**Facultat de Matemàtiques i Estadística, UPC**

# Estadística

## Problemes i Pràctiques

Marta Pérez, Josep A. Sánchez i Jordi Valero

Primavera 2019-2020

# Índex

# 1 Descriptiva.

## 1.1 Introducció a R i descriptiva univariant.

**Exercici:**

Crea una matriu de dades fictícia amb les següents indicacions:

- La base de dades correspon a una enquesta de 100 estudiants seleccionats aleatòriament d'una certa facultat universitària (població).
- La primera variable és l'identificador de l'enquestat que correspon a un nombre aleatori de 3 xifres.
- La segona variable és el gènere, pot ser "H" o "D" i la proporció de dones en la població enquestada és del 62%
- La tercera variable és l'edat. A la població les edats van de 19 a 24 anys. Les proporcions són 5: 4: 2: 2: 1: 1
- La quarta variable és el pes en kg. La distribució del pes en la població segueix una distribució gamma amb paràmetre de forma 300 i d'escala 1/5.
- La cinquena variable és l'alçada en cm. A la població, l'alçada es distribueix com una Normal amb mitjana 170cm i desviació estàndard 8cm
- La sisena variable és el grau de coneixement d'anglès. Pot ser: "Cap", "Baix", "Mitjà", "Alt" i "Molt Alt". La proporció de cada categoria en la població és 2%, 7%, 40%, 30% i 21% respectivament.
- La setena variable és el nombre de germans. A la població, aquesta variable té una distribució binomial negativa amb paràmetre de mida 5 i probabilitat 0.3

Obté la representació gràfica adient per a les variables gènere, edat, pes, alçada, anglès i germans i verifica que la distribució de la mostra s'assembla a la teòrica de la població:

- Pels histogrames: incorpora la corba de la distribució teòrica
- Per als diagrames de barres: fes també la representació del gràfic amb les proporcions teòriques.
- Tant per a les continues com per les discretes, fes també la comparació de les funcions de distribucions mostral i teòrica per comprovar que les dades tenen la distribució que els hi correspon.

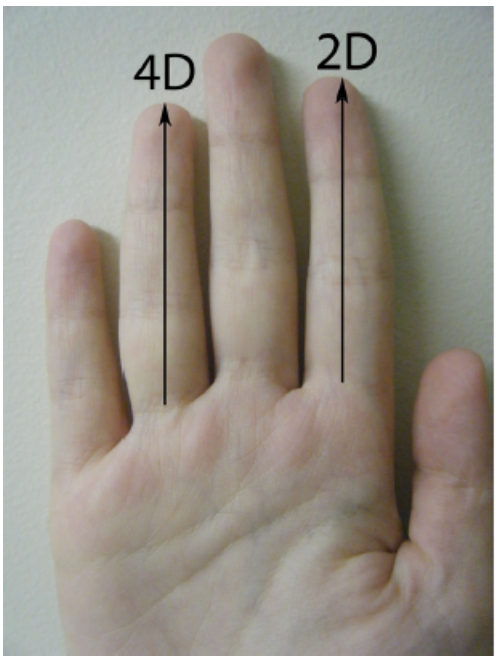
## 1.2 Estadística Descriptiva bivariant

La base de dades "`enquesta.csv`" conté informació d'una enquesta realitzada a 500 persones.

Les variables recollides són:

| Variable | Descripció | Valors |
|---|---|---|
| Edat | Edat de l'individu | |
| Sexe | Sexe de l'individu | M: Home, F:Dona |
| H | Alçada en cm | |
| Pes | Pes en Kg | |
| D2 | Longitud en cm del dit índex | |
| D4 | Longitud en cm del dit anular | |
| MO | Separes, i diposites al seu contenidor, la matèria orgànica? | 1: No, 2: Sí |
| PC | Separes, i diposites al seu contenidor, el paper i cartró? | 1: No, 2: Sí |
| Envàs | Separes, i diposites al seu contenidor, el plàstic i els envasos? | 1: No, 2: Sí |
| Vidre | Separes, i diposites al seu contenidor, el vidre? | 1: No, 2: Sí |
| Aspecte | Quin és l'aspecte que dificulta més aquesta classificació? | E: Falta d'espai per a tantes bosses<br>T: Pèrdua de temps<br>M: Desconeixement de la manera de fer-ho<br>F: Desconeixement dels avantatges a futur |

Les longituds dels dos dits es relacionen amb la presència d'estròg ens i testosterona en el període de gestació, a partir de l'índex de Manning D2/D4. `https://en.wikipedia.org/wiki/Digit_ratio`



L'objectiu és fer un anàlisi descriptiu de la informació continguda en la mostra.

Per llegir les dades, s'ha de tenir en compte que el fitxer té format `csv`, creat amb excel. El format situa els camps separats per punt i coma i les comes es reserven per al delimitador decimal.

```
> dades=read.csv2("Enquesta2.csv")
```

### Exercici:

En aquests exercicis, inclou una breu interpretació dels resultats obtinguts

### Anàlisi descriptiva univariant

**1)** Crea una nova variable amb l'Index de Massa Corporal (BMI) que és el pes en kg dividit per l'alçada al quadrat mesurada en metres

**2)** Crea una nova variable de tipus qualitatiu que reflecteixi si l'individu te un BMI baix (per sota de 19), normal (entre 19 i per sota de 27) i alt (la resta).

**3)** Fes la descriptiva numèrica univariant corresponent a cada variable i inclou la representació gràfica que consideris adient en cada cas.

**4)** Fes un diagrama de caixes (boxplot) per representar l'Index de Massa Corporal. Afegeix dues línies per marcar les categories mitjançant la comanda `abline`.

**5)** Per a les variables de pes i alçada, fes un histograma i superposa la distribució normal amb la mitjana i desviació estàndard de la mostra.

### Anàlisi descriptiva una variable quantitativa amb una o més qualitatives

**1)** Analitza com varia l'alçada en funció del sexe i del grup de BMI

a) Resums numèrics:

(1) Fes una taula de mitjanes de H en les combinacions de sexe i grup de BMI, en la que també hi hagi les mitjanes marginals. (Comanda `tabular` del paquet de R `tables`)

(2) Fes una taula de variàncies de H en les combinacions de sexe i grup de BMI, en la que no hi hagi les variàncies marginals.

b) Resum gràfic:

(1) Fes la gràfica de mitjanes, amb la desviació tipus, sobre les combinacions de sexe i grup de BMI. (comanda `plotMeans` del paquet de R `RcmdrMisc`)

### Anàlisi descriptiva bivariant

**1)** a) Analitza la relació entre pes i alçada i entre el BMI i la edat. Fes servir la instrucció `scatterplot` del package car.

b) Calcula la recta de regressió de Pes respecte H, pel mètode de mínims quadrats, és a dir, minimitzant la funció dels paràmetres $\alpha$ i $\beta$ $SQ(\alpha, \beta) = \sum_{k=1}^{N} (Pes_k - \alpha - \beta \cdot H_k)^2$. Fes servir la instrucció `nlm`, o bé `optim`, per calcular numèricament el mínim. Compara el resultat que has obtingut amb el del `scatterplot`.

**2)** Fes la descriptiva numèrica i gràfica per a l'alçada, pes i rati, segmentant per sexe. Interpreta breument els resultats.

**3)** Considera les variables relacionades amb els hàbits de reciclatge (MO, PC, Envàs, Vidre i Aspecte) i fes una descriptiva segmentada per sexe

# 2 Població i Mostra. Distribucions relacionades amb la Normal.

## 2.1 Teorema de Fisher. Distribucions de mostreig.

**1)** En determinada població les alçades dels homes segueixen una distribució $N(170, 7^2)$ i la de les dones $N(160, 6^2)$. Escollim un home i una dona a l'atzar. Calculeu la probabilitat que l'home sigui més alt que la dona.

**2)** Sigui $\overline{X}$ la mitjana d'una mostra de 16 variables aleatòries. independents i normals (0,1). Determineu c tal que

$$\Pr\left(\left|\overline{X}\right| < c\right) = 0.5 \ .$$

**3)** Siguin $\{X_1, \cdots, X_{18}\}$ v.a. independents amb distribució normal amb esperança $\mu$ i desviació estàndard $\sigma$ desconegudes. Useu la distribució $\chi^2$ per calcular

$$\Pr\left(a < \tfrac{S^2}{\sigma^2} < b\right) \ .$$

Calculeu aquesta probabilitat per a $a = 0.51$ i $b = 1.62$ .

**4)** Suposeu que hem de prendre una mostra d'una distribució normal amb esperança $\mu$ desconeguda i desviació estàndard 2. Calculeu la grandària de la mostra necessària, en cadascuna de les següents situacions, per tal que per a cada possible valor de $\mu$:

a) $E_\mu\left(\mid \overline{X}_n - \mu \mid^2\right) \leq 0.1$ .

b) $E_\mu\left(\mid \overline{X}_n - \mu \mid\right) \leq 0.1$ .

c) $\Pr_\mu\left(\mid \overline{X}_n - \mu \mid \leq 0.1\right) \geq 0.95$ .

**5)** Suposem que $X_1, \ldots, X_{16}$ formen una mostra aleatòria d'una distribució normal d'esperança $\mu$ i variància $\sigma^2$. Calculeu les següents probabilitats:

a) $\Pr\left(\frac{\sigma^2}{2} \leq \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \leq 2\sigma^2\right)$ .

b) $\Pr\left(\frac{\sigma^2}{2} \leq \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \leq 2\sigma^2\right)$ .

**6)** Donada una v.a. $Y$ a $\mathbb{R}$, es diu que els seus resultats "normals" són el conjunt $(a, b)$ amb $\Pr(Y \in (a, b)) = p_0$ on $p_0$ és una probabilitat predeterminada alta, per exemple 95%, 99%... Els conjunts $(-\infty, a]$ i $[b, \infty)$ són els resultats "estranys", generalment es demana que $\Pr(Y \in (-\infty, a]) = \Pr(Y \in [b, \infty))$, també es pot planejar que una de les cues tingui probabilitat 0, és a dir que els resultats normals siguin de la forma $[a, \infty)$ o bé $(-\infty, b]$.

a) Si $Y \sim N(\mu, \sigma^2)$ amb $\mu$ i $\sigma^2$ coneguts, trobeu-ne els resultats normals al 99% on els valors estranys són dues cues equiprobables.

b) Si $Y \sim N(\mu, \sigma^2)$ amb $\mu$ i $\sigma^2$ desconeguts, però tenim una m.a.s. $\{Y_1, ..., Y_{10}\}$ de la que s'ha obtingut $\overline{Y} = m$ i $S^2 = v$ ,trobeu-ne (aproximadament, "interval de predicció") els resultats normals al 99%, on els valors estranys són dues cues equiprobables.

c) En la mateixa situació de l'apartat (2), trobeu els resultats normals al 99% de dues cues iguals: de $\overline{Y}$ suposant $\mu$ conegut, i també de $S_Y^2$ però suposant $\sigma^2$ conegut.

**7)** Suposeu que les v.a. $X_1, ..., X_n$ formen una mostra aleatòria simple d'una distribució contínua en la recta real amb funció de densitat $f$.

- Trobeu l'esperança del nombre d'observacions en la mostra que estaran a dins d'un interval especificat $a \leq x \leq b$.

**8)** *Sigui $X$ una v.a. amb distribució $N(0, 1)$.*

a) *Sigui $A = \{|X| \leq 0.7\}$ i $B = \{|X| \geq 0.7\}$. Quin conjunt té probabilitat més gran?*

b) *Què és més gran:* $\Pr(-0.5 \leq X \leq 0.1)$ *o* $\Pr(1 \leq X \leq 2)$*?*

# 3 Estimació puntual

## 3.1 Mètodes

### 3.1.1 Mètode dels moments

**1)** Suposeu que $X$ segueix una distribució lognormal de paràmetres $\mu$ i $\sigma^2$. Prenem una mostra de grandària $n$ d'aquesta distribució.

Trobeu els estimadors de $\mu$ i $\sigma^2$ pel mètode dels moments.

**2)** Suposeu que $X$ segueix una distribució exponencial d'esperança $\tau$ i taxa de fallida $\lambda$. Prenem una mostra de grandària $n$ d'aquesta distribució. Amb el mètode dels moments:

a) Trobeu l'estimador de $\tau$ . Anomeneu-lo $\hat{\tau}_1$.

b) Trobeu l'estimador de $\lambda$ . Anomeneu-lo $\hat{\lambda}_1$.

**3)** Suposeu que $X$ segueix una distribució de Rayleigh (cas particular de la Weibull quan el paràmetre de forma és 2) de paràmetre $\theta$, $F(x) = 1 - e^{-\frac{1}{2}\frac{x^2}{\theta}}$, amb $x \geq 0$ i $\theta > 0$ . Prenem una mostra de grandària $n$ d'aquesta distribució.

Utilitzant el mètode dels moments:

a) Trobeu l'estimador de $\theta$.

b) Amb la parametrització $F(x) = 1 - e^{-\beta x^2}$, $x \geq 0\, i\, \beta > 0$, trobeu l'estimador de $\beta$.

**4)** Sigui $X_1, \ldots, X_n$ vv.aa. independents i idènticament distribuïdes amb funció de densitat:

$$f(x;\theta) = \frac{x^3}{6\theta^4} e^{-\frac{x}{\theta}} \quad per \quad 0 < x < \infty \quad \text{i} \quad \theta > 0.$$

Obtingueu, mitjançant el mètode dels moments, un estimador per a $\theta$.

### 3.1.2 Màxima versemblança.

**5)** Suposeu que $X$ segueix una distribució lognormal de paràmetres $\mu$ i $\sigma^2$. Prenem una mostra de grandària $n$ d'aquesta distribució.

Trobeu els estimadors de $\mu$ i $\sigma^2$ pel mètode de la màxima versemblança.

**6) Continuació del problema 2)**

Suposeu que $X$ segueix una distribució exponencial d'esperança $\tau$ i taxa de fallida $\lambda$. Prenem una mostra de grandària $n$ d'aquesta distribució. Utilitzant el mètode de màxima versemblança:

a) Trobeu l'estimador de $\tau$ . Anomeneu-lo $\hat{\tau}_2$.

b) Trobeu l'estimador de $\lambda$ . Anomeneu-lo $\hat{\lambda}_2$.

**7)** Suposeu que $X$ segueix una distribució de Rayleigh (cas particular de la Weibull quan el paràmetre de forma és 2) de paràmetre $\theta$, $F(x) = 1 - e^{-\frac{1}{2}\frac{x^2}{\theta}}$, amb $x \geq 0$ i $\theta > 0$ . Prenem una mostra de grandària $n$ d'aquesta distribució.

Utilitzant el mètode de màxima versemblança:

a) Trobeu l'estimador de $\theta$ .

b) Amb la parametrització $F(x) = 1 - e^{-\beta x^2}$, $x \geq 0\, i\, \beta > 0$, trobeu l'estimador de $\beta$ .

**8)** Disposem d'una mostra $X_1, \ldots, X_n$ d'una distribució de Pareto. La distribució de Pareto s'utilitza en ciències econòmiques com a model per a una funció de densitat amb una cua que decreix a poc a poc. Aquesta llei té per densitat ($\theta > 0$, $\tau > 0$):

$$f(x;\theta,\tau) = \begin{cases} \frac{\theta\tau^\theta}{x^{\theta+1}} & x \geq \tau \\ 0 & x < \tau \end{cases}$$

a) Calculeu l'esperança d'una v.a. $X$ que segueix una distribució de Pareto.

b) Suposem que el valor $\tau$ és conegut.

- Trobeu l'estimador $\hat{\theta}_1$ de $\theta$ pel mètode dels moments.
- Trobeu l'estimador $\hat{\theta}_2$ de $\theta$ pel mètode de la màxima versemblança.

**9)** Si les freqüències genètiques estan en equilibri, els genotipus AA, Aa i aa ocorren amb probabilitats $p^2$, $2p(1-p)$ i $(1-p)^2$ respectivament. Plato i *al.* (1964) publicaren les següents dades sobre el tipus d'haptoglobina en una mostra de 190 persones:

| genotip | AA | Aa | aa |
|---|---|---|---|
| freqüència | 10 | 68 | 112 |

Trobeu l'estimador de màxima versemblança de $p$.

**10)** **Propietat**: L'estimador de màxima versemblança no sempre existeix.

Considerem una v.a. $X$ que pot provenir d'una distribució $N(0,1)$ amb probabilitat 0.5 i d'una distribució $N(\mu, \sigma^2)$, amb $\mu$ i $\sigma^2$ desconeguts, amb probabilitat 0.5.

Demostreu que no existeix l'estimador de màxima versemblança.

**11)** Una variable aleatòria $X$ segueix una distribució normal d'esperança $\mu$ i de variància 1. Es fan 20 observacions de $X$ però en comptes d'escriure el valor de les $X$ només s'observa si $X$ és negativa o no.

a) Estimeu $\mu$ per màxima versemblança basant–se tan sols en la informació disponible.

b) Suposant que l'esdeveniment $\{X < 0\}$ ha ocorregut exactament 14 vegades, trobeu el valor de l'estimador de màxima versemblança de $\mu$.

**12)** Suposeu que $X$ segueix una distribució uniforme $[0, \theta]$. Prenem una mostra de grandària $n$ d'aquesta distribució.

a) Trobeu l'estimador de $\theta$ pel mètode dels moments, així com la seva esperança i variància.

b) Trobeu l'estimador de $\theta$ pel mètode de la màxima versemblança.

c) Trobeu la funció de densitat de l'estimador de màxima versemblança, i calculeu la seva esperança i la seva variància. Compareu la variància, el biaix i l'error quadràtic mig amb l'estimador obtingut pel mètode dels moments.

d) Trobeu una modificació de l'estimador de màxima versemblança que el faci no esbiaixat.

**13)** **Continuació problema 8) d'aquest Tema.**

Disposem d'una mostra $X_1, \ldots, X_n$ d'una distribució de Pareto. La distribució de Pareto s'utilitza en ciències econòmiques com a model per a una funció de densitat amb una cua que decreix a poc a poc. Aquesta llei té per densitat ($\theta > 0$, $\tau > 0$):

$$f(x; \theta, \tau) = \begin{cases} \frac{\theta \tau^\theta}{x^{\theta+1}} & x \geq \tau \\ 0 & x < \tau \end{cases}$$

a) Suposem ara que el valor de $\tau$ és desconegut i el de $\theta$ conegut.

Calculeu l'estimador $\hat{\tau}_1$ de màxima versemblança de $\tau$.

b) Suposem que els dos valor de $\tau$ i de $\theta$ són desconeguts.

Calculeu els estimadors de màxima versemblança de $\theta$ i de $\tau$.

**14)** Considerem la variable aleatòria $T$ que mesura el temps de supervivència en mesos d'un pacient després d'un tractament. Suposem que $T$ es distribueix segons un model exponencial traslladat.

El model exponencial traslladat de paràmetres $\lambda$ ($\lambda > 0$) i $G$ ($G > 0$) té per funció de densitat:

$$f(t;\lambda,G) = \begin{cases} \lambda e^{-\lambda(t-G)} & si \quad t \geq G \geq 0 \\ 0 & si \quad t < G \end{cases}$$

$G$ s'interpreta com el temps de garantia o el mínim temps de vida abans del qual no hi ha morts.

Hem fet un estudi amb 11 pacients i llurs temps de supervivència han estat:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 13 | 13 | 13 | 13 | 13 | 14 | 14 | 15 | 15 | 17 |

Suposem que aquests pacients són una mostra d'una llei exponencial traslladada de paràmetres $\lambda$ i $G$.

a) Calculeu la funció de versemblança d'aquestes dades en funció dels paràmetres desconeguts.

b) Suposeu $G$ conegut.

- Estimeu la taxa de fallida $\lambda$ mitjançant el mètode de la màxima versemblança.
- Calculeu el valor de l'estimador de $\lambda$ en funció de $G$ pels pacients de l'estudi.
- Trobeu l'estimador de màxima versemblança de la mediana del temps de supervivència i anomeneu-lo $\hat{m}$. Calculeu la mediana del temps de supervivència dels 11 pacients.
- Determineu la distribució asimptòtica de l'estimador $\hat{m}$. Calculeu la variància asimptòtica de $\hat{m}$.

c) Suposeu $\lambda$ conegut.

- Estimeu $G$ mitjançant el mètode de la màxima versemblança.
- Calculeu el valor de l'estimador de $G$ pels pacients de l'estudi.

d) Suposeu $\lambda$ i $G$ desconeguts.

- Trobeu els estimadors de màxima versemblança per a $\lambda$ i $G$. Calculeu-los pels pacients de l'estudi.
- Feu servir les estimacions obtingudes per estimar la probabilitat de sobreviure 18 mesos després del tractament.

### 3.1.3 Mètode de Bayes

**15)** Suposem que la proporció $\theta$ de components defectuoses d'un gran lot és o bé 0.1 o bé 0.2, i la funció de probabilitat *a priori* de $\theta$ és $\psi(0.1) = 0.7$ i $\psi(0.2) = 0.3$. Es seleccionen 8

components del lot a l'atzar i es troben 2 defectuoses. Determineu la funció de probabilitat *a posteriori* de $\theta$.

**16)** Suposem que la proporció $\theta$ de components defectuoses d'un gran lot és desconeguda, i la distribució *a priori* de $\theta$ és uniforme en l'interval $(0, 1)$. Es seleccionen 8 components del lot a l'atzar i es troben 3 defectuoses. Determineu la llei *a posteriori* de $\theta$.

**17)** Suposem que la proporció $\theta$ de components defectuoses d'un gran lot és desconeguda, i la distribució *a priori* de $\theta$ és Beta($\alpha = 2,\ \beta = 200$). Es seleccionen 100 components del lot a l'atzar i es troben 3 defectuoses.

a) Determineu la llei *a posteriori* de $\theta$.

b) Suposem que després d'observar 3 peces defectuoses la distribució *a posteriori* és una beta de mitjana .0392 i de variància $3.658 \times 10^{-4}$. Determineu la llei *a priori* de $\theta$.

**18)** Les alçades dels individus d'una certa població segueixen una distribució normal amb mitjana desconeguda $\theta$ i amb desviació estàndard 5 cm. Se suposa que la distribució a priori de $\theta$ és una normal de mitjana 170 cm i de desviació estàndard 2.5 cm.

a) La mitjana de 10 persones seleccionades a l'atzar és de 174 cm. Determineu la llei *a posteriori* de $\theta$.

b) Calculeu l'interval d'amplada 2.5 cm que contingui el valor de $\theta$ amb la màxima probabilitat *a priori* i doneu aquesta probabilitat.

c) Calculeu l'interval d'amplada 2.5 cm que contingui el valor de $\theta$ amb la màxima probabilitat *a posteriori* i doneu aquesta probabilitat.

**19)** El temps (en minuts) que es triga en atendre un client en una botiga segueix una llei exponencial de paràmetre $\theta$ desconegut. la distribució *a priori* de $\theta$ és una gamma amb mitjana 0.2 i desviació estàndard igual a 1. Si el temps que s'ha trigat en atendre a 20 clients ha sigut de 3.8 minuts, quina és la distribució *a posteriori* de $\theta$?

## 3.2 Avaluació d'estimadors

### 3.2.1 Error quadràtic mitjà. Biaix i variància.

**1)** Aquest problema fa referència a l'estimació de la variància d'una distribució normal de mitjana desconeguda a partir d'una mostra $X_1, \ldots, X_n$ de v.a. normals. Considerem tres possibles estimadors de la variància poblacional:

$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$

$\sigma_*^2 = \rho \sum_{i=1}^n (X_i - \overline{X})^2$

a) Calculeu el biaix de cadascun d'aquests tres estimadors. Quins d'aquests són no esbiaixats?

b) Calculeu la variància de cadascun d'aquests tres estimadors.

c) Calculeu l'error quadràtic mitjà dels dos primers estimadors.

d) Per a quin valor de $\rho$ l'estimador $\sigma_*^2$ té error quadràtic mitjà més petit?

e) Quin estimador escolliríeu i per què?

**2)** Considerem una mostra de mida $n$, $X_1, \ldots, X_n$ d'una població amb moment d'ordre $k$ finit.

a) Demostreu que $\frac{1}{n}\sum_{i=1}^n X_i^k$ és un estimador sense biaix del moment d'ordre $k$.

b) Determineu el valor de $c$ per tal que $c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2$ sigui un estimador sense biaix de $\sigma^2$.

**3)** Demostrar que si $X$ és v.a. amb esperança $\mu \neq 0$ i $Var(X) > 0$, d'entre tots els estimadors lineals sense biaix $U = a_1X_1 + \ldots + a_nX_n$ de $\mu$, el de variància mínima és la mitjana mostral $\overline{X}$.

**4)** Donada la v.a. $X \sim U[0, b]$, amb $b$ desconegut, considerem els estimadors $V_1 = \max\{X_1, \ldots, X_n\}$, $V_2 = 2\overline{X}$, $V_3 = \frac{n+1}{n}\max\{X_1, \ldots, X_n\}$.

a) Demostreu que $V_1$ és un estimador amb biaix de $b$.

b) Demostreu que $V_2$ és un estimador sense biaix de $b$.

c) Demostreu que $V_3$ és un estimador sense biaix de $b$ millor que $V_2$.

d) Feu una estimació de $b$ usant els tres estimadors, a partir de la mostra:

| | | | | | |
|---|---|---|---|---|---|
| 0.53 | 0.73 | 1.54 | 2.48 | 1.30 | 0.20 |

**5)** Considerem una mostra $X_1, \ldots, X_n$ d'una llei amb densitat

$f_\theta(x) = e^{\theta - x}1_{(\theta,\infty)}(x)$, on $\theta > 0$.

a) Sigui $T_1 = \text{mín}\,(X_1, \ldots, X_n)$. És $T_1$ un estimador sense biaix de $\theta$? En cas que tingui biaix, deduïu-ne un de sense biaix, que designarem per $T_2$, i calculeu l'error quadràtic mitjà.

b) Considerem ara l'estimador $T_3 = \overline{X} - 1$. Té biaix l'estimador de $T_3$? És millor que $T_1$?

**6)** Sigui $X_1, \ldots, X_n$ una mostra aleatòria d'una distribució uniforme en l'interval $(\alpha, \alpha+1)$. Definim

$$\hat{\alpha}_1 = X_1 - \tfrac{1}{2}, \quad \hat{\alpha}_2 = X_{(n)} - \tfrac{n}{n+1}.$$

a) Demostreu que són estimadors no esbiaixats del paràmetre $\alpha$.

b) Quin estimador és millor?

### 3.2.2 Eficiència (Informació de Fisher i cota de Cramer-Rao).

**7)** Sigui $X$ una v.a. normal d'esperança 0 i amb desviació estàndard $\sigma$ ($\sigma > 0$) desconeguda.

a) Trobeu un estimador no esbiaixat per a $\sigma$.

b) Calculeu la variància d'aquest estimador.

c) Calculeu la informació de Fisher $I(\sigma)$ continguda en $X$.

d) Calculeu la informació de Fisher $I(\sigma^2)$ continguda en $X$.

e) Demostreu que la variància per aquest estimador de $\sigma$ és superior a $1/I(\sigma)$ per cada valor $\sigma > 0$.

**8)** Considerem una mostra de mida $n$ d'una llei de Poisson de paràmetre $\lambda$. Demostreu que $\overline{X}$ és un estimador sense biaix i eficient de $\lambda$.

**9)** Sigui $X_1, \ldots, X_n$ una mostra aleatòria d'una distribució de Bernoulli de paràmetre $p$ desconegut:

a) Demostreu que $\bar{X}_n$ és un estimador eficient de $p$.

b) (♠) Per a $n \geq 2$ trobeu un estimador sense biaix de $p^2$. Demostreu que si $n = 1$ aleshores no existeix cap estimador sense biaix de $p^2$.

**10)** Sigui $X_1, \ldots, X_n$ una mostra d'una llei de Poisson de paràmetre $\lambda$. Volem estimar $g(\lambda) = \lambda^2$. Definim

$$T = \frac{1}{n^2}\left(\left(\sum_{i=1}^{n} X_i\right)^2 - \sum_{i=1}^{n} X_i\right).$$

És $T$ un estimador sense biaix de $g(\lambda)$? Calculeu-ne la fita de Cramer-Rao.

**11)** La distribució de Weibull, de paràmetres $(\lambda, \alpha)$, $\lambda > 0$ *i* $\alpha > 0$, té per densitat

$$f(x, \lambda, \alpha) = \frac{\alpha}{\lambda} x^{\alpha-1} e^{\frac{-x^\alpha}{\lambda}} \cdot 1_{(0,\infty)}(x)$$

a) Calculeu la matriu de variàncies asimptòtiques dels estimadors de màxima versemblança de $\lambda$ i $\alpha$. (Nota: No hi ha expressions explícites per aquests estimadors.)

b) Disposem d'una mostra $X_1, \ldots, X_n$ d'aquesta distribució, amb $\alpha$ conegut. Calculeu la informació de Fisher $I(\lambda)$ de la mostra i trobeu un estimador eficient de $\lambda$.

### 3.2.3 Suficiència, Consistència i distribució asimptòtica dels estimadors de màxima versemblança.

**12)** Sigui $X_1, \ldots, X_n$ una mostra d'una distribució geomètrica, i sigui $T = \sum_{i=1}^n X_i$.

Demostreu que $T$ és un estadístic suficient per al paràmetre de la distribució.

**13)** Sigui $X_1, \ldots, X_n$ una mostra d'una distribució de Poisson d'esperança $\lambda$, i sigui $T = \sum_{i=1}^n X_i$.

- Demostreu que la distribució de $X_1, \ldots, X_n$ donat $T$ és independent de $\lambda$, i concloeu que $T$ és suficient per a $\lambda$.
- Demostreu que $X_1$ no és suficient.

**14)** Useu el teorema de factorització per a trobar un estadístic suficient per al paràmetre de la distribució exponencial.

**15)** Disposem d'una mostra $X_1, \ldots, X_n$ d'una llei que té per funció de densitat:

$$f(x;\theta) = \frac{\theta}{(1+x)^{\theta+1}}, \quad \theta > 0 \;\; x \geq 0.$$

Doneu un estadístic suficient per a $\theta$. Raoneu la resposta.

**16)** Disposem d'una mostra $X_1, \ldots, X_n$ d'una llei que té per funció de densitat:

$$f(x;\theta) = (\theta+1)x^\theta, \quad 0 \leq x \leq 1, \; \theta > -1.$$

a) Calculeu l'esperança i la variància de $X$.

b) Trobeu l'estimador $\hat{\theta}_1$ de $\theta$ pel mètode dels moments.

c) Quina llei asimptòtica segueix $\hat{\theta}_1$? Preciseu en particular la seva variància.

**Indicació**: *Feu servir el Teorema Central del Límit i el mètode delta.*

d) Trobeu l'estimador $\hat{\theta}_2$ de $\theta$ pel mètode de la màxima versemblança.

e) Quina llei asimptòtica segueix $\hat{\theta}_2$? Preciseu en particular la seva variància.

f) Calculeu l'eficiència relativa asimptòtica de $\hat{\theta}_1$ respecte $\hat{\theta}_2$ i deduïu quin dels dos estimadors és més eficient.

**17) Continuació del problema 9) del Tema 3.1 .**

Si les freqüències genètiques estan en equilibri, els genotipus AA, Aa i aa ocorren amb probabilitats $p^2$, $2p(1-p)$ i $(1-p)^2$ respectivament. A partir de les dades del problema

| genotip | AA | Aa | aa |
|---|---|---|---|
| freqüència | 10 | 68 | 112 |

Trobeu la variància asimptòtica de l'estimador de màxima versemblança de $\theta$.

**18) Distribució Gamma, exemple de família exponencial.**

Sigui $X$ una variable aleatòria de distribució Gamma amb densitat $f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$

a) Proveu que és una família exponencial, és a dir, que per una parametrització adequada anomenada canònica, $\theta = (\theta_1, \theta_2)$, la funció de densitat es pot escriure de la forma $f(x,\theta) = \frac{h(x)e^{\sum \theta_i T_i(x)}}{C(\theta)}$ on $T(x) = (T_1(x), T_2(x))$ i $C(\theta) = \int h(x)\, e^{\sum \theta_i T_i(x)} dx$. Doneu els paràmetres canònics $\theta$, els estadístics $T(x)$, la funció $h(x)$ i la constant $C(\theta)$.

b) Calculeu $E[T(x)]$, i la matriu de covariàncies $Cov(T(x))$.
Nota: Comproveu que per fer aquests càlculs només necessiteu algunes derivades de $\log(C(\theta))$, per exemple $E[T_i(x)] = \frac{\partial}{\partial \theta_i}\log(C(\theta))$.

c) D'una mostra aleatòria de grandària 250 s'ha obtingut entre altres $\overline{X} = 10.3372$ i $\overline{\log(X)} = 2.0641$. Comproveu que aquests estadístics són suficients i estimeu per màxima versemblança $\mu_x = E[X]$ i $\mu_\ell = E[\log(X)]$. Doneu també $\hat{\alpha}$ i $\hat{\beta}$.

d) Doneu la matriu de covariàncies de l'estimador $(\hat{\mu}_x, \hat{\mu}_\ell)$, en general, i avaluada a $(\mu_x, \mu_\ell) = (\hat{\mu}_x, \hat{\mu}_\ell)$.

e) Calculeu la matriu d'informació de $X$ en els paràmetres canònics.

f) A partir de la matriu de l'apartat anterior, calculeu la fita de Cramer-Rao de l'estimador $(\hat{\mu}_x, \hat{\mu}_\ell)$.

g) Comproveu que $(\hat{\mu}_x, \hat{\mu}_\ell)$ és un estimador UMVUE de $(\mu_x, \mu_\ell)$.

# 4 Intervals de confiança i Test d'hipòtesis

## 4.1 Intervals de Confiança.

### 4.1.1 IC clàssics

**1)** Després d'un tractament contra l'obesitat, els pesos en Kg. de vuit dones eren 58, 50, 60, 65, 64, 62, 56, 57. Suposeu normalitat.

a) Trobeu un interval de confiança al 0.95 per la mitjana sabent que $\sigma = 3$.

b) Trobeu un interval de confiança al 0.9 per la mitjana si $\sigma$ és desconeguda.

**2)** Considerem una mostra de mida $n$, $X_1, \ldots, X_n$ d'una llei normal de mitjana $\mu$ i variància $\sigma^2 = 16$. Trobeu el menor valor de $n$ perquè $[\overline{X} - 1, \overline{X} + 1]$ sigui un interval de confiança per a $\mu$ de nivell 0.95.

**3)** Es fan 5 determinacions de les quantitats d'argent d'un mineral i s'obté 5.2, 4.8, 5.3, 5.7 i 5.0 mg. d'argent. Determineu intervals de confiança amb nivell de 0.95 de la mitjana teòrica (quantitat d'argent) i de la variància teòrica (precisió de l'experiment). Suposeu normalitat.

**4)** Es calcula que la quantitat de nicotina que tenen els cigarrets de determinada marca segueix una distribució $N(30, \sigma^2)$. Agafem deu cigarrets d'aquesta marca a l'atzar i obtenim $\frac{1}{10}\sum_{i=1}^{n}(X_i - 30)^2 = 12.4$ Trobeu un interval de confiança per a $\sigma$ amb nivell de confiança de 0.9.

**5)** Es pren una mostra de mida 3 d'una variable aleatòria normal i dóna $3, 4$ i 5. Determineu un interval de confiança de $\sigma^2$ al 90 per cent.

**6)** Un material s'empaqueta en caixes per dos proveïdors $A$ i $B$. Els pesos de les caixes de $A$ presenten una distribució $N(\mu_1, 0.07^2)$ i les de $B$ una $N(\mu_2, 0.04^2)$. Una mostra de 100 caixes de $A$ dóna una mitjana de 0.99 Kg., i una mostra de 300 caixes de $B$ dóna una mitjana de 1.01 Kg. Determineu un interval de confiança per a $\mu_1 - \mu_2$ al 95 per cent.

### 4.1.2 IC basats en quantitats pivotals.

**7) Continuació problemes 2) i 6) (tema 3.1).**

Suposeu que $X$ segueix una distribució exponencial d'esperança $\tau$ i taxa de fallida $\lambda$. Prenem una mostra de grandària $n$ d'aquesta distribució.

a) Doneu un interval de confiança per a $\tau$.

b) Doneu un interval de confiança per a $\lambda$.

**8)** Siguin $X_1, \ldots, X_m$ i $Y_1, \ldots, Y_n$ dues mostres independents d'una distribució normal amb la mateixa variància desconeguda $\sigma^2$ i amb mitjanes $\mu_1$ i $\mu_2$, respectivament, també desconegudes.

a) Doneu un interval de confiança per a la diferència entre les dues mitjanes de les dues concentracions amb un nivell de confiança del 90%. Indicació: necessitareu la distribució $t_{m+n-2}$ .

b) Hem donat una certa droga A a 8 malalts seleccionats a l'atzar, i després d'una estona hem mesurat la concentració d'aquesta droga en la sang, obtenint:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.23 | 1.42 | 1.41 | 1.62 | 1.55 | 1.51 | 1.60 | 1.76 |

Suposem que administrem una altra droga B a sis malalts diferents i els resultats obtinguts són:

| | | | | | |
|---|---|---|---|---|---|
| 1.76 | 1.41 | 1.87 | 1.49 | 1.67 | 1.81 |

Suposant que totes les observacions provenen d'una distribució normal amb la mateixa variància, Trobeu l'interval de confiança (95%) de la diferència dels valors esperats dels grups, $IC_{95\%}(\mu_B - \mu_A)$.

**9) Continuació del problema 9) (tema 3.1) i 17) (tema 3.2).**

Si les freqüències genètiques estan en equilibri, els genotipus AA, Aa i aa ocorren amb probabilitats $(1-\theta)^2$, $2\theta(1-\theta)$ i $\theta^2$ respectivament. A partir de les dades del problema

| genotip | AA | Aa | aa |
|---|---|---|---|
| freqüència | 10 | 68 | 112 |

i basant-vos en la llei asimptòtica de l'estimador de màxima versemblança de $\theta$, doneu un interval de confiança aproximat al 99% per a $\theta$.

**10)** Sigui $X_1, \ldots, X_n$ una mostra de mida $n$ d'una llei uniforme $U(0, \theta)$, $\theta > 0$.

Designem per $X_{(n)}$ el màxim de $X_1, \ldots, X_n$. Calculeu la distribució de $X_{(n)}/\theta$, i utilitzeu aquest resultat per a construir un interval de confiança de $\theta$ de nivell 0.9.

**11)** Sigui $X_1, \ldots, X_n$ una mostra d'una exponencial de mitjana $\lambda$ desconeguda. Quan ha de valer $n$ per tal que l'interval de confiança al nivell 95% obtingut utilitzant la desigualtat de Txebitxev per a la mitjana poblacional sigui $\left[\frac{20\bar{X}_n}{21}, \frac{20\bar{X}_n}{19}\right]$?

## 4.2 Test d'hipòtesis.

### 4.2.1 Test d'hipòtesis: Tipus d'error, Potència, Regió crítica, p-valor,...

**1)** Suposeu que la variable aleatòria $X$ segueix una distribució binomial de paràmetres $n = 100$ i $p$ desconegut. Considereu el procediment que rebutja $H_0 : p = 0.5$ en favor de $H_A : p \neq 0.5$ si $\mid X - 50 \mid > 10$. Utilitzeu l'aproximació normal a la binomial amb la correcció per continuïtat per a fer tots els càlculs.

a) Calculeu la probabilitat d'un error de tipus I.

b) Feu una gràfica de la funció de potència en funció de $p$.

**2)** Suposeu que es pren una observació $X$ d'una distribució uniforme en l'interval $[\theta - 0.5, \theta + 0.5]$, i suposeu que voleu resoldre la següent prova d'hipòtesi:

$$\begin{cases} H_0 : \theta \leq 3 \\ H_A : \theta \geq 4 \end{cases}$$

Doneu un procediment que tingui una funció de potència que prengui els següents valors: $\pi(\theta) = 0$ per $\theta \leq 3$ i $\pi(\theta) = 1$ per $\theta \geq 4$.

**3)** Sigui $X_1, \ldots, X_n$ una mostra de la distribució uniforme en l'interval $(0, \theta)$. Volem resoldre $H_0 : \theta \geq 2$ versus $H_1 : \theta < 2$. Sigui $Y_n = \max(X_1, \ldots, X_n)$ i considerem aquell procediment que té per regió crítica tots aquells resultats tals que $Y_n \leq 1.5$.

a) Determineu la funció de potència d'aquest procediment.

b) Determineu la talla del procediment.

**4)** Suposeu que $X_1, \ldots, X_n$ és una mostra d'una distribució normal de mitjana $\mu$ desconeguda i de variància 1 i es desitja resoldre la prova d'hipòtesi: $H_0 : 0.1 \leq \mu \leq 0.2$ versus $H_1 : \mu < 0.1$ *o* $\mu > 0.2$. Considerem un procediment $\delta$ que rebutja la hipòtesi nul·la si $\bar{X}_n \leq c_1$ o si $\bar{X}_n \geq c_2$, i sigui $\pi(\mu|\delta)$ la funció de potència de $\delta$. Suposeu que $n = 25$.

a) Determineu els valors de $c_1$ i de $c_2$ per tal que

$$\pi(0.1|\delta) = \pi(0.2|\delta) = 0.07.$$

b) Determineu els valors de $c_1$ i de $c_2$ per tal que

$$\pi(0.1|\delta) = 0.02 \qquad i \quad \pi(0.2|\delta) = 0.05$$

**5)** Suposeu que $X_1, \ldots, X_n$ és una mostra d'una distribució uniforme en l'interval $(0, \theta)$, a on el valor de $\theta$ és desconegut i es desitja realitzar la prova d'hipòtesi: $H_0 : \theta \leq 3$ versus $H_A : \theta > 3$.

a) Demostreu que per a qualsevol nivell de significació $\alpha_0 \in (0, 1)$, existeix un procediment U.M.P. que rebutja $H_0$ si $\max_i \{X_i\} > c$, per alguna constant $c$.

b) Trobeu el valor de $c$ per cada valor possible de $\alpha_0$.

c) Per a una grandària $n$ dibuixeu la funció de potència del procediment anterior.

**6)** La v.a. $X$ segueix una distribució exponencial de paràmetre $\lambda = 1$ o $\lambda = 2$. Volem contrastar amb una mostra de mida 1 les hipòtesis $H_0 : \lambda = 1$ contra $H_1 : \lambda = 2$.

a) Calculeu les probabilitats dels errors de tipus I i de tipus II si la regió d'acceptació és $A_0 = \{x \leq 1\}$.

b) Calculeu les probabilitats dels errors de tipus I i de tipus II si la regió d'acceptació és $A_1 = \{x \geq 0.07\}$.

c) Quina de les dues proves definides per les regions d'acceptació dels apartats 1 i 2 triaríeu?

d) Trobeu el valor de k per tal que la suma dels errors sigui mínima si la regió d'acceptació és $A_k = \{x \geq k\}$.

### 4.2.2 Construcció de Tests

**7)** Donada una v.a. $X$ amb funció de densitat

$$f(x;\theta) = \begin{cases} \frac{1}{6\theta^4} x^3 e^{-\frac{x}{\theta}} & si \quad x > 0 \\ 0 & altrament \end{cases}$$

i el test:

$$\begin{cases} H_0 : \theta = 2 \\ H_1 : \theta = 3 \end{cases}$$

basant–se en una mostra $X_1, \ldots, X_n$ d'aquesta població.

a) Determineu un procediment $\delta$ que minimitzi $a\alpha(\delta) + b\beta(\delta)$, on $\alpha(\delta)$ i $\beta(\delta)$ són les probabilitats d'un error de tipus I i de tipus II, respectivament, si s'utilitza el procediment $\delta$.

b) Calculeu el valor de $\alpha(\delta)$ per $a = 11$ , $b = 2$ , $n = 3$.

c) Es desitja ara realitzar la següent prova d'hipòtesis:

$$\begin{cases} H_0 : \theta \geq 2 \\ H_1 : \theta < 2 \end{cases}$$

Determineu un procediment UMP amb nivell de significació $\alpha_0 = 0.05$ per $n = 3$.

**8)** Sigui $X_1, \ldots, X_n$ una mostra de la distribució de Poisson.

a) Utilitzeu el mètode de la raó de versemblança per a contrastar $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda = \lambda_1$, a on $\lambda_1 > \lambda_0$.

**Indicació:** *La suma de v.a. independents de Poisson segueix una distribució de Poisson.*

b) Feu els càlculs per $n = 20$, $\lambda_0 = 0.25$ i $\lambda_1 = 0.5$.

c) Demostreu que el procediment trobat a l'apartat a) és uniformement més potent per a contrastar $H_0 : \lambda = \lambda_0$, versus $H_A : \lambda > \lambda_0$.

**9)** a) A partir d'una mostra de mida $n = 10$ d'una distribució de Poisson de paràmetre $\lambda$ desconegut, es vol contrastar

$$\begin{cases} H_0 : \lambda = 0.3 \\ H_1 : \lambda = 0.4 \end{cases}$$

Construïu un test amb nivell $\alpha = 0.05$. Calculeu la potència d'aquest test. És UMP per contrastar

$$\begin{cases} H_0 : \lambda = 0.3 \\ H_1 : \lambda > 0.3 \end{cases} \quad ?$$

b) En una mostra d'una distribució de Poisson s'ha observat:

| dades | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| freqüències | 38 | 36 | 17 | 8 | 1 |

Contrasteu:

$$\begin{cases} H_0 : \lambda = 0.3 \\ H_1 : \lambda > 0.3 \end{cases}$$

**10)** A partir d'una observació d'una v.a. $X$ volem fer un test per a contrastar

$$\begin{cases} H_0 : X \sim N\left(0, \frac{1}{2}\right) \\ H_1 : X \sim N\left(0, 1\right) \end{cases}$$

amb nivell $\alpha = 0.01$. Quina potència té aquest test?

**11)** Amb una observació d'una v.a. $X$ volem contrastar les hipòtesis

$$\begin{cases} H_0 : X \quad \textit{té densitat} \quad f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \\ H_1 : X \quad \textit{té densitat} \quad f_1(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \end{cases}$$

Estudieu pels diferents valors de $\alpha \in (0, 1)$ quina és la regió d'acceptació d'un test UMP.

**12)** Suposeu que $X_1, \ldots, X_n$ és una mostra d'una distribució de Poisson de paràmetre $\lambda$ desconegut ($\lambda > 0$).

a) Demostreu que la funció de probabilitat conjunta de $(X_1, \ldots, X_n)$ té una raó de versemblança monòtona en l'estadístic $\sum_{i=1}^{n} X_i$.

b) Demostreu que per $n = 10$ existeix un procediment U.M.P per la prova d'hipòtesi: $H_0 : \lambda \leq 1$ versus $H_A : \lambda > 1$ amb nivell de significació $\alpha_0 = 0.0143$.

c) Demostreu que per $n = 10$ existeix un procediment U.M.P per la prova d'hipòtesi: $H_0 : \lambda \geq 1$ versus $H_A : \lambda < 1$ amb nivell de significació $\alpha_0$ per alguns $\alpha_0$ tals que $0 < \alpha_0 < 0.03$.

**13)** Sigui $X_1, \ldots, X_m$ una mostra d'una distribució exponencial amb taxa de fallida $\theta$. Feu un test de la raó de versemblança per a contrastar:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

**Indicació:** *La regió de rebuig és de la forma* $\left\{\overline{X}e^{-\theta_0 \overline{X}} \leq c\right\}$, *a on* $c$ *és una constant donada pel nivell de significació.*

**14)** Sigui $X_1, \ldots, X_n$ una mostra d'una distribució que té per funció de densitat:

$$f(x;\theta) = \theta x^{\theta-1} \quad 0 < x < 1$$

a on el paràmetre $\theta > 0$ és desconegut. Es desitja saber si es pot concloure que $\theta > 1$ basant-nos en una mostra de grandària 8.

a) Plantegeu formalment el problema.

b) Demostreu que la funció de densitat conjunta de la mostra té una raó de versemblança monòtona en l'estadístic $T$. Preciseu l'expressió d'aquest estadístic.

c) Demostreu que existeix una prova que és uniformement més potent (UMP) i determineu genèricament la regió de rebuig d'aquesta prova.

d) Es desitja concretar la regió de rebuig per un nivell de significació $\alpha_0 = 0.05$.

**Indicació:** *Sigui* $X$ *una v.a. distribuïda segons una llei gamma de paràmetres* $(n, t)$. *Cal utilitzar la següent relació:*

$$\Pr(X \leq a) = \Pr(Y \geq n)$$

*on* $Y$ *és una v.a. que es distribueix segons la llei d'una Poisson de paràmetre* $a \cdot t$, *i utilitzar les taules de la Poisson.*

### 4.2.3 Tests clàssics

**15)** Sigui $X_1, \ldots, X_n$ una mostra d'una població normal de mitjana $\mu$ desconeguda i de variància $\sigma^2$ desconeguda. Desenvolupeu la prova de la $t$ de Student per a contrastar

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

mitjançant el procediment de la raó de versemblança. Discutiu quines són les propietats de la prova $t$ de Student.

**16) Continuació del problema 8) del Tema 4.1 .**
Siguin $X_1, \ldots, X_m$ i $Y_1, \ldots, Y_n$ dues mostres independents d'una distribució normal amb la mateixa variància desconeguda $\sigma^2$ i amb mitjanes $\mu_1$ i $\mu_2$, respectivament, també desconegudes.

a) Donat un valor constant $\lambda$ $(-\infty < \lambda < \infty)$, construïu una prova de la $t$ amb $m+n-2$ graus de llibertat per a: $H_0 : \mu_1 - \mu_2 = \lambda$ versus $H_A : \mu_1 - \mu_2 \neq \lambda$.

b) Hem donat una certa droga A a 8 malalts seleccionats a l'atzar, i després d'una estona hem mesurat la concentració d'aquesta droga en la sang, obtenint:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.23 | 1.42 | 1.41 | 1.62 | 1.55 | 1.51 | 1.60 | 1.76 |

Suposem que administrem una altra droga B a sis malalts diferents i els resultats obtinguts són:

| | | | | | |
|---|---|---|---|---|---|
| 1.76 | 1.41 | 1.87 | 1.49 | 1.67 | 1.81 |

Suposant que totes les observacions provenen d'una distribució normal amb la mateixa variància, feu una prova d'hipòtesi per a saber si podem concloure que la mitjana de la concentració de la droga B és més gran que la de la droga A.

**17)** Considerem dues poblacions normals amb mitjanes $\mu_1$ i $\mu_2$ i variàncies $\sigma_1^2$ i $\sigma_2^2$ desconegudes i suposem que volem fer la següent prova d'hipòtesi: $H_0 : \sigma_1^2 \leq \sigma_2^2$ versus $H_A : \sigma_1^2 > \sigma_2^2$. Suposem que prenem una mostra de grandària 16 de la primera població donant com a resultats $\sum_{i=1}^{16} X_i = 84$ i $\sum_{i=1}^{16} X_i^2 = 563$, i una mostra de grandària 10 de la segona població tal que $\sum_{i=1}^{10} Y_i = 18$ i $\sum_{i=1}^{10} Y_i^2 = 72$

a) Determineu els estimadors de màxima versemblança de $\sigma_1^2$ i de $\sigma_2^2$.

b) Feu una prova de la F amb $\alpha = 0.05$ i concloeu si $H_0$ pot o no rebutjar-se.

c) Suposem ara que volem contrastar $H_0 : \sigma_1^2 \leq 3\sigma_2^2$ versus $H_A : \sigma_1^2 > 3\sigma_2^2$. Descriure com dur a terme un test F per aquestes hipòtesis.

### 4.2.4 Bondat d'ajust

**18)** La distribució de Poisson es pot utilitzar com a model de tràfic quan aquest és lleuger, ja que es basaria en el fet que si la taxa d'arribada és aproximadament constant i el tràfic no és gaire dens (és a dir, els cotxes poden circular independentment un dels altres), la distribució del nombre de cotxes en un interval de temps donat, o en una àrea donada, és aproximadament Poisson. A continuació donem una taula que mostra el nombre de girs a la dreta durant 300 intervals de 3 minuts a un encreuament donat. Aquests 300 intervals estan distribuïts durant vàries hores del dia i varis dies de la setmana. Ajusteu una distribució de Poisson i feu una prova d'ajustament usant l'estadístic $\chi^2$ de Pearson.

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| freqüència | 14 | 30 | 36 | 68 | 43 | 43 | 30 | 14 | 10 | 6 | 4 | 1 | 1 | 0 |

**19)** En un estudi ecològic sobre el comportament alimentari dels ocells, es va comptar el nombre de salts ($n$) entre vols per a uns quants ocells. A partir de les dades següents, ajusteu la distribució geomètrica, trobeu un interval de confiança per a p i feu una prova d'ajustament.

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| freqüència | 48 | 31 | 20 | 9 | 6 | 5 | 4 | 2 | 1 | 1 | 2 | 1 |

**20)** Considereu una prova d'ajustament per a una distribució multinomial amb dues classes. Denotem el nombre d'observacions a cada classe per $X_1$ i $X_2$ i les probabilitats de cada classe per $p_1$ i $p_2$. L'estadístic $\chi^2$ de Pearson ve donat per

$$Q = \sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i}.$$

Demostreu que

$$Q = \frac{(X_1 - np_1)^2}{np_1(1-p_1)}$$

i que sota $H_0 : p_1 = p_1^0$, $Q \sim \chi_1^2$.

**21)** Al 1965, un diari va publicar una història sobre un estudiant que afirmava que havia llençat una moneda 17950 vegades i que havia obtingut 9207 cares i 8743 creus.

a) Representa aquest resultat una discrepància important sobre la hipòtesi $H_0 : p = 0.5$?

b) Un estadístic es va posar en contacte amb ell i li va preguntar que com havia fet l'experiment. L'estudiant per estalviar-se temps havia fet grups de 5 monedes i havia registrat els següents resultats:

| nombre de cares | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| freqüència | 100 | 524 | 1080 | 1126 | 655 | 105 |

Són aquestes dades consistents amb la hipòtesi que les monedes eren equilibrades?

c) Són aquestes dades consistents amb la hipòtesi que les 5 monedes tenien la mateixa probabilitat de cares encara que aquesta no fos necessàriament 0.5?

**22)** La següent taula dóna 50 valors. Feu servir la prova de Kolmogorov-Smirnov per a contrastar que:

a) Les 50 dades provenen d'una distribució normal de mitjana 26 i de variància 4.

b) Les 50 dades provenen d'una distribució normal de mitjana 24 i de variància 4.

| | | | | |
|---|---|---|---|---|
| 25.088 | 26.615 | 25.468 | 27.453 | 23.845 |
| 25.996 | 26.516 | 28.240 | 25.980 | 30.432 |
| 26.560 | 25.844 | 26.964 | 23.382 | 25.282 |
| 24.432 | 23.593 | 24.644 | 26.849 | 26.801 |
| 26.303 | 23.016 | 27.378 | 25.351 | 23.601 |
| 24.317 | 29.778 | 29.585 | 22.147 | 28.352 |
| 29.263 | 27.924 | 21.579 | 25.320 | 28.129 |
| 28.478 | 23.896 | 26.020 | 23.750 | 24.904 |
| 24.078 | 27.228 | 27.433 | 23.341 | 28.923 |
| 24.466 | 25.153 | 25.893 | 26.796 | 24.743 |

*Mostra 1, Dades exercici 22*

**23)** En un estudi antropològic s'estudia la distribució del grup sanguini entre tres races humanes. Es van triar una m.a.s d'individus de cada raça i es va comptar la freqüència en què es trobava cada grup. El resum de la mostra ve donada en la següent taula de contingència:

| | 0 | A | B | AB |
|---|---|---|---|---|
| Caucasians | 45 | 31 | 15 | 9 |
| Pigmeus | 23 | 17 | 20 | 20 |
| Esquimals | 22 | 17 | 25 | 16 |

a) Trobeu la taula d'efectius esperats suposant que no hi ha diferències per les distribucions de grups sanguinis entre les diferents races.

b) Resoleu el test associat a la taula de contingència, per decidir si hi ha diferències entre les races.

c) De quin tipus de test es tracta, d'un test d'independència o d'homogeneïtat? Justifiqueu la resposta.

**24)** La següent taula recull el nombre d'empreses classificades d'acord a la grandària (nombre d'empleats) i la seva situació geogràfica en una certa zona industrialitzada molt extensa.

| | Petita | Mitjana | Gran | Molt Gran |
|---|---|---|---|---|
| Nord | 20 | 10 | 17 | 9 |
| Centre | 30 | 15 | 23 | 10 |
| Sud | 55 | 37 | 25 | 15 |

Hi ha diferències en la distribució del tipus d'indústria en funció de la zona que considerem?

**25)** Els estudiants d'un institut responen a un qüestionari on indiquen el grau de dedicació a l'estudi i els tipus d'estudis dels seus pares. Es vol determinar si existeix relació entre ambdues característiques. La mostra ve resumida en la següent taula:

| | Mai | 4 hores/set | 6 hores/set | 8 hores/set |
|---|---|---|---|---|
| Sense estudis | 50 | 36 | 28 | 19 |
| Primaris | 35 | 18 | 23 | 12 |
| Secundaris | 45 | 47 | 35 | 25 |
| Superiors | 15 | 33 | 45 | 26 |

# 5 Models lineals.

## Exercicis de pràctiques

### 5.1 Regressió lineal simple i múltiple.

**1)** En un estudi per estudiar el nivell de colesterol en sang, en persones en edat de creixement (9-20 anys), s'ha plantejat el model:

$$C_i = \beta_0 + \beta_1 P_i + \beta_2 H_i + \beta_3 E_i + \varepsilon_i \quad \text{amb}\, \varepsilon_i \sim N\left(0, \sigma^2\right)$$

on $C$, $P$, $H$ i $E$ són: el nivell de colesterol, el pes, l'alçada i l'edat, respectivament.
Les dades experimentals són independents i estan en el fitxer "`col.csv`".
Escrivint aquest model de la forma matricial $y = X\beta + \varepsilon$, ($dim\,(X) = (n, p)$ i té rang màxim), contesteu els apartats següents donant les expressions matricials que heu utilitzat i els resultats numèrics d'aquest model. A més comproveu que amb el programa estadístic `R`, obteniu els mateixos resultats. Quan es necessiti $\alpha = 0.05$.

a) Trobeu la funció de regressió $C = \hat{\beta}_0 + \hat{\beta}_1 P + \hat{\beta}_2 H + \hat{\beta}_3 E$.

b) Trobeu els *"valors predits"*, $\hat{y} = \widehat{E}\,[y|\beta]$, i també els "*valors residuals*", $\hat{r} = y - \hat{y}$. Amb ells feu la gràfica de *"residuals"* en front de *"predits"* i dieu si us sembla que el model resumeix bé el comportament del colesterol.

c) Contrasteu el test *"omnibus"* $\left.\begin{array}{l} H_0 : (\beta_1, \beta_2, \beta_3) = (0, 0, 0) \\ H_1 : (\beta_1, \beta_2, \beta_3) \neq (0, 0, 0) \end{array}\right\}$ per saber si el model que hem plantejat millora el model nul $C_i = \beta_0 + \varepsilon_i$ en el que el colesterol no depèn del pes ni de l'alçada ni de l'edat.
Nota: Haureu de calcular $GLE$, $GLM$, $SQE$, $SQM$, $MQE$, $MQM$ i $F_{test}$.

- Per a cadascuna de les tres variables independents, dieu si, en variar-la però mantenint constant les altres dues, afecta al colesterol i si ho fa de forma positiva o negativa. És a dir, heu de contrastar els tests $\left.\begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array}\right\}$, $\left.\begin{array}{l} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{array}\right\}$ i $\left.\begin{array}{l} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{array}\right\}$.
Comproveu que els test coincideixen amb el que dona el `summary(model)`, que coincideix amb el que dona la comanda `Anova(model)`, però no coincideixen amb els de la comanda `anova(model)`, que a més canvia si canviem l'ordre de les variables en

la fórmula del model

*Notes:*

Necessitareu la distribució de $\hat{\sigma}^2 = MQE$ que ja teniu calculada, la distribució de $\hat{\beta}$.

La comanda `Anova(model)` és de la llibreria **"car"** i calcula l'anova amb les sumes de quadrats tipus II, també se li podria demanar les tipus III. La comanda `anova(model)` calcula l'anova amb les sumes de quadrats tipus I (seqüencial) que depenen de l'ordre de les variables en la fórmula.

d) Torneu a fer l'apartat anterior de forma alternativa, calculant els intervals de confiança $IC_{1-\alpha}(\beta_1)$, $IC_{1-\alpha}(\beta_2)$ i $IC_{1-\alpha}(\beta_3)$.

e) Per una persona amb $P = 65$, $E = 15$ i $H = 150$ calculeu-ne:

- Regió de "*predicció del 95%*" del colesterol.
- L'interval de confiança $IC_{1-\alpha}(E[C|P = 65, E = 15, H = 150])$.
- Contrasteu el test $H_0 : E[C|P = 65, E = 15, H = 150] = 190$ vs $H_1 : E[C|P = 65, E = 15, H = 150] \neq 190$.

Nota: Necessitareu la distribució de $\widehat{E}[C|P = 65, E = 15, H = 150] = (1, 65, 15, 150)\,\hat{\beta}$ .

**2)** En el fitxer **"`reg8.csv`"** hi teniu les dades de $X$ i $Y$, en 8 situacions diferents indicades per la variable REG, per a calcular les rectes de regressió de Y respecte $X$.
Nota: Teniu les mateixes dades en el fitxer `reg8h.csv`, però les 8 situacions estan en columnes diferents.

a) Calculeu les 8 rectes de regressió i observeu que totes donen les mateixes funcions de regressió i el mateix resultat del test.

b) Per a cada un dels casos, feu les gràfiques:

- "residuals" front "predits.
- valors observats de $Y$ i $X$, amb la recta de regressió.

Observeu que de les 8 situacions només n'hi ha una en que el model sigui del tot correcte. Trobeu els motius pels que en els demés casos el model no és correcte, o té algun problema.

**3)** En un laboratori de química és molt corrent utilitzar potenciòmetres com a pH-metres, però si canviem els elèctrodes també es poden utilitzar per a determinar altres ions o molècules.

Així es pot connectar l'anomenat elèctrode selectiu d'amoníac que té una membrana especial que, com el seu nom indica, és sensible a l'amoníac gas. En aquest cas les lectures que dona el potenciòmetre es relacionen amb la concentració de nitrogen amoniacal mitjançant uns patrons ( solucions de concentració de $N$. amoniacal coneguda ) i la funció:

$$lectura = \beta_0 + \beta_1 \cdot \log(concentració)$$

En una experiència realitzada, els patrons, de concentracions conegudes han donat :

| *Concentració de NNH3 en ppm* | *2* | *5* | *10* | *20* | *50* | *100* |
|---|---|---|---|---|---|---|
| *lectura de l'elèctrode en mV* | *85* | *107* | *124* | *140* | *161* | *179* |

i un extracte aquós d'un compost ha donat una lectura de : $167mV$.

L'objectiu és estimar la concentració de N. amoniacal que té l'extracte, pel que necessitarem fer la regressió de les lectures (mV) dels patrons, en funció de les seves concentracions (ppm), segons la funció indicada. Quan es necessiti utilitzeu $\alpha = 0.05$.

a) Digueu quin és el model lineal d'aquesta regressió, les condicions que s'ha de suposar que es compleixen i plantegeu les hipòtesis del test ANOVA.

b) Calculeu la funció de regressió i contrasteu el test ANOVA.

c) Estudieu els punts estranys i els punts influents. Digueu si el model que hem escollit és adequat o no, i si cal revisar algun dels resultats obtinguts.

d) Digueu quina és la concentració que segons la regressió dona la mateixa lectura que l'extracte (167 mV).

e) És possible que la concentració de l'extracte sigui 60 ppm?

f) Digueu quina concentració pot tenir l'extracte, es a dir, trobeu totes les concentracions que poden donar la lectura de l'extracte (167 mV) amb probabilitat 0,95.

**4)** Per estudiar com evoluciona la quantitat de llet diària produïda per una vaca en funció dels dies que han passat des del part, s'han obtingut les dades del fitxer "`pllet.csv`", en el que la primera columna són els dies i la segona la producció en l/dia. Quan es necessiti utilitzeu $\alpha = 0.05$.

a) Amb el model de la recta $E[prod] = \beta_0 + \beta_1 dies$ :

(1) Trobeu la funció de regressió i

- Plantegeu i contrasteu el test ANOVA.
- Estimeu la variància de l'Error.
- Calculeu el coeficient de determinació ajustat.

(2) Dibuixeu:

- La funció de regressió amb els punts.
- Els residuals front els valors predits.
- La banda de predicció$_{(95\%)}$ amb la funció de regressió i els punts observats.

(3) Justifiqueu si us sembla que el model és adequat o no.

b) Amb el model de la paràbola $E[prod] = \beta_0 + \beta_1 dies + \beta_2 dies^2$ :

(1) Trobeu la funció de regressió i

- Plantegeu i contrasteu el test ANOVA.

- Estimeu la variància de l'Error.
- Calculeu el coeficient de determinació ajustat.

(2) Dibuixeu:

- La funció de regressió amb els punts.
- Els residuals front els valors predits.
- La banda de predicció$_{(95\%)}$ amb la funció de regressió i els punts observats.

(3) Justifiqueu si us sembla que el model és adequat o no.

c) Amb el model de la funció gamma transformant amb el logaritme $E\left[\log\left(prod\right)\right] = \beta_0 + \beta_1 dies + \beta_2 \log\left(dies\right)$ :

(1) Trobeu la funció de regressió i

- Plantegeu i contrasteu el test ANOVA.
- Estimeu la variància de l'Error.
- Calculeu el coeficient de determinació ajustat.
- Plantegeu i contrasteu el test per contestar les preguntes :
  - · ajusta millor la funció gamma que la funció exponencial?
  - · ajusta millor la funció gamma que la funció potencial?

(2) Dibuixeu:

- La funció de regressió amb els punts (amb transformació i sense).
- Els residuals front els valors predits.
- La banda de predicció amb la funció de regressió i amb els punts. ( log i sense)
- Detecteu punts estranys i punts influents.

(3) Justifiqueu si us sembla que:

- el model és adequat o no.
- el model és més adequat, o no, que el de la funció exponencial.
- el model és més adequat, o no, que el de la funció potencial.

d) Entre tots els models d'aquest exercici, quin us sembla que és més adequat? per què?

**5)** En la mateixa situació que l'exercici **1)**.

És ben conegut que l'excés de pes es un dels factors que afecta negativament al nivell de colesterol de les persones. En un estudi amb nens i joves de 9 a 20 anys, s'ha obtingut el seu nivell de colesterol, $C$, pes, $P$, alçada, $H$, i edat, $E$. Els resultats són al fitxer "`col.csv`" i $\alpha = 0.05$.

a) Plantegeu i calculeu la recta de regressió del colesterol en funció del pes.

b) Dibuixeu la gràfica de la recta de regressió amb les bandes de confiança i de predicció.

c) Utilitzeu les gràfiques adequades pel diagnòstic de:

- Tendències i homogeneïtat de variàncies.
- Valores estranys.
- Valores influents.

d) Interpreteu els resultats, trobeu alguna contradicció? justifiqueu les conclusions. Indicació: us pot ajudar fer el diagrama de dispersió de $C$ vs $P$, afegint les rectes de regressió agrupant per edats, la comanda és:

```
scatterplot(C~P|EDAT, reg.line=lm, smooth=F, spread=F, boxplots=F,
    span=0.5, data=dades)
```

**6)** Com en l'exercici **1)**.

En un estudi per estudiar el nivell de colesterol en sang, en persones en edat de creixement (9-20 anys), s'ha plantejat el model:

$$C_i = \beta_0 + \beta_1 P_i + \beta_2 H_i + \beta_3 E_i + \varepsilon_i \quad \text{amb}\, \varepsilon_i \sim N\left(0, \sigma^2\right)$$

on $C$, $P$, $H$ i $E$ són: el nivell de colesterol, el pes, l'alçada i l'edat, respectivament.

Les dades experimentals són independents i estan en el fitxer "`col.csv`". Quan es necessiti utilitzeu $\alpha = 0.05$.

a) Comproveu, amb notació matricial, que si canviem les dades experimentals per combinacions lineals d'elles, $X_C = X \cdot C$ on $C$ és una matriu $(p, p)$ invertible, aleshores, en ajustar el model $Y = X_C\beta_C + \varepsilon$ s'obté que:

(1) Els valors predits, els residuals, $\hat{\sigma}^2$, $R^2$ i tots els elements del test "omnibus", donen exactament el mateix que sense transformar les dades.

(2) $\hat{\beta}_C = C^{-1}\hat{\beta}$, pel que també canvia la seva distribució, la col·linealitat, els tests dels paràmetres i els intervals de confiança dels paràmetres.

b) És ben conegut que l'excés de pes es un dels factors que afecta negativament al nivell de colesterol de les persones, però l'excés de pes no és exactament el pes. Suposant que $P_0 = -10+0.5H$ és un patró del pes respecta a l'alçada en les persones d'edat en el rang de la nostra experiència, definim l'excés de pes $EP = P-P_0 = P+10-0.5H$. Ajusteu la regressió de $C$ en funció de $EP$, $H$ i $E$ i comproveu que:

(1) Els valors predits, els residuals, $\hat{\sigma}^2$, $R^2$ i tots els elements del test "omnibus", donen exactament el mateix que sense transformar les dades.

(2) El *VIF* (col·linealitat) a canviat.

(3) Els resultats dels tests $\left.\begin{array}{l} H_0 : \beta_{C2} = 0 \\ H_1 : \beta_{C2} \neq 0 \end{array}\right\}$, $\left.\begin{array}{l} H_0 : \beta_{C3} = 0 \\ H_1 : \beta_{C3} \neq 0 \end{array}\right\}$ i $\left.\begin{array}{l} H_0 : \beta_{C4} = 0 \\ H_1 : \beta_{C4} \neq 0 \end{array}\right\}$ han canviat.

(4) Reinterpreteu com cadascuna de les variables independents afecta al colesterol, mantenint constants les demés variables independents.

(5) Tindria sentit eliminar l'alçada de la regressió i deixar només l'excés de pes i l'edat? quins avantatges i quins inconvenients tindria?

## 5.2 Anàlisi de la variància.

**7)** Para veure si afegir edulcorant millora l'engreix de garrins s'ha fet una experiència en la que s'ha mesurat el guany mig diari, *GMD*, de garrins en les mateixes condicions però afegint edulcorant. S'han provat les 5 dosis: *D00*, *D08*, *D15*, *D20* i *D30*, incloent la dosi D00 que de fet és la dieta sense edulcorant, i cada dieta s'ha provat en 5 garrins. Els resultats experimentals obtinguts són al fitxer "`gmd.csv`".
Tractant la dosis com els nivells d'un factor i $\alpha = 0.05$, calculeu:

a) Les matrius dels següents models lineals, les matrius del canvi dels demés models al primer, i per cada model, operant amb les matrius, el valor estimat dels paràmetres:

(1) $y_{ij} = \mu_i + \varepsilon_{ij}$

(2) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ amb $\sum_{i=1}^{k} \alpha_i = 0$

(3) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ amb $\alpha_1 = 0$

b) Utilitzant `R` ajusteu el model utilitzant algun dels models anteriors.
Notes: Per defecte el programa utilitza la 3a parametrització, per fer la primera s'ha d'indicar que es vol sense terme independent `lm(GMD~0+DOSI...)` i per la segona s'ha d'indicar que o faci amb la parametrització SUM, per fer-ho hi ha dues opcions, una és afegir a la comanda `lm(....,contrasts=list(DOSI="contr.sum"))` o bé a l'inici de la sessió executar la comanda `options(contrasts = c("contr.sum", "contr.treatment"))` així en aquesta sessió utilitzarà sempre aquesta parametrització.

c) Plantegeu, contrasteu i interpreteu, el test de l'anàlisi de la variància.
Comproveu que amb la 1a parametrització no contrasta el test per veure si els valors esperats de cada nivell són iguals o diferents, però si acceptem el terme independent, aleshores no importa amb quin dels models contrastem el test.

d) Utilitzant el mètode de Tukey:
Notes: No importa la parametrització escollida i es necessita el paquet del `R: emmeans`.

- Per cada nivell de dosi, estimeu els valors esperats del *GMD* i el seu interval de confiança.
  Nota: comanda `emmeans`.
- Feu les comparacions de totes les dosis entre elles.
  Dieu quines parelles de nivell de dosi, donen un GMD esperat diferent.
  Nota: comandes `emmeans i pairs`.
- Expresseu el resultat anterior de forma compacta.
  Nota: comandes `emmeans i cld`.

**8)** En un tast de patés, cada persona del panell ha tastat i puntuat 5 patés. Per a no influir sobre el sobre les persones del panell, els patés es varen presentar de forma aleatòria i identificats amb codis numèrics que indueixin a ser ordenats. En el fitxer "`pate.csv`" hi ha els resultats del tast, les seves columnes són:

- tastador

- codi del paté
- les valoracions de la qualitat de: color, aroma, textura, sabor i la ordenació. Les puntuacions són de 0 a 10 i l'ordenació és en ordre de preferència ( de 1 a 5).

Per la puntuació del color:

a) Plantegeu el model lineal additiu amb els factors tastador i paté, i escriviu-ne la matriu del model.

b) Ajusteu el model utilitzant el R, i mireu quina matriu ha utilitzat.
Nota: per poder veure quina matriu del model a utilitzat heu d'afegir a la comanda `lm(...,x=T,...)` així amb `"nom del model"$x` ens donarà la matriu.

c) Suposant que es compleixen les condicions per fer el test anova, contrasteu el test adequat per veure si els patés es poden distingir, o no, pel color, feu-ne les comparacions múltiples i interpreteu els resultats.

d) Compareu els resultats de l'apartat anterior amb els que s'obtindrien si ho féssim amb el model de només factor paté, sense tenir en compte els tastadors.

e) Creus que es compleixen les condicions del test anova? Justifiqueu-ho teòricament i amb gràfics de diagnòstic.

Per a cadascuna de les demés valoracions, torneu a fer els apartats anteriors.
Nota: El cas de l'ordenació és molt diferent de les altres valoracions, i queda una versió asimptòtica del test de Friedmann.

**9)** En l'elaboració de formatges, el rendiment és la relació entre el pes del formatge obtingut i el de la llet utilitzada. Es vol veure, per separat en llet de cabra, d'ovella i de vaca; com canvia el rendiment amb un tractament tèrmic (crua/pasteuritzada) i amb l'addició de CaCl2 (si/no). Les dades són al fitxer "`formatges.csv`".
Per cadascun dels tres tipus de llet per separat, suposant que es compleixen les hipòtesis dels models lineals i amb $\alpha = 0.05$:

a) Plantegeu el model factorial i trobeu-ne la matriu del model.
Ajusteu el model amb `R` i mireu quina és la matriu del model que ha utilitzat.
Tenim els subespais vectorials $V_1$, $V_2$, $V_3$ i $V_4$, on $V_1$ és el generat per les columnes de la matriu del model corresponents al terme independent, $V_2$ per les de l'efecte principal del tractament tèrmic, $V_3$ per les de l'efecte principal del clorur càlcic i $V_4$ per les de la interacció. Aquests 4 subespais són ortogonals entre ells? això depèn de la parametrització escollida?

b) Tenen algun efecte els tractaments (combinacions dels nivells dels dos factors) utilitzats?
En cas afirmatiu responeu justificadament les preguntes següents:

- Quin és el millor o els millors tractaments?
- El CaCl2 té algun efecte sobre el rendiment?

- Podem dir que afegir CaCl2 fa augmentar el rendiment?
- El tractament tèrmic té algun efecte sobre el rendiment?
- Podem dir que pasteuritzar fa augmentar el rendiment?
- Podríem utilitzar un altre model lineal de dimensió més petita? per què? quin? canviarien els resultats del test anova? i els de les comparacions múltiples?

**10)** Tenim els resultats de mesurar l'àrea que cobreixen plantes entapissants a talussos de carreteres. S'han estudiat dues especies *E1* i *E2* (factor *ESPECIE*) i haver posat, no, compost al talús (factor *COMPOST*). Les dades experimentals són al fitxer "`area.csv`". Utilitzant $\alpha = 0.05$:

a) Plantegeu el model lineal additiu de la variable *AREA* respecte els factors *ESPECIE* i *COMPOST*.

- Ajusteu aquest model.
- Contrasteu el test anova. Què ens diu el test?
- Contrasteu el test adequat per comprovar que en el model no era necessària la interacció dels dos factors.
- Feu les comparacions múltiples. Què ens diuen?
- Mitjançant gràfics de diagnòstic, comproveu que no es compleixen les hipòtesis dels models lineals. On és el problema?

b) Plantegeu el model lineal additiu de la variable $\log(AREA)$ en lloc de *AREA*.

- Ajusteu aquest model.
- Contrasteu el test anova. Què ens diu el test?
- Contrasteu el test adequat per comprovar que en el model és necessària la interacció dels dos factors.
- Feu les comparacions múltiples del model adequat. Què ens diuen?
- Mitjançant gràfics de diagnòstic, comproveu que ara ja no hi ha el problema que hi havia en el model de la variable *AREA*.
- En estudiar el logaritme de la variable, canvia el sentit dels efectes del model?

**11)** En el fitxer prac2f.csv, hi teniu 8 columnes, las dues primeres $(F1, F2)$ són dos factors de 3 i 4 nivells respectivament, i les 6 restants $(V1, V2, \ldots, V6)$ són resultats experimentals (simulats) de 6 variables aleatòries.
Per cadascuna de les variables aleatòries poseu l'enunciat d'una experiència, de forma que les seves dades experimentals puguin ser les del fitxer. Per separat, amb $\alpha = 0.05$, feu els apartats següents:

a) Descriviu el comportament de les dades mitjançant:

- una taula de mitjanes de l'estil:

| $F1 \diagdown F2$ | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| Total | | | | | |

- el diagrama de mitjanes-errors estàndard, $\overline{X} - S_{\overline{X}}$ .

b) Plantegeu i ajusteu el model lineal factorial.

(1) Contrasteu el test anova.

(2) Feu les comparacions múltiples dels tractaments i les dels factors.
Nota: Pot ajudar transcriure el resultat de les comparacions múltiples a la taula de mitjanes.

(3) Dibuixeu el diagrama de dispersió dels residus vs els valors predits per poder veure si hi ha tendències no explicades i/o heterogeneïtat de variàncies.

c) Amb els resultats de l'apartat anterior contesteu justificadament les preguntes següents:

- Hi ha algun efecte? Com es veu en la taula de mitjanes?
- Quina és la variància de l'error?
- El factor *F1* té efecte? Com es veu a la taula de mitjanes? i al diagrama $\overline{X} - S_{\overline{X}}$?
- El factor *F2* té efecte? Com es veu a la taula de mitjanes? i al diagrama $\overline{X} - S_{\overline{X}}$?
- Hi ha interacció? Com es veu a la taula de mitjanes? i al diagrama $\overline{X} - S_{\overline{X}}$?
- Quin és el millor o millors tractaments?
- Quin és el millor o millors nivells del factor *F1*?
- Quin és el millor o millors nivells del factor *F1*?
- El millor tractament és la combinació dels millors nivells de *F1* i *F2*? per què?
- Si haguessis d'utilitzar el nivell 1 del factor *F1*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 2 del factor *F1*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 3 del factor *F1*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 1 del factor *F2*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 2 del factor *F2*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 3 del factor *F2*, quin tractament escolliries?
- Si haguessis d'utilitzar el nivell 4 del factor *F2*, quin tractament escolliries?
- Es veu alguna tendència anòmala en els residus?
- Es veu alguna tendència en la variància dels residus?

d) Utilitzant el model de dos factors sense interacció feu l'equivalent de l'apartat b) i responeu les preguntes que tinguin sentit de l'apartat c). Dieu si aquest model és adequat.

e) Utilitzant el model de dos factors encaixats ($F1 + F1 : F2$) feu l'equivalent de l'apartat b) i responeu les preguntes que tinguin sentit de l'apartat c). Dieu si aquest model és adequat.

**12)** Es vol estudiar si la degradació, *W2* , de les farines de blat depèn de la varietat del blat i també si afecta la presència, o no, d'insectes a la farina. Les dades experimentals obtingudes són al fitxer "`blat.csv`", on *VAR* indica la varietat, *W2* el valor de la degradació i *PRES* la presència d'insectes. Una característica d'aquestes dades és que hi diferències grans entre el nombre de repeticions dels tractaments.

a) Suposant que es compleixen les condicions del test anova, plantegeu i ajusteu un model lineal factorial de la variable *W2* amb els factors *VAR* i *PRES*. Amb els resultats obtinguts contesteu les preguntes següents:

- La degradació es veu afectada per la varietat i/o la presència d'insectes?
- La degradació es veu afectada per la presència d'insectes?
- La degradació es veu afectada per la varietat?
- Compara els resultats utilitzant sumes de quadrats tipus I i les de tipus II. L'ordre dels factors en la fórmula del model, té algun efecte sobre els resultats dels tests?
- En aquest cas, quin paper juga la interacció? podríem prescindir d'ella?

b) Suposant que es compleixen les condicions del test anova, plantegeu i ajusteu un model lineal de la variable *W2* amb els factors *VAR* i *PRES* sense interacció. Amb els resultats obtinguts contesteu les preguntes de l'apartat a) que tinguin sentit.

## 5.3 Anàlisi de la covariància.

**13)** Es vol comparar les valoracions, $v$, de dos mètodes pedagògics, $m$, en funció del coeficient d'intel·ligència, $c$, amb les dades del fitxer "`comrect.csv`" .
Plantegeu i ajusteu el model lineal factorial de la variable $v$ en funció del factor $m$ i de la covariable $c$. Suposant que es compleixen les condicions dels models lineal, amb $\alpha = 0.05$, a més d'estimar els paràmetres i contrastar els test de cada paràmetre (paràmetre=0, o no), contrasteu també els test adequats per contestar les preguntes següents:

a) Les dues rectes són iguals? o no?

b) Les dues rectes són paral·leles? o no?

c) Les dues rectes tenen el mateix terme independent? o no?

d) Per cadascun dels següents valores de la variable $c$: 90, 105 i 120, quines diferències hi ha en la valoració segons el mètode?

Repetiu l'exercici, però amb les valoracions *vv*.

**14)** En un estudi per veure com afecta, a primera hora del matí, l'estrès hídric sobre la fisiologia d'unes plantes, s'ha plantejat una experiència per mesurar la fotosíntesi, *FS*, i la transpiració, *TR*, de les plantes sotmeses a un factor estrès hídric, *SHIDR*, amb 4 nivells *S1*, *S2*, *S3* i *S4*, corresponents al nombre de dies sense ser regades. En total s'han mesurat 20 plantes, 5 per cada nivell d'estrès. Un problema que es va presentar en fer l'experiència és que per mesurar la *FS* i la *TR* d'una planta es triga aproximadament 3 minuts, o sigui en total una hora. Però a primera hora del matí les condicions atmosfèriques canvien ràpidament, per poder compensar-ho es va mesurar també el temps, $T$, des de l'inici de prendre mesures i la radiació solar, *RAD*, en el moment de fer la mesura. Totes les dades són al fitxer "`shidr.csv`" on hi ha també la variable *REP*, que és el número de repetició de la planta dins del nivell de SHIDR.
Suposant que es compleixen les condicions dels models lineals, amb $\alpha = 0.05$, per a cascuna de les parelles $(FS, RAD)$, $(FS, T)$, $(TR, RAD)$ i $(TR, T)$, que anomenarem $(Y, X)$:

a) Dibuixeu la gràfica de dispersió $(Y, X)$, junt amb les rectes de regressió de $Y \sim X$ dins de cada nivell del factor *SHIDR*.

b) Plantegeu, i ajusteu, el model factorial de la variable $Y$ en funció del factor $SHIDR$ i de la covariable $X$.

c) Contrasteu els test adequats per a contestar les preguntes:

- El factor $SHIDR$ i la covariable $X$ afecten a la variable $Y$?
- Les 4 rectes són paral·leles? en el cas que no ho siguin dieu quines diferències hi ha en els pendents.

- En el valor mitja de la covariable $X$, hi ha diferències en els valors estimats de la $Y$? en cas afirmatiu, quines diferències hi ha?
- Per un valor petit de $X$, $RAD = 350$ o bé $T = 0$, hi ha diferències en els valors estimats de la $Y$? en cas afirmatiu, quines diferències hi ha?
- Per un valor alt de $X$, $RAD = 650$ o bé $T = 60$, hi ha diferències en els valors estimats de la $Y$? en cas afirmatiu, quines diferències hi ha?
- Es podria plantejar un altre model de dimensió més petita? per què?

d) Contesteu també:

- Acompanyant el factor $SHIDR$, quina covariable, $RAD$ o $T$, afecta més a la variable $FS$?
- i a la variable $TR$?

**15)** En el fitxer "`pracovar.csv`", hi teniu les dades (simulades) d'una experiència amb:

- un factor, *FACTOR*.
- una covariable o variable independent, *X*.
- diverses variables dependents *Y1, Y2, ..., Y8*.

Per a cadascuna de les variables dependents, *Y1, Y2, ..., Y8*, per separat:

a) Poseu un enunciat d'una experiència, de forma que les seves dades experimentals puguin ser les del fitxer i feu la descriptiva següent:

- Gràfica dels punts amb les rectes de regressió de cada nivell del factor.

b) Plantegeu un model d'anàlisi de la covariància factorial, ajusteu-lo i contrasteu els tests necessaris per contestar les preguntes següents:

- Hi ha algun efecte?
- Hi ha algun efecte del factor? $\leftrightarrow$ Les rectes són iguals?
- Hi ha algun efecte de la covariable? $\leftrightarrow$ Les rectes són horitzontals?
- El factor i la covariable interaccionen? $\leftrightarrow$ Les rectes són paral·leles?
- El model és adequat? segons la gràfica de diagnòstic: residuals vs predits.

c) Plantegeu un model d'anàlisi de la covariància sense interacció, ajusteu-lo i contrasteu els tests necessaris per contestar les preguntes següents:

- Hi ha algun efecte?
- Hi ha algun efecte del factor? $\leftrightarrow$ Les rectes són iguals?
- Hi ha algun efecte de la covariable? $\leftrightarrow$ Les rectes són horitzontals?
- El model és adequat? segons la gràfica de diagnòstic: residuals vs predits.
- Quin model us sembla més adequat? el factorial o sense interacció.

- Hi ha algun altre model de dimensió més petita que sigui adequat? quin?

**16)** Es vol comparar com evoluciona, en passar el temps, el contingut de vitamina C d'un suc de taronja natural, depenent del tipus d'envàs i la temperatura de conservació, dels que hem escollit tres tractaments, "***a***", "***b***" i "***c***" .

Durant 12 setmanes, començant al cap d'una setmana després de l'envasat dels sucs, s'-ha analitzat el contingut de vitamina C de dues unitats de cada tractament. Les dades experimentals (simulades) són al fitxer "`vitc.csv`".
En general sabem que el contingut de vitamina C evoluciona seguint una funció exponencial, $vitc = \alpha e^{-\beta \cdot set}$,on $\alpha > 0$, $\beta > 0$ , i el valor d'aquests paràmetres depèn del tractament de conservació. Amb $\alpha = 0.05$, feu:

a) Sabent que a l'envasat tots els sucs tenien el mateix contingut de vitamina C, plantegeu, i ajusteu, un model lineal adequat per veure si en els tres tractaments la pèrdua de vitamina C es fa amb la mateixa rapidesa, o no. Amb aquest model:

- Estimeu el contingut de vitamina C esperat, en el moment de l'envasat.
- Per a cada tractament calculeu el valor estimat de la $\beta$.
- Dieu si les tres $\beta's$ són totes iguals, o no, i quines diferències hi ha.
- Quin tractament, o tractaments, conserven més la vitamina C?

b) Test blanc: Plantegeu, i ajusteu, un altre model de dimensió més gran, amb el que es pugui comprovar que a l'envasat, els sucs dels tres tractaments tenien realment el mateix nivell de vitamina C. Amb aquest model:

- Estimeu, per cada tractament, el contingut esperat de vitamina C que tenien en ser envasats, es a dir quan $set = 0$.
- Els continguts de vitamina C a $set = 0$, són iguals? o no?
- Amb aquest model, quin tractament, o tractaments, conserven més la $vitc$? El resultat coincideix amb la mateixa pregunta de l'apartat a)? per què?

c) Justifiqueu si els models lineals utilitzats compleixen totes les hipòtesis dels *"models lineals"* i dieu si ajusten bé les dades.