NOM:_____COGNOMS:_____

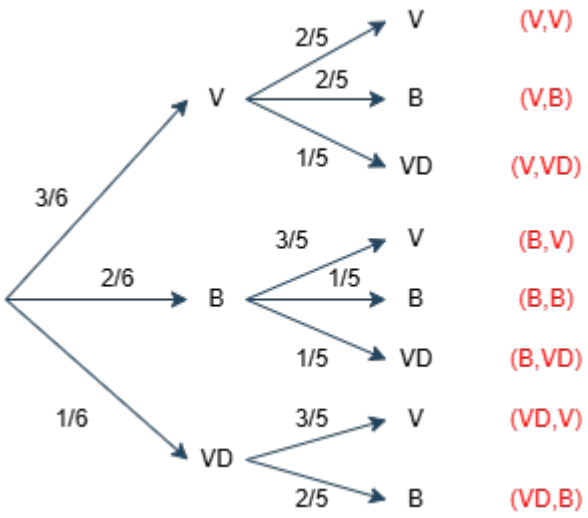
Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs.

Problema 1 (Bloc A)

L'Àlex i la Diana s'han aficionat a un joc amb una bossa opaca que conté 6 boles de diferents colors: 3 vermelles, 2 blaves i 1 verda. En una ronda del joc un jugador extreu dues boles aleatòriament, sense reposició. El jugador guanya si obté una bola verda o dues boles del mateix color; en qualsevol altre cas, perd. Un cop finalitzada la ronda, les boles es tornen a introduir a la bossa per a que l'altre jugador provi sort.

1) **(1 punt)** Dibuixa l'arbre de probabilitats del procés corresponent a una ronda del joc.

Definim els esdeveniments següents:
V = "Extreure una bola vermella", B = "Extreure una bola blava" i VD = "Extreure una bola verda".



2) **(1.5 punts)** Quina és la probabilitat que un jugador guanyi?

Si definim els següents esdeveniments:
G = "Guanyar el joc"
A = "Extreure una bola verda"
M= "Extreure dues boles del mateix color"

Fent càlculs a partir de les probabilitats recollides a l'arbre obtenim:

P(A)= P((VD,V)) + P((VD,B)) + P((V,VD)) + P((B,VD)) =

$$= \frac{1}{6}\cdot\frac{3}{5} + \frac{1}{6}\cdot\frac{2}{5} + \frac{3}{6}\cdot\frac{1}{5} + \frac{2}{6}\cdot\frac{1}{5} = \frac{3+2+3+2}{30} = \frac{10}{30} = \mathbf{\frac{1}{3}}$$

P(M)= P((V,V)) + P((B,B)) = $\frac{3}{6}\cdot\frac{2}{5} + \frac{2}{6}\cdot\frac{1}{5} = \frac{6+2}{30} = \frac{8}{30} = \mathbf{\frac{4}{15}}$

Així doncs,

P(G) =P (A ∪M) = P(A) + P(M) = $\frac{10}{30} + \frac{8}{30} = \frac{18}{30} = \mathbf{\frac{3}{5}}$**= 0.6**

3) **(1punt)** Si el jugador guanya, quina és la probabilitat que hagi obtingut una bola verda?

Pel Teorema de Bayes i utilitzant els càlculs de l'apartat anterior

P(A | G) = (P(G | A) · P(A)) / P(G) = $\left(1 \cdot {}^{1}/_{3}\right)/\left({}^{3}/_{5}\right) = \mathbf{\frac{5}{9}} \approx \mathbf{0.56}$

4) **(1 punt)** L'Àlex juga una ronda i després és el torn de la Diana. Quina és la probabilitat que ambdós guanyin el joc? Justifica la resposta.

Com que els esdeveniments que cada jugador guanyi són independents entre ells (perquè el joc comença de nou des de 0):

P({"Àlex guanya"} ∩ {"Diana guanya"}) =P({"Àlex guanya"})· P({"Diana guanya"}) = $P(G)^2 = \left(\frac{3}{5}\right)^2 = \mathbf{\frac{9}{25}}$**= 0.36**

Una modalitat alternativa del joc és donar puntuacions a cadascuna de les boles. Les boles vermelles valen 1 punt, les blaves 2 punts i la verda 3 punts. Definim la variable aleatòria X="Puntuació total obtinguda en una ronda".

5) (**1.5 punts**) Defineix la funció de probabilitat de la variable X.

Primer de tot cal trobar el conjunt de resultats possibles per X, Ω_X.

Prenem cadascuna de les possibles combinacions d'extraccions de dues boles (esdeveniments representats en l'arbre) i veiem quin valor de la variable aleatòria X li correspon.

ω	$P(\omega)$	X
(V,V)	$\frac{3}{6}\cdot\frac{2}{5}=\frac{1}{5}$	2
(V, B)	$\frac{3}{6}\cdot\frac{2}{5}=\frac{1}{5}$	3
(V, VD)	$\frac{3}{6}\cdot\frac{1}{5}=\frac{1}{10}$	4
(B, V)	$\frac{2}{6}\cdot\frac{3}{5}=\frac{1}{5}$	3
(B, B)	$\frac{2}{6}\cdot\frac{1}{5}=\frac{1}{15}$	4
(B, VD)	$\frac{2}{6}\cdot\frac{1}{5}=\frac{1}{15}$	5
(VD, V)	$\frac{1}{6}\cdot\frac{3}{5}=\frac{1}{10}$	4
(VD, B)	$\frac{1}{6}\cdot\frac{2}{5}=\frac{1}{15}$	5

Utilitzant les probabilitats de cadascun dels esdeveniments (que es poden extreure de l'arbre de probabilitats), tenim que la funció de probabilitat de X és la següent:

X	P_X
2	$\mathbf{\frac{1}{5}}$
3	$\frac{1}{5}+\frac{1}{5}=\mathbf{\frac{2}{5}}$
4	$\frac{1}{10}+\frac{1}{15}+\frac{1}{10}=\frac{8}{30}=\mathbf{\frac{4}{15}}$
5	$\frac{1}{15}+\frac{1}{15}=\mathbf{\frac{2}{15}}$

6)(**1 punt**) Quina és la esperança i variància de la puntuació total d'un jugador?

$E(X) = 2\cdot\frac{1}{5}+\ 3\cdot\frac{2}{5}+\ 4\cdot\frac{4}{15}+\ 5\cdot\frac{2}{15}=\ \frac{50}{15}=\ \mathbf{\frac{10}{3}}\approx \mathbf{3.33}$

$E(X^2) = 2^2\cdot\frac{1}{5}+\ 3^2\cdot\frac{2}{5}+\ 4^2\cdot\frac{4}{15}+\ 5^2\cdot\frac{2}{15}=\ 12$

$Var(X) = E(X^2) - \ E(X)^2 = 12 - \left({}^{10}/_{3}\right)^2 = \mathbf{\frac{8}{9}}\approx \mathbf{0.89}$

Definim les variables aleatòries: X_1 = "Puntuació obtinguda en la primera extracció" i X_2 ="Puntuació obtinguda en la segona extracció".

7) (**1 punt**) Troba la funció de probabilitat conjunta de X_1 i X_2, $P_{X_1,X_2}(x_1,x_2)$.

Mostrem la funció de probabilitat conjunta de X_1 i X_2 mitjançant la següent taula, d'on els valors de les probabilitats es poden obtenir de l'arbre de probabilitats de l'apartat 1) o dels càlculs de probabilitats dels esdeveniments de l'apartat 5).

$X_1 \backslash X_2$	1	2	3
1	$\frac{3}{6}\cdot\frac{2}{5}=\mathbf{\frac{1}{5}}$	$\frac{3}{6}\cdot\frac{2}{5}=\mathbf{\frac{1}{5}}$	$\frac{3}{6}\cdot\frac{1}{5}=\mathbf{\frac{1}{10}}$
2	$\frac{2}{6}\cdot\frac{3}{5}=\mathbf{\frac{1}{5}}$	$\frac{2}{6}\cdot\frac{1}{5}=\mathbf{\frac{1}{15}}$	$\frac{2}{6}\cdot\frac{1}{5}=\mathbf{\frac{1}{15}}$
3	$\frac{1}{6}\cdot\frac{3}{5}=\mathbf{\frac{1}{10}}$	$\frac{1}{6}\cdot\frac{2}{5}=\mathbf{\frac{1}{15}}$	**0**

8) (**0.5 punts**) Calcula les probabilitats marginals de X_1 i X_2.

Sumant files i columnes de la taula de l'exercici anterior obtenim

x_1	P_{X_1}
1	$\frac{1}{5}+\frac{1}{5}+\frac{1}{10}=\frac{5}{10}=\mathbf{\frac{1}{2}}$
2	$\frac{1}{5}+\frac{1}{15}+\frac{1}{15}=\frac{5}{15}=\mathbf{\frac{1}{3}}$
3	$\frac{1}{10}+\frac{1}{15}+0=\frac{5}{30}=\mathbf{\frac{1}{6}}$

Per simetria, el resultat per X_2 és exactament el mateix.

9) (**1 punt**) Calcula la covariància $\mathrm{Cov}(X_1, X_2)$.

Utilitzant les funcions de probabilitat marginals de l'apartat anterior, calculem l'esperança de X_1 i X_2:

$$E(X_1) = E(X_2) = 1\cdot\frac{1}{2} + 2\cdot\frac{1}{3} + 3\cdot\frac{1}{6} = \frac{10}{6} = \mathbf{\frac{5}{3}}$$

O també, veient que X = $X_1 + X_2$ i que tenen la mateixa esperança perquè tenen la mateixa funció de probabilitat, aleshores $E(X_1) = E(X_2) = E(X)/2 = \mathbf{\frac{5}{3}}$

Per trobar la covariància, calculem també $E(X_1 \cdot X_2)$:

$$E(X_1\cdot X_2) = \sum_{i,j=1}^{3} i\cdot j\cdot P_{X_1,X_2}(i,j) = 1\cdot1\cdot\frac{1}{5} + 1\cdot2\cdot\frac{1}{5} + 1\cdot3\cdot\frac{1}{10} + 2\cdot1\cdot\frac{1}{5} + 2\cdot2\cdot\frac{1}{15} + 2\cdot3\cdot\frac{1}{15} + 3\cdot1\cdot\frac{1}{10} + 3\cdot2\cdot\frac{1}{15} + 0$$
$$= \frac{80}{30} = \mathbf{\frac{8}{3}}$$

$$\mathrm{Cov}(X_1, X_2) = E(X_1\cdot X_2) - E(X_1)\cdot E(X_2) = \frac{8}{3} - \frac{5}{3}\cdot\frac{5}{3} = \mathbf{-\frac{1}{9}} \approx \mathbf{-0.11}$$

Alternativament, V(X) = $V(X_1 + X_2) = V(X_1) + V(X_2) + 2\cdot\mathrm{Cov}(X_1, X_2)$.

V(X_2) = V(X_1) = E(X_1^2) $- E(X_1)^2 = \left(1^2\cdot\frac{1}{2} + 2^2\cdot\frac{1}{3} + 3^2\cdot\frac{1}{6}\right) - \left(\frac{5}{3}\right)^2 = \frac{10}{3} - \frac{25}{9} = \frac{5}{9}$

Així doncs, $\mathrm{Cov}(X_1, X_2) = (V(X) - V(X_1) - V(X_2))/2$ = (8/9 – 2 · 5/9)/2 = $\mathbf{-\frac{1}{9}} \approx \mathbf{-0.11}$

10) (**0.5 punts**) Interpreta el signe del resultat obtingut en el context en què ens trobem.

El signe negatiu de la covariància entre X_1 i X_2 indica una relació inversa entre ambdues variables, és a dir, valors més grans de X_1 tendeixen a anar associats amb valors més petits de X_2.

En el context en què ens trobem, com que es realitzen extraccions consecutives sense reemplaçament, té sentit que si en la primera extracció s'obté una puntuació alta, la probabilitat d'obtenir una puntuació alta en la segona extracció disminueix, ja que queden menys boles amb puntuació elevada a la bossa.

Problema 3 (Bloc C)

Un grup d'estudiants de l'assignatura Estadística i Probabilitat vol comparar el rendiment de dues estratègies d'indexació per a consultes SQL complexes: l'estratègia **A (*Hash Index*)** i l'estratègia **B (*B-Tree*)**. El rendiment es mesura pel **Temps de Resposta** (T) en mil·lisegons. Els estudiants seleccionen 35 consultes aleatòries i les executen sota ambdues estratègies en condicions idèntiques. Definim a continuació les següents variables aleatòries:

- T_A: temps de resposta amb *Hash index*,
- T_B: temps de resposta amb *B-Tree*,
- $D = T_A - T_B$: diferència de temps de resposta entre *Hash index* i *B-Tree*.

Disposem de les següents dades resumides:

$$\sum_{i=1}^{35} D_i = -125.5, \qquad \sum_{i=1}^{35} D_i^2 = 980.25.$$

(a) Es tracta d'un disseny d'estudi de mostres independents o aparellades? Raoneu la resposta. **(0.5 punts)**

Solució

Es tracta d'un disseny de **mostres aparellades** donat que totes les consultes s'executen amb ambdues estratègies.

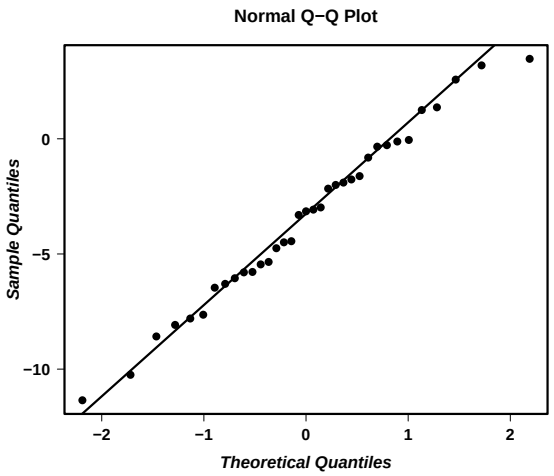
(b) Calculeu un interval de confiança (IC) al 98% per a la diferència $\boldsymbol{\mu_{T_B} - \mu_{T_A}}$. Indiqueu l'estimació puntual, l'error estàndard i la conclusió de l'IC. **(2 punts)**

Solució

- **Estimació puntual:** $\hat{\mu}_{D'} = \frac{1}{35}\sum_{i=1}^{35}(-D_i) = \frac{125.5}{35} = 3.59$.
- **Error estàndard:** $se(\hat{\mu}_{D'}) = \frac{1}{35}\sqrt{\frac{1}{34}\left(\sum_{i=1}^{35} D_i^2 - 35 \cdot \hat{\mu}_{D'}^2\right)} = 0.67$.
- **Interval de confiança:** $\text{IC}(\mu_{D'};\, 0.98) = 3.59 \mp 2.441 \cdot 0.67 = [1.95.\, 5.23]$.

Pel fet que l'IC no conté el 0, podem assumir, amb un nivell de confiança del 98% que, en mitjana, l'estratègia A és més ràpid que l'estratègia B. En concret, podem estar força segurs que la diferència mitjana es troba entre 2.52 i $4.66\,\text{ms}$.

(c) Quina premissa del càlcul de l'IC anterior es pot comprovar amb el següent gràfic? Es pot suposar que es compleix aquesta premissa? Raoneu la resposta. **(1 punt)**



Solució

El QQ-plot permet comprovar gràficament si una mostra prové d'una variable amb distribució normal; en aquest cas, els punts se situen aproximadament sobre una recta. En el nostre cas, el fet que els punts estiguin per sobre o molt a prop de la recta del QQ-plot suggereix que es compleix la premissa de normalitat de D'.

(d) L'equip vol repetir l'estudi de l'estratègia A de forma independent per estudiar el seu temps de resposta mitjà (μ_{T_A}). Si sabem per estudis previs que la variància poblacional és $\mathrm{V}(T_A) = \sigma^2_{T_A} = 64\,\mathrm{ms}$, quina hauria de ser la mida mínima de la mostra n perquè l'error màxim de l'estimació (és a dir, el marge d'error de l'IC amb confiança del 95%) no superi els $2\,\mathrm{ms}$? **(1 punt)**

Solució

Marge d'error: $z_{0.975} \cdot \frac{\sigma_{T_A}}{\sqrt{n}}$. Per tant,

$$1.96 \cdot \frac{8}{\sqrt{n}} \leq 2 \iff n \geq \left(1.96 \cdot \frac{8}{2}\right)^2 = 61.5.$$

És a dir, la mida mínima de la mostra ha de ser de **62** consultes.

(e) Quin seria l'error estàndard de l'estimador $\hat{\mu}_{T_A}$ en aquest cas? **(0.5 punts)**

Solució

Error estàndard: $se(\hat{\mu}_{T_A}) = \frac{8}{\sqrt{62}} = 1.02$.

El mateix grup d'estudiants també vol analitzar la variabilitat del sistema i per tant fan 40 execucions més en un servidor secundari (C). Amb aquesta mostra obtenen una variància mostral de $s^2_{T_C} = 45.2$. Se sap que la variància mostral de l'estratègia B en el servidor principal (amb $n = 35$) va ser $s^2_{T_B} = 30.8$.

(f) Calculeu l'IC al 95% per al quocient de desviacions estàndards $\sigma_{T_C}/\sigma_{T_B}$. A partir del resultat, és raonable suposar que ambdós servidors tenen la mateixa variabilitat? **(2.5 punts)**

Solució

L'IC per a la raó $\sigma^2_{T_B}/\sigma^2_{T_C}$:

$$\mathrm{IC}(\sigma^2_{T_B}/\sigma^2_{T_C}; 0.95) = \left[\frac{30.8/45.2}{1.922},\ \frac{30.8/45.2}{0.513}\right].$$

Per tant,

$$\mathrm{IC}(\sigma_{T_C}/\sigma_{T_B}; 0.95) = \left[1\Big/\sqrt{\frac{30.8/45.2}{0.513}},\ 1\Big/\sqrt{\frac{30.8/45.2}{1.922}}\right] = [0.868,\ 1.679].$$

Pel fet que $1 \in \mathrm{IC}(\sigma_{T_C}/\sigma_{T_B}; 0.95)$ no podem descartar la possibilitat que la variabilitat dels dos servidors sigui la mateixa.

(g) En un altre estudi de 90 consultes aleatòries, els temps de 77 consultes van baixar del llindar de $100\,\mathrm{ms}$. Calculeu l'estimació puntual i un IC al 90% per al percentatge de consultes que compleixen aquest criteri. Doneu una interpretació de l'interval obtingut. **(1.5 punts)**

Solució

Qualsevol dels dos intervals de confiança a continuació es considera correcte:

- **Estimació puntual:** $\frac{77}{90} \cdot 100 = 85.6\%$.
- **Interval de confiança (I):** $\left(\frac{77}{90} \mp 1.645 \cdot \frac{0.5}{\sqrt{90}}\right) \cdot 100 = [76.9,\ 94.2]$.
- **Interval de confiança (II):** $\left(\frac{77}{90} \mp 1.645 \cdot \frac{\sqrt{0.856 \cdot 0.144}}{\sqrt{90}}\right) \cdot 100 = [79.5,\ 91.6]$.

L'IC (II) conté la probabilitat (expressat en percentatge) que el temps de resposta d'una consulta qualsevol estigui per sota del llindar de $100\,\mathrm{ms}$ amb un nivell de confiança del 90%.

(h) Com canviaria aquest IC si augmentéssim la mida mostral? Raoneu la resposta. **(1 punt)**

Solució

- El centre de l'IC, és a dir, l'estimació puntual, podria ser menor, igual o superior a 85.6%.
- L'amplada de l'IC (I) seria més petit en reduir-se l'error estàndard de l'estimació.
- L'amplada de l'IC (II) seria més petit si l'estimació puntual és igual o superior a 85.6%. En canvi, si és més petit, l'amplada podria ser més gran o més petit, ja que tant el numerador com el denominador de l'error estàndard incrementarien.

`qt(0.98, 35)`	2.133	`qt(0.99, 35)`	2.438	`qt(0.98, 34)`	2.136	`qt(0.99, 34)`	2.441
`qnorm(0.9)`	1.282	`qnorm(0.925)`	1.44	`qnorm(0.95)`	1.645	`qnorm(0.975)`	1.96
`qf(0.025, 34, 39)`	0.513	`qf(0.975, 34, 39)`	1.922	`qf(0.05, 39, 34)`	0.579	`qf(0.95, 39, 34)`	1.749

Problema D.

75 directoris trets de Github s'han comprimit amb tres aplicacions (rar, tar i zip), recollint la mida inicial del directori, la final i el tempsde CPU consumit. L'espai es mesura en megabytes (MB). Hem fet amb R un model on la resposta és el percentatge de reducció de la mida del directori, i les variables explicatives són el *tipus de compressor*(tipo) i la *mida inicial* en escala logarítmica (X2 = log(PES_INICIAL_MB)).

1. Què ens diu la següent informació obtinguda de les dades, en relació a l'estudi que es planteja fer?:

```
>tapply(Tb$PES_INICIAL_MB, Tb$tipo, mean)
    rar     tar     zip
1222.80  582.32  524.96
>tapply(Tb$X2, Tb$tipo, mean)

5.739022 5.573800 5.355103
>tapply(Tb$X2, Tb$tipo, median)

5.129899 5.641907 4.905275
```

En primer lloc, veiem que la mida inicial sembla una mica desequilibrada, potser per algun directori molt gran en el grup 'rar' (mitjana 1223, més del doble que a 'tar' i 'zip'). Però quan s'observa la variable transformada es veu que les mitjanes són molt més semblants, i ens interessa que els directoris s'hagin distribuït equitativament entre els tres grups, per tant la transformació logarítmica resulta útil.

2. D'acord amb la descripció de l'enunciat, expliqueu quins són els paràmetres del model i el seu significat (el de cada paràmetre).

$Y = \beta_0 + \beta_1 X2 + \beta_2 T + \varepsilon$

β_0: valor mitjà de la resposta per el tipus de referència (suposem 'rar') i per X2=0

β_1: increment de la resposta per un increment d'1 en X2, i tipus fixe

β_2: en realitat dos paràmetres; canvi mitjà de la resposta comparat amb tipus 'rar', quan el tipus T és 'tar', o quan és 'zip'. Suposant un valor fixe de X2.

V(ε): variància del terme d'error, o variància residual.

Un error habitual és confondre els paràmetres amb les variables: és a dir, donar només les variables que intervenen en el model estadístic.

3. A la dreta podem veure el resultat de `confint()` a partir del model que hem ajustat de les dades disponibles. Interpreteu els valors de la línia "(Intercept)".

```
                2.5 %     97.5 %
(Intercept)  10.178842 39.228647
tipotar      -1.783757 13.332576
tipozip     -14.151894  1.052221
X2           -3.205826  1.501847
```

Amb un 95% de confiança, creiem que la reducció **mitjana** de la mida d'un directori comprimit amb el 'rar' es troba entre 10.2% i 39.2%, concretament un directori de mida X2=0 (equivalent a 1 MB).

Un error habitual és oblidar-se de dir que la interpretació es vàlida quan X2 val zero.

4. A continuació interpreteu els valors de les dues línies següents. Quina conclusió pràctica es pot treure d'aquests resultats?

"tipotar": amb 95% de confiança, creiem que la diferència entre 'rar' i 'tar' es troba entre un descens de la reducció de 1.8% i un increment de 13.3%. No es pot concloure que 'tar' comprimeixi més els directoris que 'rar', l'interval és massa ample.

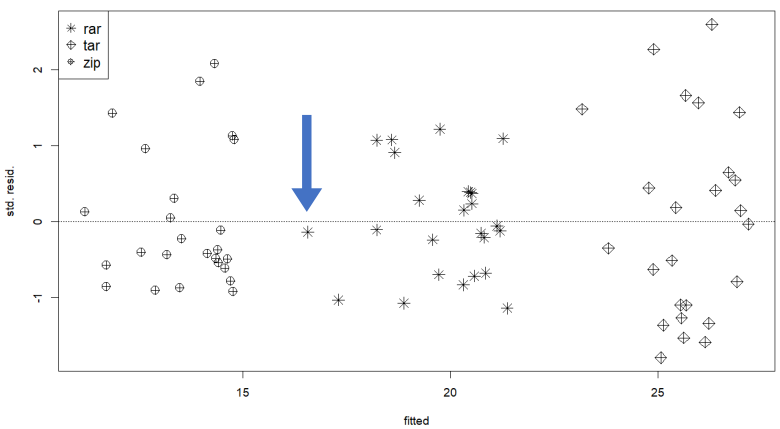
"tipozip": la interpretació seria similar, per a un rang entre -14.2% i 1.1%, és a dir, entre 'rar' i 'zip', no hem pogut trobar evidències de superioritat d'un compressor sobre l'altre. En tot cas, no es pot concloure que els tres compressors són equivalents.

5. A partir del resultat de `confint()`, estimeu puntualment el coeficient associat a X2 i el seu error tipus. Calculeu quin efecte esperaríem si apliquem a un compressor donat un directori que dupliqués la mida, i aporteu una breu valoració del paper d'aquesta variable.

El centre de l'IC ((-3.205826 + 1.501847)/2), **−0.852**, ens dona l'estimació puntual. La meitat de l'amplada dividida per el quantil $t_{0.975,\,71}$, 1.994, ens dona l'error tipus: **1.180**. Com a aproximació, també serveix el quantil d'una normal, 1.96.

Incrementar una unitat X2 o, en termes de mida real, multiplicar la mida per 2.718..., disminueix el percentatge de reducció en 0.85%. Si es duplica la mida del directori és com incrementar X2 en ln(2)=0.6931, per tant la disminució del percentatge de reducció serà 0.852·0.6931 = **0.59**%. Però la incertesa de l'estimació és tan gran que tant podria disminuir com créixer. De fet, el valor neutre (0) és a dins de l'interval.

6. Interpreteu els resultats anteriors conjuntament amb la gràfica de la dreta ('residus estandaritzats' versus 'prediccions'). Quines premises del model emprat es poden verificar amb l'ajut de la gràfica? Justifiqueu la resposta.



Hem dit que no es poden trobar diferències significatives entre 'rar' i 'tar', o entre 'rar' i 'zip', però a la gràfica les prediccions ('fitted') dels tres compressors es distingeixen nítidament en tres grups separats (les petites variacions dintre del grup s'expliquen per diferències de mida, encara que no és una variable rellevant). No hi ha contradicció, perquè la categoria de

referència es troba al mig, i possiblement sí podríem trobar una diferència important entre les dues categories extremes.

La premisa que es pot verificar millor és la *homocedasticitat*, comprovant que la dispersió dels tres grups és semblant (potser 'zip' té més variabilitat i 'rar' menys). La *normalitat* no es pot estudiar satisfactòriament, encara que en els tres casos la simetria és acceptable. La *linealitat* (amb l'única variable numèrica, X2) no és fàcil de verificar, però no s'observa un patró no lineal a cap dels tres blocs.

7. Quants MBs tenia inicialment un directori abans de ser comprimit? Es pot deduir de la figura adjunta? Expliqueu el procediment i apliqueu-ho al punt senyalat amb la fletxa (*fitted* aprox: 16.56, *std. resid* aprox: −0.14).

A partir dels intervals de confiança dels coeficients del model trobem les estimacions puntuals (els centres dels IC) i obtenim un model:

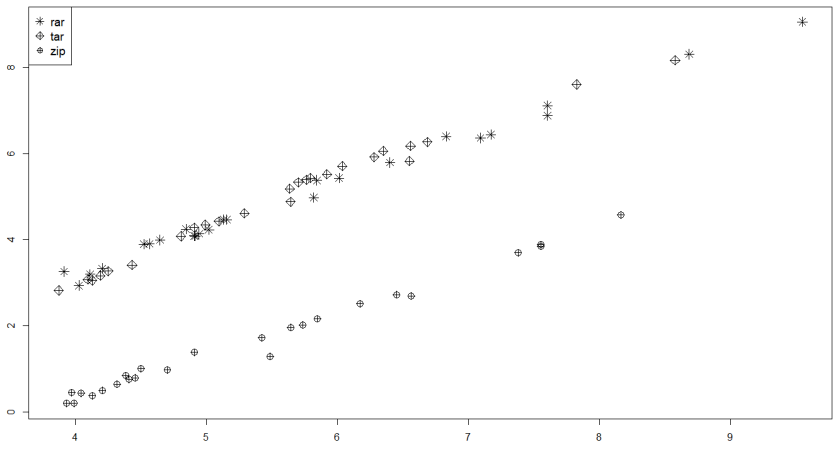
Y = 24.704 + (−0.852) *X2* + 5.774 *es_tar?* + (-6.550) *es_zip?*, on *es_tar?* val 1 si s'utilitza el 'tar' i 0 si no, i *es_zip?*val 1 si s'utilitza el 'zip' i 0 si no. Com el punt de la fletxa utilitza 'rar', el model queda simplement com a Y = 24.704 + (−0.852) *X2*. Igualant a la predicció o valor en l'eix X del punt: 16.56 = 24.704 −0.852 *X2* obtenim *X2* = 9.56, i la mida inicial aproximada del directori serà exp(9.56) = 14167 MB. Noteu que el valor en l'eix Y del punt (−0.14) no és necessari.

8. El resultat de la dretaprové de les mateixes dades i el mateix model, però prenent la categoria 'tar' com a referència. Quin és el canvi més destacat d'aquest resultat, en comparació al model previ, i com s'interpreta?

```
             Estimate Std. Error
(Intercept)    30.478      7.104
Tzip          -12.324      3.794
Trar           -5.774      3.791
X2             ******      *****
```

Obviament, el intercept ha canviat de 24.7 (el centre del primer IC) a 30.5 perquè la referència ara es 'tar' en lloc de 'rar'. Però el canvi més notable és que ara tenim l'estimació del canvi entre 'tar' i 'zip', un canvi de 12.32 punts amb error tipus 3.794 (t = −3.248), que seria estadísticament significatiu. Podem concloure així que almenys una variable del model pot explicar la variabilitat de la resposta.

9. Hem generat el gràfic del costat amb la instrucció `plot(X2, log(T_s))` (i una mica més), on T_s és la variable que recull el temps emprat per les aplicacions al comprimir un directori. Quin model faries servir per considerar aquestes variables i a més el tipus?



El que hem utilitzat ha trobat:

```
Residual standard error: 0.1795 on 71 deg. of fr.
Multiple R-sq.: 0.9932, Adjusted R-sq.: 0.9929
F-statistic: 3458 on 3 and 71 DF, p-value: <
2.2e-16
```

Què expressen els dos indicadors subratllats en aquest cas, en relació a les variables que intervenen al model?

El model seria: $\log(T_s) = \beta_0 + \beta_1 X2 + \beta_2 T + \varepsilon$, i en aquest cas sembla que les variables explicatives sí explicarien molt be la variabilitat de la resposta, el logaritme del temps de càlcul, perquè el coeficient de determinació val 99.32%. La variància del terme d'error s'estimaria per 0.18^2, indicant un valor molt baix. Les prediccions del temps utilitzant la mida del directori i el tipus de compressor serien molt acurades, encara que diferenciar entre tar i rar no aportaria molta informació.