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.

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

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

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.

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.

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

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, $P_{X_1,X_2}(x_1, x_2)$.

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

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

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

NOM:_____ COGNOMS:_____

Problema B

(Justifiqueu les respostes)

Amb l'objectiu de millorar la qualitat del servei, una empresa de repartiment de paqueteria està fent una prova pilot amb 40 vehicles durant un mes. Cada vehicle incorpora un dispositiu GPS nou. Se sap que, a causa de defectes de fabricació, el 4% d'aquests dispositius presenten fallades durant el primer mes d'ús.

1. Quina distribució de probabilitat s'adequa al fenomen "nombre de dispositius GPS que fallen"? Quin és el nombre esperat i la desviació típica del nombre de dispositius defectuosos? **(1,5 punts)**

2. Quina és la probabilitat que falli almenys un dispositiu GPS? **(1,5 punts)**

El temps d'espera (en minuts) perquè el sistema GPS es connecti en arrencar el vehicle és una variable aleatòria. Suposa que aquest temps d'espera segueix una distribució uniforme entre 0,5 i 3 minuts.

3. Quina és la probabilitat que el temps de connexió sigui inferior a 45 segons? **(1 punt)**

En un dels centres logístics de l'empresa, arriben 12 comandes per minut de mitjana.

4. Quin és el model de probabilitat que s'adequa al nombre de comandes que arriben en un minut? Aproximadament, quina és la probabilitat que arribin entre 8 i 15 comandes en un minut? (Fes servir el gràfic corresponent) **(1 punt)**

5. Considerem ara el temps (en minuts) entre l'arribada de dues comandes consecutives al centre logístic. Quin és el model de probabilitat que descriu aquest fenomen? Quin és el temps esperat entre arribades en segons? **(1 punt)**

Durant una campanya especial, la intensitat de comandes es multiplica per 5 respecte d'un dia normal.

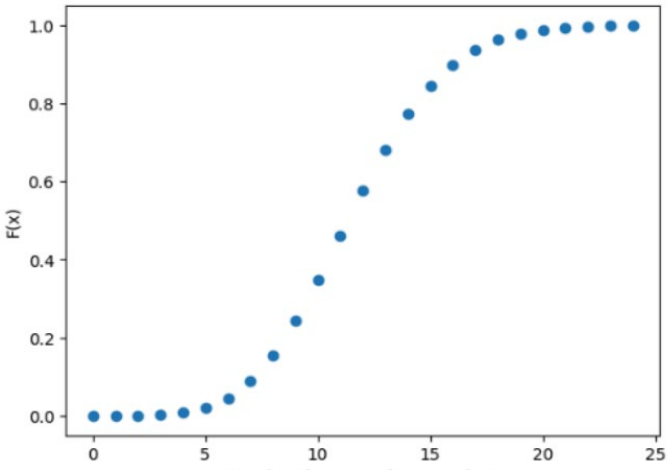
6. Quina és la probabilitat que arribin més de 70 comandes en un minut? Quina és la probabilitat que arribin entre 50 i 90 comandes per minut? (Fes servir el gràfic corresponent) **(2 punts)**

En un dia de campanya, l'empresa disposa de 40 centres logístics, tots amb el mateix comportament pel que fa a les arribades de comandes.

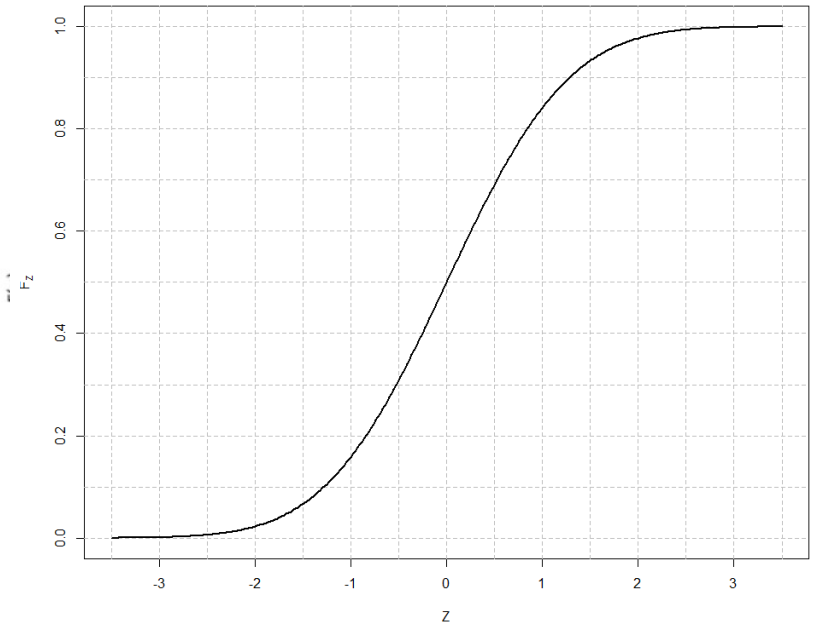
7. Es vol estudiar el nombre total de comandes que arriben als 40 centres logístics durant un minut. Suposant que les arribades són independents i idènticament distribuïdes, indica quina és la distribució de la variable aleatòria que permet estudiar aquest total. **(1 punt)**

8. Aproximadament, quina és la probabilitat que arribin més de 2500 comandes en total als 40 centres en un minut? (Fes servir el gràfic corresponent) **(1 punt)**

Gràfic 1. Funció de distribució acumulada de Poisson ($\lambda = 12$)



Gràfic 2. Funció de distribució acumulada de la normal estàndard Z~N(0,1)



Nom i cognoms:

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*.

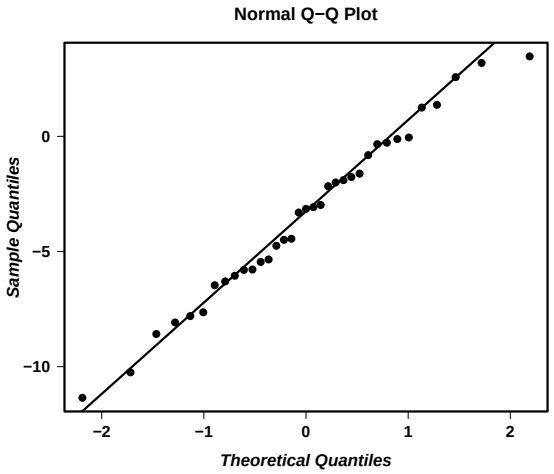
Disposem de les següents dades resumides:

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

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

(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)**

(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)**



(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)**

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

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)**

(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)**

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

`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. *(Justifiqueu les respostes. Totes les preguntes compten igual)*

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 temps de 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
```

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

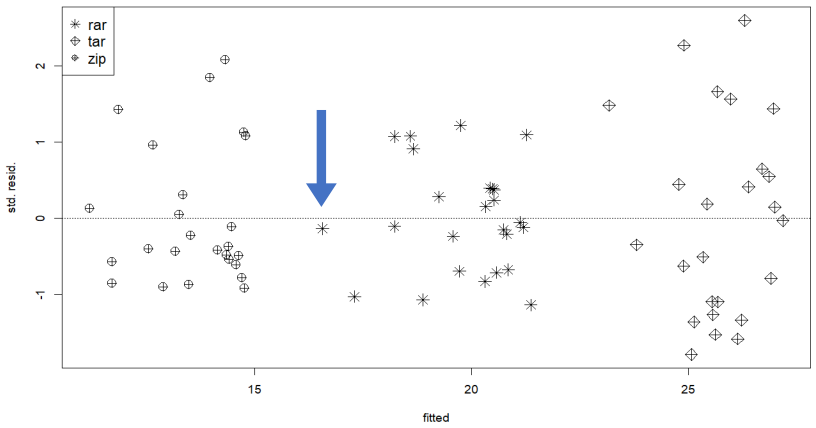
3. A la dreta podem veure el resultat de `confint()` a partir del model que hem ajustat amb les dades disponibles. Interpreteu els valors de la línia "(Intercept)".

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

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

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.

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.

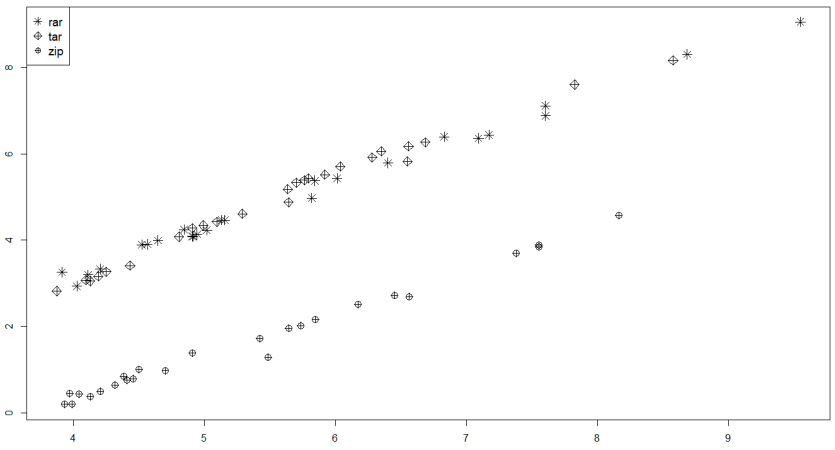


7. Quants MBs tenia inicialment un directori abans de ser comprimit? Es pot deduir de la figura adjunta? Expliqueu el procediment i apliqueu-ho al punt senyalat amb la fletxa (*fitted* aprox: 16.56, *std. resid* aprox: −0.14).

8. El resultat de la dreta prové 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             ******      *****
```

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?