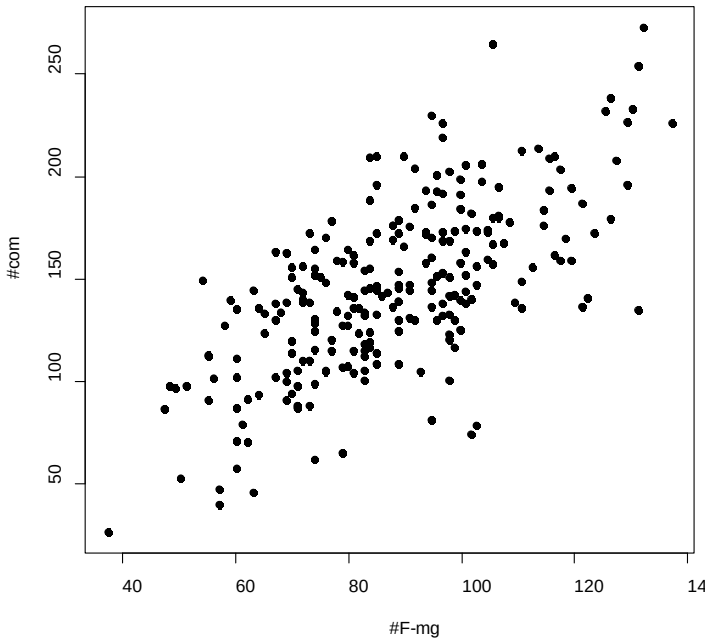
Al risc del 5% bilateral, la zona de rebuig és ±1.96 aproximadament. En el nostre cas, primer trobarem l'estimació pooled de la variància: $(249\ 7.8^2 + 249\ 10.6^2)/498 = 86.6 = 9.306^2$. L'estadístic val: (9.7 – 5.8) / (9.306 0.08944) = 4.68. Per tant, podem rebutjar la hipòtesi nul·la, els increments en mitjana no han estat iguals als dos grups (el P-valor amb les taules no es pot trobar, però com tenim 4 xifres de precisió segur que és menor que 0.0001).

E. Obteniu una mesura de l'efecte del perfil, mitjançant un interval de confiança al 95%, i feu una interpretació global dels resultats.

$$\mathrm{IC}(\mu_C - \mu_N, 95\%) = \overline{y_C} - \overline{y_N} \pm z_{0.975} S\sqrt{1/n + 1/n} = 9.7 - 5.8 \pm 1.96\ 9.306\ 0.08944 = (2.27, 5.53)$$

Hi ha forta evidència de que el perfil *cool* aporta més *followers* que el perfil *normal*. Concretament, amb una confiança del 95%, en mitjana s'aconsegueixen entre 2.27 i 5.53 *followers* més.

F. La figura adjunta mostra la relació entre el nombre de comentaris i *followers* a l'inici en el grup *cool*. Tenim la dada de que la correlació mostral a aquestes variables és 0.67214.



a. Estimeu quina és la recta de regressió.

$$b_1 = r\,{}^{s_Y}/_{s_X} = 0.672\,{}^{41.4}/_{19} = 1.4646$$

$$b_0 = 146 - 1.4646\ 87.8 = 17.41$$

b. Interpreteu els coeficients que heu trobat.

b_0 representa una estimació del nombre de comentaris per a un usuari amb 0 *followers*; el pendent b_1 representa que s'espera que un *follower* addicional per a un usuari suposi 1.46 comentaris més.

c. Un model lineal sembla un model apropiat? Repasseu breument les premisses del model.

Amb el gràfic podem dir que la tendència lineal és bastant clara, però per poder criticar les premisses d'homoscedasticitat i normalitat ens caldria un gràfic basat en els residus. A partir del plot potser s'observa un lleu increment de la variància de la resposta per a valors alts. Per la manera d'obtenir les dades (usuaris escollits a l'atzar), sembla que no hauríem de dubtar de la independència de les respostes.

G. Què val la desviació residual estimada per al model? I el coeficient de determinació? Estan relacionats aquests dos indicadors? És a dir, si un puja l'altre ha de pujar o baixar? Si l'objectiu del model és fer previsions del nombre de comentaris a un mes, segons el nombre de *followers* d'un usuari, com s'interpreten els dos indicadors?

$$S^2 = \frac{n-1}{n-2} s_Y^2 \,(1 - r^2) = 943.43 = \ 30.72^2$$

$$R^2 = 0.672^2 = 0.4518 \quad (45.18\%)$$

La desviació residual indica la incertesa del nombre de comentaris per a un usuari amb cert nombre de *followers* determinat (31 comentaris amunt o avall de la previsió puntual seria una fluctuació habitual). El coeficient de determinació indica que un 45% de la variabilitat dels comentaris s'explica per el nombre de *followers* de l'usuari, i un 55% per altres causes. Si aquest coeficient augmenta, disminueix la part incerta i, per tant, ha de disminuir la desviació residual.

H. Estimeu el pendent amb un interval de confiança del 95%.

$$\mathrm{IC}(\beta_1, 95\%) = b_1 \pm z_{0.975} s_{b1} = b_1 \pm z_{0.975} \frac{S^2}{(n-1)s_X^2} = 1.4646 \pm 1.96 \frac{943.43}{249\ 19^2} = (1.264, 1.665)$$