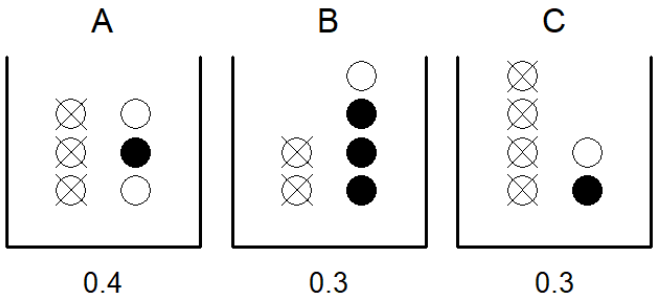
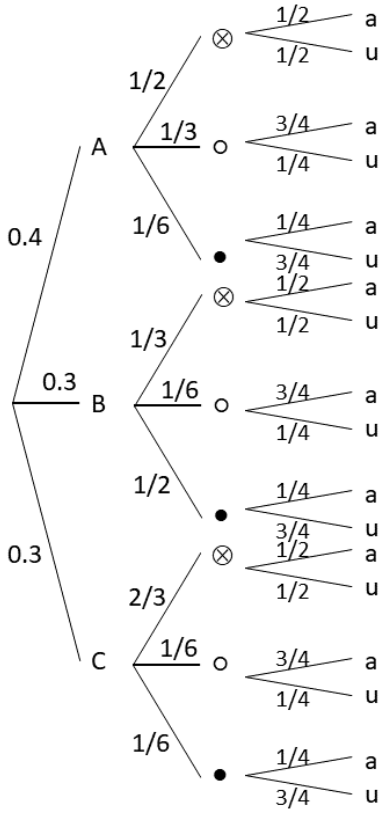
Problema 1 (Bloc A). Solució

(*heu de justificar formalment el desenvolupament de les respostes*)

Tenim tres urnes A, B I C que contenen cadascuna sis monedes, com es veu a l'esquema de la dreta. Les monedes marcades amb ⊗ són monedes normals, amb la mateixa probabilitat per cara que per creu. El cercle blanc representa monedes trucades: la cara és 3 vegades *més* probable que la creu. El cercle negre també, però la cara és 3 vegades *menys* probable que la creu. Una urna es escollida a l'atzar, amb probabilitat 40% per a A, i les altres amb probabilitat 30% cadascuna. De l'urna escollida traiem una moneda (totes són indistingibles al tacte), que es llença a l'aire per observar un dels seus costats aleatòriament.



1. Representeu a la dreta l'arbre de esdeveniments i probabilitats de l'experiència descrita anteriorment. (1.5pt)



"a" representa "cara"
"u" representa "creu"

2. Poseu un exemple de *resultat* i un altre d'*esdeveniment*, clarament diferenciats, extrets de l'experiència anterior. (1pt)

Resultat: ω = {urna A, moneda ⊗ #1, cara}, si cada moneda és individualment distingible de les altres de la mateixa classe, però tampoc és necessari.

Quan demanem un resultat, esperem que podeu donar un element ω ∈ Ω.

Esdeveniment: D = "obtenir creu amb la moneda"; un altre exemple: E = "elegir una moneda trucada". Aquests són casos que inclouen múltiples resultats.

3. Si l'urna escollida és la B, calculeu la probabilitat que de la moneda extreta observem cara. (1pt)

P(cara | B) =
P(cara | B i ⊗) P(⊗ | B) + P(cara | B i ○) P(○ | B) + P(cara | B i ●) P(● | B) =
1/2 · 1/3 + 3/4 · 1/6 + 1/4 · 1/2 = 5/12

És coherent, ja que en aquesta urna les monedes més abundants són les trucades que retornen creu més sovint que cara.

4. Quina és la probabilitat d'extreure una moneda trucada? (1pt)

P(¬ ⊗) = P(¬ ⊗ | A) P(A) + P(¬ ⊗ | B) P(B) + P(¬ ⊗ | C) P(C) = (1 - 1/2) 0.4 + (1 – 1/3) 0.3 + (1 – 2/3) 0.3 = 1/2

Fixeu-vos que no és correcte deduir que, com hi ha 9 monedes trucades d'un total de 18, el resultat és el 50%, encara que aquesta vegada casualment hem coincidit.

5. El costat de la moneda que hem extret d'una urna mostra una *cara*. Trobeu la probabilitat que l'urna hagi estat la A. (1.5pt)

P(A | cara) = P(A i cara) /P(cara).

P(A i cara) = P(A i ⊗ i cara) + P(A i ○ i cara) + P(A i ● i cara) = (0.4 · 1/2 · 1/2) + (0.4 · 1/3 · 3/4) + (0.4 · 1/6 · 1/4) = 13/60

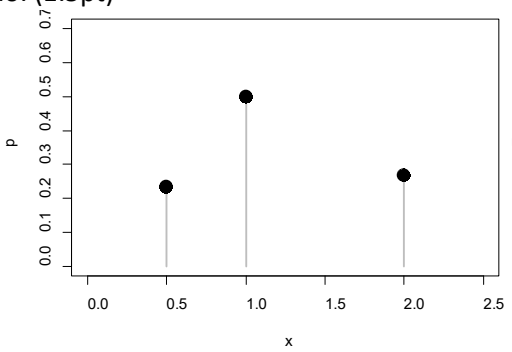
P(cara) = P(A i cara) + P(B i cara) + P(C i cara). De manera similar al cas anterior:
P(B i cara) = (0.3 · 1/3 · 1/2) + (0.3 · 1/6 · 3/4) + (0.3 · 1/2 · 1/4) = 1/8
P(C i cara) = (0.3 · 2/3 · 1/2) + (0.3 · 1/6 · 3/4) + (0.3 · 1/6 · 1/4) = 3/20

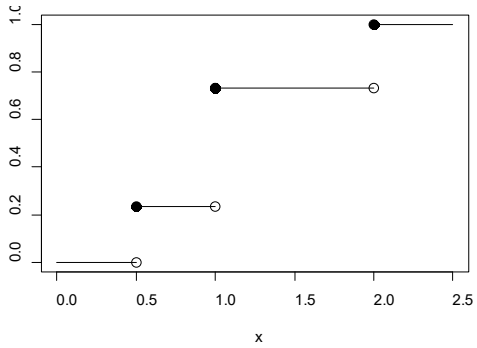
P(A | cara) = (13/60) / { 13/60 + 1/8 + 3/20 } = (26/120) / (59/120) = 26/59 = 0.4407...

Suposeu que les monedes no trucades (⊗) són de 1€; les monedes (○) són de 50 cèntims, i les (●) són de 2€. Com abans, la ma que treu la moneda de l'urna seleccionada no és capaç de diferenciar unes monedes d'altres.

6. Especifiqueu la distribució de probabilitat de la variable aleatòria X "valor de la moneda"; representeu gràficament les funcions de probabilitat i de distribució. (1.5pt)

	x	p(x)	F(x)
○	0.50	7/30	7/30
⊗	1	1/2	11/15
●	2	4/15	1





7. Trobeu el valor esperat de X i la seva desviació tipus. (1pt)

$E(X) = 0.50 \cdot 7/30 + 1 \cdot 1/2 + 2 \cdot 4/15 = \underline{1.15}$€

$E(X^2) = 0.50^2 \cdot 7/30 + 1^2 \cdot 1/2 + 2^2 \cdot 4/15 = 1.625$

$V(X) = 1.625 - 1.15^2 = 0.3025$

$\sigma_X = \sqrt{(V(X))} = \underline{0.55}$€

8. Demostreu utilitzant les probabilitats adients si la variable X és independent o no de si surt cara o creu amb la moneda extreta. (1.5pt)

Una possible manera és comprovar si es satisfà: P(cara | X=1) = P(cara). Si no és així ja estaria demostrat que no són independents, però si fos cert hauríem de comprovar-ho també amb X=0.50 i X=2.

P(cara | X=1) equival a P(cara | ⊗) = P(cara i ⊗) /P(X=1)

P(cara i ⊗) = P(A i cara i ⊗) + P(B i cara i ⊗) + P(C i cara i ⊗) = $0.4 \cdot 1/2 \cdot 1/2 + 0.3 \cdot 1/3 \cdot 1/2 + 0.3 \cdot 2/3 \cdot 1/2 = 1/4$

P(cara | ⊗) = (1/4) / (1/2) = 1/2 *(en realitat és trivial, ja que si sabem que tenim una moneda normal la probabilitat de cara es 50%)*

Però, com hem vist en l'apartat 5, P(cara) = 59/120, que no és exactament igual a ½ (tot i que està molt a prop). Així i tot, això és suficient per demostrar que no són independents.

NOM:_____ COGNOMS:_____

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

Problema 2 (Bloc B)

A la imaginària ciutat de X, els delictes poden donar-se en qualsevol època de l'any, però són poc freqüents, i no relacionats entre ells. La policia estima que s'hi produeixen 9 delictes anualment en mitjana. L'alcalde vol conèixer com de tranquil es pot considerar a X un mes qualsevol (per simplificar, preneu que els 12 mesos d'un any tenen 30 dies).

8. Sigui D el nombre de delictes comesos a X en un any. Quin model de probabilitat és adequat per a D, i perquè? Especifiqueu tots els paràmetres que siguin necessaris. (1pt)

   D ~ Poisson(9). Perquè l'esdeveniment (un delicte) és esporàdic, independent del moment i d'altres delictes que s'hagin produït.

9. Sigui R el nombre de mesos d'un període d'un any en els que la policia de X ha registrat algun delicte. Determineu el model de probabilitat adient a R, amb tots els paràmetres necessaris. Calculeu el nombre esperat de mesos *conflictius* al cap d'un any. (1.5pts)

   Quina és la probabilitat *p* que un mes qualsevol es produeixi un delicte? Si D s'aplica a un any, a un mes hem de prendre com a taxa la part proporcional, ja que el model continua sent Poisson:

   M ~ Poisson(9/12), nombre de delictes comesos en un mes. Llavors, $p = P(M > 0) = 1 - P(M=0) = 1 - \exp(-3/4) = 0.5276$

   Finalment, tenim que R ~ B(12, p), ja que tots els mesos són independents uns dels altres.

   $E(R) = 12 \times 0.5276 = 6.3316$

10. Calculeu la probabilitat que en un any tinguem un màxim de tres delictes a X. (1pt)

    $P(D \le 3) = P(D=0) + P(D=1) + P(D=2) + P(D=3) = \exp(-9)[\,1 + 9/1! + 9^2/2! + 9^3/3!\,] = \exp(-9)[1 + 9 + 81/2 + 243/2] =$ **0.02123**

11. Calculeu la probabilitat que en un any tinguem un màxim de tres mesos amb algun delicte. (1pt)

    $P(R \le 3) = P(R=0) + P(R=1) + P(R=2) + P(R=3) = \binom{12}{0} p^0(1-p)^{12} + \binom{12}{1} p^1(1-p)^{11} + \binom{12}{2} p^2(1-p)^{10} + \binom{12}{3} p^3(1-p)^9$

    = 0.000123410 + 0.001654187 + 0.010162494 + 0.037838347 = **0.04978**

El trajecte recomanat per una aplicació per anar de la ciutat de X a la ciutat de Y consisteix en: 1) F_1, agafar un tren, que suposa un temps amb distribució U[100, 200]; 2) F_2, caminar, que suposa un temps amb distribució N(100, 25); 3) F_3, agafar un autobús, que suposa un temps amb distribució Exp(1/60). Tots els temps són minuts. Per qüestions diverses, els temps del tren i el autobús estan correlacionats amb ρ=0.5; caminar és independent dels altres mitjans.

12. Es demana que trobeu el temps esperat del trajecte complet, així com la seva desviació tipus, detallant el desenvolupament necessari. (2pts)

    $T = F_1 + F_2 + F_3 \Rightarrow E(T) = E(F_1) + E(F_2) + E(F_3)$

    $E(F_1) = (100+200)/2 = 150$

    $E(F_2) = 100$ (directe)

    $E(F_3) = 1/(1/60) = 60$, per tant, $E(T) =$ **310** minuts.

    Trobar la variància no és tan directe, per la correlació existent entre F_1 y F_3.

    $V(F_1 + F_3) = V(F_1) + V(F_3) + 2\,\rho\,\sigma_1\,\sigma_3$; $V(T) = V(F_1 + F_2 + F_3) = V(F_1 + F_3) + V(F_2)$

    $V(F_1) = (200-100)^2/12 = 833{,}33...$; $\sigma_1 = 28{,}8675$

    $V(F_2) = 25^2 = 625$; $\sigma_2 = 25$

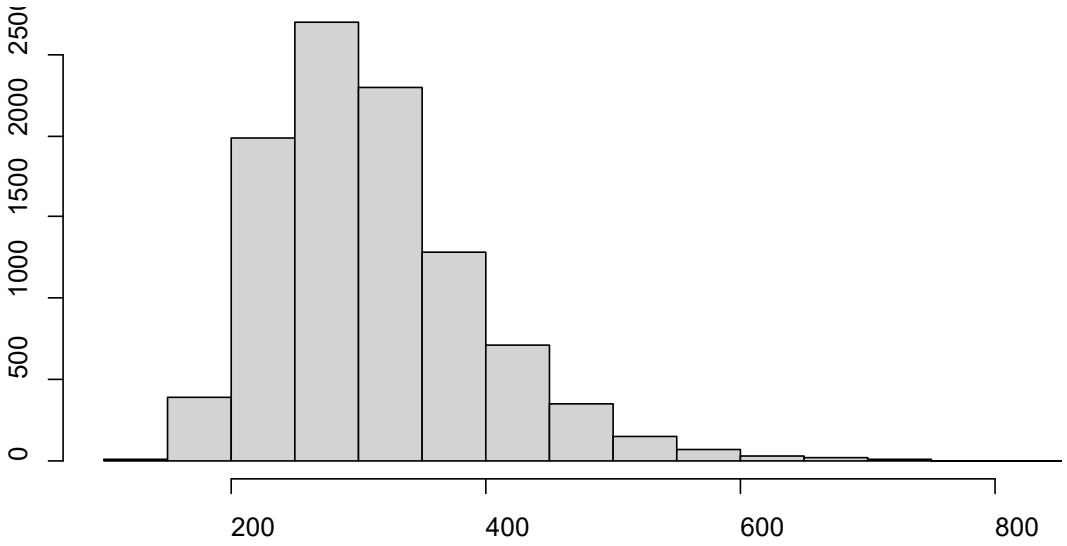
    $V(F_3) = 1/(1/60)^2 = 3600$; $\sigma_3 = 60$

    $V(F_1 + F_3) = 833{,}33 + 3600 + 2 \cdot 0{,}5 \cdot 28{,}8675 \cdot 60 = 6165{,}38$

    $V(F_1 + F_2 + F_3) = 6165{,}38 + 625 = 6790{,}38$; $\sigma_T =$ **82.404** minuts

13. La distribució de T és una Normal, o aproximadament Normal? Justificar la resposta. Tant si la resposta és sí o si és no, suposeu-ho cert per trobar la probabilitat que un viatge de X a Y es pugui fer en menys de quatre hores. (1.5pts)

    No és Normal, i podem sospitar que la suma d'una variable Exponencial amb una Uniforme i una Normal tampoc s'assemblarà massa a una Normal, per l'extrema asimetria de la Exponencial. Si fossin moltes variables independents es podria aplicar el teorema del límit central, però tres són un nombre molt baix.



Una distribució simulada de la variable T

Però *imaginem que sí segueix* una Normal, només per a poder fer el càlcul demanat. Llavors, T ~ N(310, 82.4), i la qüestió és: P(T < 240), passant el temps d'hores a minuts.

$$P(T \leq 240) = P\left(Z \leq \frac{240 - 310}{82.4}\right) = P(Z \leq -0.8495) = 1 - P(Z \leq 0.8495) = 1 - 0.8023 = 0.1977$$

(amb la taula de valors de probabilitats de la Normal estàndard de sota podríem prendre el valor 0.80, ja que el quantil de 0.80 (0.8416) és proper al valor 0.8495 per al que es vol calcular la solució; llavors, 1-0.80 = 0.20 és una resposta vàlida)

14. Un viatger fa aquest trajecte 25 vegades en un any. Un dia, aquest individu vol calcular quant temps passa en total a l'any en el tren (comptant només els viatges d'aquests trajectes). Com el temps de cada viatge és aleatori, busca un interval centrat a la mitjana que contengui amb probabilitat 0.80 el temps *en hores* total anual dels seus viatges en tren. Calculeu-ho i expliqueu tot el procediment. (2pts)

Diguem G_1, G_2, ... G_{25} als temps de viatge en tren per cada un dels trajectes, tots ells U[100, 200] i suposadament independents uns dies dels altres (és raonable). El temps total *en minuts* (de moment) és:

$$H = \sum_{i=1}^{25} G_i$$

25 és un nombre prou alt, tenint en compte el tipus de distribució de les variables G: la uniforme és una distribució continua i simètrica, la qual cosa facilita la convergència de H cap a un model Normal, tal como diu el TCL.
Ara obtindrem la mitjana i la desviació tipus:
E(H) = 25 E(F_1) = 3750 min = 62,5 hores
V(H) = 25 V(F_1) = 20833,33 min^2 ; σ_H = 144,34 min = 2,4056 hores. Per tant, el temps total *en hores* és aproximadament:

$$H_h \approx N(62{,}5;\ 2{,}4056)$$

La solució és: 62,5 ± $z_{0.90}$ 2.4056 = 62,5 ± 1,2816 · 2.4056 = **(59,42; 65,58).** Amb probabilitat 80%, aquest és el temps total que el viatger passa en el tren en un any.

```
> p = seq(0.60, 0.99, by=0.01)
> round(qnorm(p), 4)
 [1] 0.2533 0.2793 0.3055 0.3319 0.3585 0.3853 0.4125 0.4399 0.4677 0.4959
[11] 0.5244 0.5534 0.5828 0.6128 0.6433 0.6745 0.7063 0.7388 0.7722 0.8064
[21] 0.8416 0.8779 0.9154 0.9542 0.9945 1.0364 1.0803 1.1264 1.1750 1.2265
[31] 1.2816 1.3408 1.4051 1.4758 1.5548 1.6449 1.7507 1.8808 2.0537 2.3263
```

17-06-2026

NOM: _____ COGNOM: _____

Contesteu cada pregunta en els espais en blanc. Expliqueu i justifiqueu els càlculs.

BLOC C. Una empresa està desenvolupant un sistema basat en intel·ligència artificial per respondre preguntes sobre documentació jurídica. Abans de desplegar-lo, es duen a terme diversos estudis per avaluar-ne el rendiment.

Es realitza una primera prova pilot on es selecciona una mostra aleatòria de 10 documents jurídics i es calcula el percentatge de respostes correctes del sistema en cadascun d'ells.

Document	1	2	3	4	5	6	7	8	9	10
Precisió (%)	78	82	75	89	84	77	86	80	83	78

$\sum x_i = 812$

$\sum x_i^2 = 66108$

1. (1.5 pts.) Calcula els estimadors puntuals de la mitjana poblacional i de la desviació estàndard poblacional.
   Primer calculem la mitjana mostral, com a estimador puntual de la mitjana poblacional:

$$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n} = \frac{812}{10} = 81.2$$

   Per estimar la desviació estàndard poblacional utilitzem la variància mostral:

$$\hat{\sigma} = s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{66108 - 10 \cdot 81.2^2}{9}} = \sqrt{\frac{173.6}{9}} = \sqrt{19.29} \approx 4.39$$

2. (1.5 pts.) Calcula un interval de confiança del 95% per a la precisió mitjana del sistema.

   Com que no coneixem la desviació estàndard poblacional teòrica i $n = 10$, fem servir la t de Student amb 9 graus de llibertat. Així doncs:

$$IC_{95\%}(\mu) = \hat{\mu} \pm t_{9,0.975}\frac{s}{\sqrt{n}} = 8.12 \pm 2.262 \cdot \frac{4.39}{\sqrt{10}} = 81.2 \pm 3.14 = [78.06,\ 84.34]$$

Després de l'estudi pilot, l'empresa decideix ampliar la mostra. El nou interval de confiança del 95% per a la precisió mitjana del sistema és:

$$IC_{95\%}(\mu) = [79.2,\ 83.2]$$

3. (1.5 pts.) Suposa que en el nou estudi s'assumeix coneguda la desviació estàndard poblacional i que aquesta coincideix amb la desviació estàndard estimada en l'estudi pilot. Determina la mida mostral utilitzada en aquest nou estudi.

L'amplada de l'interval és $A = \frac{83.2-79.2}{2} = 2$. Suposem que la desviació estàndard poblacional és la mateixa que l'estimada en l'estudi pilot, és a dir, $\sigma \approx 4.39$. Per a un interval de confiança del 95% amb desviació coneguda tenim $A = z_{0.975}\frac{\sigma}{\sqrt{n}}$ i, per tant,

$$2 = 1.96\frac{4.39}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{1.96 \cdot 4.39}{2} = 4.302 \Rightarrow n = (4.302)^2 = 18.5$$

Com que la mida mostral ha de ser un nombre enter, determinem $n = 19$.

Per analitzar si el comportament del sistema és igual d'estable en diferents entorns, es prova en dos departaments de l'empresa. Les puntuacions obtingudes són les següents:

	1	2	3	4	5	6	7	8	9	10
Dept. A	82	76	85	79	88	81	74	83	80	86
Dept. B	91	72	84	69	95	78	88	73	81	98

$\sum x_A = 814, \sum x_A^2 = 66332$

$\sum x_B = 829, \sum x_B^2 = 69629$

4. (2 pts.) Calcula un interval de confiança del 99% per al quocient de variàncies σ_A^2/σ_B^2.

$$\bar{x}_A = \frac{\sum x_A}{10} = \frac{814}{10} = 81.4, \quad s_A^2 = \frac{\sum x_A^2 - n\bar{x}_A^2}{n-1} = \frac{66332 - 10(81.4)^2}{9} = \frac{172.4}{9} = 8.04.$$

Anàlogament,

$$\bar{x}_B = \frac{\sum x_B}{10} = \frac{829}{10} = 82.9, \quad s_B^2 = \frac{\sum x_B^2 - n\bar{x}_B^2}{n-1} = \frac{69629 - 10(82.9)^2}{9} = \frac{958.9}{9} = 100.54.$$

Així,

$$\frac{s_A^2}{s_B^2} = \frac{8.04}{100.54} = 0.08.$$

L'interval de confiança per al quocient de variàncies és

$$IC_{99\%}\left(\frac{\sigma_A^2}{\sigma_B^2}\right) = \left[\frac{s_A^2/s_B^2}{F_{9,9,0.995}}, \frac{s_A^2/s_B^2}{F_{9,9,0.005}}\right] = \left[\frac{0.08}{6.54}, \frac{0.08}{0.153}\right] = [0.012, 0.523]$$

5. (1 pt.) A partir de l'interval obtingut, és raonable assumir que ambdós departaments tenen la mateixa variabilitat? Justifica la resposta.

si les variàncies poblacionals fossin iguals, tindríem $\frac{\sigma_A^2}{\sigma_B^2} = 1$. Com que el valor 1 queda fora de l'interval de confiança, hi ha evidència (amb un nivell de confiança del 99%) que les variàncies dels dos departaments són diferents. En particular, com que tots els valors de l'interval són inferiors a 1, es conclou que la variabilitat del Departament A és inferior a la del Departament B.

Finalment, l'empresa compara dues versions del sistema d'IA. Per a cadascuna d'elles es registra si la resposta a una pregunta és correcta o incorrecta. L'anàlisi realitzada amb R ha produït el següent resultat:

```
2-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 2.5606, df = 1, p-value = 0.1096
alternative hypothesis: two.sided

95 percent confidence interval:
 -0.0779 0.0079

sample estimates:
   prop 1    prop 2
    0.840     0.875
```

6. (1 pt.) Quin paràmetre està estimant l'interval de confiança anterior?

L'interval de confiança està estimant la diferència entre les proporcions poblacionals d'encert de les dues versions del sistema: $\pi_1 - \pi_2$, on π_1 i π_2 representen les probabilitats que les versions 1 i 2 responguin correctament una pregunta.

7. (1.5 pts.) Interpreta l'interval de confiança en el context del problema.

L'interval de confiança del 95% és $[-0.0779, -0.0079]$. Això indica que, amb un nivell de confiança del 95%, la diferència entre les proporcions d'encerts de les dues versions del sistema d'IA ($\pi_1 - \pi_2$) es troba entre -0.0779 i 0.0079.
És a dir, la proporció d'encerts de la versió 1 podria ser fins a un 7,79% inferior a la de la versió 2 o bé fins a un 0,79% superior. Com que l'interval conté el valor 0, no hi ha evidència suficient per afirmar que les dues versions tinguin una proporció d'encerts diferent.

$qnorm(0.80) = 0.8416$	$qnorm(0.90) = 1.2816$	$qt(0.90, 9) = 1.383$	$qt(0.95, 9) = 1.833$	$qf(0.95, 9, 9) = 3.18$
$qnorm(0.95) = 1.6449$	$qnorm(0.975) = 1.9600$	$qt(0.975, 9) = 2.262$	$qt(0.99, 9) = 2.821$	$qf(0.995, 9, 9) = 6.54$
$qnorm(0.99) = 2.3263$	$qnorm(0.995) = 2.5758$	$qt(0.995, 9) = 3.250$	$qt(0.999, 9) = 4.297$	$qf(0.005, 9, 9) = 0.153$

Nom i cognoms:
Contesteu cada pregunta en els espais en blanc. Expliqueu i justifiqueu els càlculs

BLOC D. Un grup d'estudiants de l'UPC estan estudiant diverses formes per comprimir eficientment fitxers GIS (dades geoespecials). En concret, han proposat 2 mètodes: A i B.
Per veure quin dels dos mètodes funciona millor, han decidit realitzar un estudi aparellat:

$$D = \%\text{ comprimit mètode A} - \%\text{ comprimit mètode B}$$

Utilitzant fitxers escollits de forma equiprobable i aleatòria. D'aquests fitxers sabem el tipus de terreny que descriuen: Ciutat, Camp o Muntanya, i la mida del fitxer (MB).
Inicialment, han proposat aquest model:

```
Call:
lm(formula = D ~ 1, data = dades)
Residuals:
     Min       1Q   Median       3Q      Max
-12.9333  -3.6933  -0.2383   4.0142  13.3367
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5733     0.4460   1.285    0.201
Residual standard error: 5.463 on 149 degrees of freedom
```

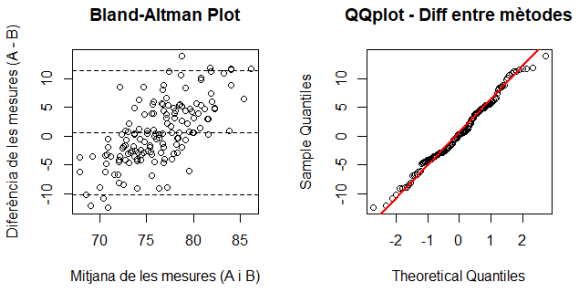


Figura 1

1 Podem afirmar estadísticament que existeix un mètode millor que l'altre? Indiqueu també l'interval del valor de la diferència del percentatge comprimit entre el mètode A i B amb un 95% de confiança. (1.5 punts)

No podem afirmar que un mètode és millor que l'altre, ja que l'intercept té una estimació de 0.57 i el seu p-valor és de 0.201, que ens indica que el valor de 0 és versemblant.
$IC = 0.5733 \pm 1.95996 * 0.4460 = (-0.301, 1.447)$
En conclusió, no podem assegurar que un mètode sigui millor que l'altre, degut a que, amb un 95% de confiança, la diferència pot ser entre 1.45% a favor de A o 0.3% a favor de B.

2 Llisteu les premisses que assumeix el model. (1 punt)
Les premisses que assumeix el model són:

- Mostra aleatòria.
- Normalitat de la diferència de mètodes.
- Existeix un efecte aditiu constant entre els dos mètodes.

3 Donats els gràfics anteriors, penseu que aquestes premisses es compleixen? Argumenteu la resposta. (1.5 punts)
Per valorar les premisses disposem de dos gràfics: Bland-Altman i QQplot.
El QQplot ens informa que efectivament les diferències segueixen una normal. Encara que les cues difereixen una mica, acceptem la normalitat.
El gràfic Bland-Altman, ens informa sobre la distribució de les diferències entre parelles, segons la magnitud d'A i B. En aquest cas, observem una diferència que no és constant, i que tendeix a crèixer amb el percentatge de compressió. Esperaríem uns punts distribuits a l'atzar al voltant d'una mitjana constant.
La premissa de la mostra aleatòria depèn del disseny de la recollida de dades. Segons l'anunciat, podem pensar que aquesta premissa es compleix.

Donat el següent fragment de codi: (El model 2 es descriu a la següent pàgina)

```
nova_obs <- data.frame( Entorn = "Muntanya", Mida_fitxer_MB = 667)
predict(model2, newdata = nova_obs, interval = "confidence")
predict(model2, newdata = nova_obs, interval = "prediction")
```

4 Quina és la diferència entre aquestes dues comandes? Què podem esperar dels resultats? (1.5 punts)
Les dues comandes retornen un interval de confiança del 95% de la diferència del percentatge comprimit entre els mètodes A i B per una observació nova d'un fitxer GIS que descriu un terreny de muntanya i que té una mida de 667MB. La diferència és que el primer interval serà per la mitjana esperada d'un fitxer d'aquestes característiques i

el segon, serà l'interval de confiança per una única predicció (una observació individual), amb molta més variabilitat. Podem esperar una diferència positiva (en concret, d'un 3.3% entre els dos mètodes: 0.0168833*667 −8.760868 +0.80154 = 3.302), ja que es tracta d'un fitxer d'un pes elevat, i que el segon interval tindrà més amplitud que el primer.

A continuació, els estudiants han decidit afegir més variables explicatives al model:

```
Call:
lm(formula = D ~ Entorn + Mida_fitxer_MB, data = dades)

Residuals:
   Min 1Q Median 3Q Max
-6.022 -1.761 -0.060 1.735 6.752

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.7608680 0.5384028 -16.272 <2e-16 ***
EntornCiutat -0.5208539 0.5368944 -0.970 0.334
EntornMuntanya 0.8015401 0.5295355 1.514 0.132
Mida_fitxer_MB 0.0168833 0.0007702 21.920 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 2.647 on 146 degrees of freedom
Multiple R-squared: 0.77
Adjusted R-squared: 0.7652
F-statistic: 162.9 on 3 and 146 DF
p-value: < 2.2e-16
```

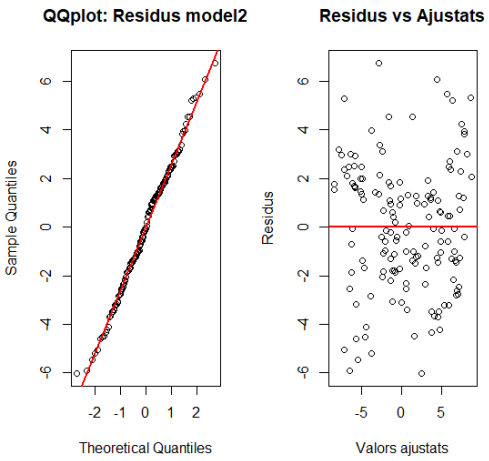


Figura 2

5 Es pot calcular el percentatge de compressió esperat del mètode A per un fitxer que descriu una Ciutat i té una mida de 426 MB únicament amb la informació proporcionada pel model? (1 punt)
No, amb les dades proporcionades podem calcular el valor esperat de la diferència del percentatge comprimit entre el mètode A i el mètode B. Sense el percentatge comprimit del mètode B és impossible poder calcular el percentatge comprimit del mètode A.

6 Interpreteu el coeficient que fa referència a la mida del fitxer, com afecta en el valor final de D? Relacioneu-lo amb el gràfic Bland-Altam del primer model (1.5 punts)
El coeficient és 0.01688, amb un error estàndar de $7.702 \cdot 10^{-4}$. Inferim que és un paràmetre significatiu (diferent a 0) ja que el seu error estàndar és molt inferior a l'estimació del coeficient. Com que el valor estimat és positiu, a mesura que augmenti la mida del fitxer, la diferència del percentatge comprimit entre el mètode A i B augmentarà. Més concretament, la diferència percentual augmentarà un 0.017 entre el mètode A i B per cada MB que augmenti de mida el fitxer.
Si mirem el Bland-Altman anterior, ens fixem que la diferència sol ser negativa per fitxers petits, i positiva per fitxers alts, així que podem llençar la hipòtesis que el mètode B funciona millor per fitxers petits, i el mètode A per fitxers grans.

7 Quines premisses està assumint el model? Donats els gràfics i la informació disponible, creieu que es compleixen? (2 punts)
En aquest model, assumim les següents premisses:

- Les dades (en aquest cas les diferències) provenen d'una mostra aleatòria.
- Linealitat: Les dades són explicades per un efecte lineal de les variables explicatives.
- Normalitat dels residus: Els residus produïts pel model han de seguir una distribució normal.
- Homoscedasticitat dels residus: Els residus han de tenir una variabilitat constant.

Amb els gràfics proporcionats podem valorar les últimes 3 premisses. Amb el QQplot valorem que efectivament els residus segueixen una distribució normal, i amb el gràfic de Residus vs Ajustats, observem una variància constant i els residus estan centrats al 0, així podem inferir un efecte de linealitat.
En resum, el model sí que compleix les premisses que assumeix.