

ESTADÍSTICA

Examen Final (Part A)

Grau Matemàtiques

Dilluns 31 de Maig de 2021, 9.00-12.00, Aules 001,002

Podeu fer ús de la calculadora, la llista de distribucions que surt a l'Atenea del curs i les taules estadístiques.

TEORIA 1 (1 punt)

La funció `rnorm(m)` de R permet generar una mostra pseudo-aleatòria de m observacions $N(0, 1)$. Indica amb codi de R o pseudocodi com fer servir aquesta funció per obtenir...

- a). ...una mostra de n observacions χ_k^2
- ```
chisq<-function(n,k) {mat<-matrix(rnorm(n*k)^2,ncol=k,nrow=n); x<-apply(mat,1,sum)}
```
- b). ...una mostra de  $n$  observacions  $t_k$
- ```
tStud<-function(n,k) {x<-rnorm(n)/sqrt(chisq(n,k)/k)}
```
- c). ...una mostra de n observacions F_{k_1, k_2}
- ```
FSned<-function(n,k1,k2) {x<-(chisq(n,k1)/k1)/(chisq(n,k2)/k2)}
```

### TEORIA 2 (1 punt)

Considerem la variable aleatòria  $X \sim N(0, \sigma^2)$ .

- a). Demostreu que la distribució Inversa Gamma és prior conjugada pel model anterior pel paràmetre  $\sigma^2$ .

Una distribució és una família conjugada pel paràmetre d'un model si quan considerem una distribució a priori d'aquesta família, la distribució a posteriori és de la mateixa família.

Distribució a priori pel paràmetre  $\sigma^2$ :  $\sigma^2 \sim IG(\alpha, \beta)$

$$\pi(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2}$$

Versemblança del model:  $X \sim N(0, \sigma^2)$

$$f(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$$

Distribució a posteriori:  $\sigma^2|X \propto \pi(\sigma^2)f(X|\sigma^2)$

$$\pi(\sigma^2|X) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2} \frac{1}{\sqrt{2\pi}} (\sigma^2)^{-1/2} e^{-x^2/2\sigma^2} \propto (\sigma^2)^{-\alpha-1/2-1} e^{-(\beta+x^2/2)/\sigma^2}$$

I per tant, la distribució a posteriori també és una Inversa Gamma:  $\sigma^2|X \sim IG(\alpha + 1/2, \beta + X^2/2)$

- b). Donada una m.a.s.  $X_1, \dots, X_n$  de  $X$ , i suposant que la distribució a priori de  $\sigma^2$  és una Inversa Gamma  $IG(\alpha, \beta)$  trobeu la distribució a posteriori

Fent el mateix càlcul però incorporant la funció de versemblança calculada a partir de la funció de densitat conjunta:

Versemblança del model:  $X \sim N(0, \sigma^2)$

$$f(x_1, \dots, x_n|\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\sum_{i=1}^n x_i^2/2\sigma^2}$$

$$\pi(\sigma^2|X) \propto (\sigma^2)^{-\alpha-n/2-1} e^{-(\beta+\sum_{i=1}^n x_i^2/2)/\sigma^2}$$

I per tant,  $\sigma^2|X_1, \dots, X_n \sim IG(\alpha + n/2, \beta + \sum_{i=1}^n X_i^2/2)$

Nota: Densitat  $IG(\alpha, \beta)$ :  $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$

### TEORIA 3 (1 punt)

Sigui  $X_1, \dots, X_n$  una m.a.s. de  $X \sim \text{Poisson}(\lambda)$ . Considerem un estimador de  $\lambda$  de la forma:

$$\hat{\lambda} = K \sum_{i=1}^n i X_i$$

- a). Trobeu el valor de  $K$  per tal que l'estimador no tingui biaix

Un estimador d'un paràmetre no té biaix si el seu valor esperat coincideix amb el paràmetre.

$$E(\hat{\lambda}) = E\left(K \sum_{i=1}^n i X_i\right) = K \sum_{i=1}^n i E(X_i) = K \lambda \sum_{i=1}^n i$$

Per tant, si l'estimador és sense biaix es compleix que:

$$K \sum_{i=1}^n i = 1 \Rightarrow K = \frac{1}{\sum_{i=1}^n i} = \frac{2}{n(n+1)}$$

- b). Determineu si l'estimador trobat a l'apartat anterior és eficient o no. Justifiqueu la resposta

Un estimador és eficient si no té biaix i la seva variància coincideix amb la fita de Cramer-Rao. Primer calculem la variància de l'estimador:

$$V(\hat{\lambda}) = V\left(\frac{2}{n(n+1)} \sum_{i=1}^n i X_i\right) = \frac{4}{n^2(n+1)^2} \sum_{i=1}^n i^2 V(X_i) = \frac{4}{n^2(n+1)^2} \frac{n(n+1)(2n+1)}{6} \lambda = \frac{2}{3} \frac{2n+1}{n(n+1)} \lambda$$

I la cota de Cramer-Rao és la inversa de la informació de Fisher:

$$\begin{aligned} I_n(\lambda) &= E\left(-\frac{\partial^2}{\partial \lambda^2} \log L(\lambda|X_1, \dots, X_n)\right) = E\left(-\frac{\partial^2}{\partial \lambda^2} (-n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \log(\prod_{i=1}^n X_i!))\right) = \\ &= E\left(\frac{\sum_{i=1}^n X_i}{\lambda^2}\right) = \frac{n}{\lambda} \end{aligned}$$

Així doncs, la cota de Cramer-Rao és:  $\frac{\lambda}{n}$  La eficiència relativa de l'estimador és:

$$Ef(\hat{\lambda}) = \frac{CCR}{V(\hat{\lambda})} = \frac{3}{2} \frac{n+1}{2n+1}$$

La eficiència relativa només pot valdre 1 (que implicaria que la variància de l'estimador assoleix la fita de Cramer-Rao) pel cas trivial  $n = 1$ . Per tant, si  $n > 1$  l'estimador no és eficient.

Indicació:  $\sum_{i=1}^n i = \frac{n(n+1)}{2}$        $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$

### TEORIA 4 (1 punt)

La publicitat d'un producte dietètic indica que la reducció esperada del pes quan es pren el tractament és superior a 5 kg en un mes. Per tal de verificar aquesta afirmació, es selecciona una mostra de 15 individus i es pren nota del seu pes basal,  $X_1, \dots, X_{15}$ . Tots els individus es sotmeten al tractament i al cap d'un mes es registra el seu pes respectiu,  $Y_1, \dots, Y_{15}$ .

- a). Es tracta d'un disseny de dues mostres aparellades o independents? Indica l'estadístic de contrast i la distribució de referència sota la hipòtesi nul·la que permet resoldre el test

Es tracta d'un disseny aparellat ja que les dues mostres provenen dels mateixos individus, abans i després del tractament. Això vol dir que hi ha  $n = 15$  unitats experimentals i per comparar els pesos abans ( $X$ ) i després ( $Y$ ) del tractament treballarem amb la variable Diferència:  $D = X - Y$ , que es pot interpretar com la reducció del pes després del tractament. El test que volem resoldre és unilateral, ja que volem veure si la reducció és superior a 5:

$$\begin{cases} H_0 : \mu_D \leq 5 \\ H_1 : \mu_D > 5 \end{cases}$$

L'estadístic de contrast i la distribució de referència sota  $H_0$  són:

$$\frac{\bar{D} - 5}{S_D/\sqrt{n}} \sim t_{n-1}$$

b). La descriptiva de les mostres ha donat el següent resultat:

| Variable       | Mitjana | Desv. Estand. |
|----------------|---------|---------------|
| X              | 77.98   | 5.43          |
| Y              | 72.71   | 5.55          |
| Diferència X-Y | 5.27    | 0.88          |

Resol el test per determinar si la reducció de pes és superior a 5 kg.

Calculem l'estadístic:

$$\hat{t} = \frac{\bar{D} - 5}{S_D/\sqrt{n}} = \frac{5.27 - 5}{0.88/\sqrt{15}} = 1.188$$

La regió crítica del test és  $C = \{D : \hat{t} > t_{14,0.95}\}$  i  $t_{14,0.95} = 1.761$ . Com l'estadístic és inferior al valor crític, vol dir que no es pot rebutjar  $H_0$  ( $\alpha = 0.05$ ). La conclusió pràctica és que les dades no presenten evidències estadístiques significatives que permetin establir que la reducció el pes és superior a 5kg.

### PROBLEMA 1 (4 punts)

Les variables aleatòries  $Y_1, Y_2, \dots, Y_n$  són independents i idènticament distribuïdes amb funció de probabilitat:

$$P_\theta(Y = y) = \binom{y-1}{r-1} \theta^r (1-\theta)^{y-r} \quad \text{per } y = r, r+1, r+2, \dots, \quad (1)$$

on  $r$  és un nombre natural conegut i on  $\theta \in (0, 1)$ . El valor esperat i la variància són:

$$E_\theta[Y] = \frac{r}{\theta} \quad \text{i} \quad V_\theta[Y] = \frac{r(1-\theta)}{\theta^2}. \quad (2)$$

a). Quina és la funció de versemblança i la del logaritme de la versemblança? Dibuixa la funció de versemblança per  $r = 2$ ,  $n = 1$  i  $Y = 2$ .

La funció de versemblança és:

$$L_y(\theta) \propto \theta^{nr} (1-\theta)^{\sum (y_i - r)}, \quad (3)$$

per  $\theta \in (0, 1)$ , i el logaritme de la funció de versemblança és:

$$l_y(\theta) = \text{const} + nr \log \theta + \left( \sum (y_i - r) \right) \log (1 - \theta). \quad (4)$$

Quan  $r = 2$ ,  $n = 1$  i  $Y = 2$ ,

$$L_{y=2}(\theta) \propto \theta^2. \quad (5)$$

b). Troba tres estimadors de  $\theta$  basats en el mètode de moments, utilitzant  $E_\theta[Y]$ ,  $E_\theta[Y^2]$  i  $V_\theta[Y]$ . Fes servir els tres estimadors en el cas en el que  $r = 2$ ,  $n = 2$  i  $Y = 2, 3$ .

Igualant  $E_\theta[Y]$  amb  $\bar{y}$  s'obté:

$$\hat{\theta}_1 = \frac{r}{\bar{y}}, \quad (6)$$

igualant:

$$E_\theta[Y^2] = \frac{r + r^2}{\theta^2} - \frac{r}{\theta} \quad (7)$$

amb  $\bar{y}^2$  s'obté:

$$\hat{\theta}_2 = \frac{-r + \sqrt{(1 + 4\bar{y}^2)r^2 + 4\bar{y}^2 r}}{2\bar{y}^2}, \quad (8)$$

igualant  $V_\theta[Y]$  amb  $s_y^2$  s'obté:

$$\hat{\theta}_3 = \frac{-r + \sqrt{r^2 + 4s_y^2 r}}{2s_y^2}. \quad (9)$$

Per  $r = 2$ ,  $n = 2$  i  $Y = 2, 3$ ,  $\hat{\theta}_1 = 0.8$ ,  $\hat{\theta}_2 = 0.8192$  i  $\hat{\theta}_3 = 0.8284$  si  $s_y^2 = 0.5$  (o  $\hat{\theta}_3 = 0.8984$  si  $s_y^2 = 0.25$ ).

c). Troba l'estimador màxim versemblant de  $\theta$  i la seva distribució assintòtica.

Resolent l'equació:

$$\frac{dl_y}{d\theta} = \frac{nr}{\theta} - \frac{\sum(y_i - r)}{1 - \theta} = 0, \quad (10)$$

s'obté:

$$\hat{\theta}_{MV} = \frac{r}{\bar{y}}, \quad (11)$$

que és un màxim perquè la segona derivada de  $l_y(\theta)$ ,

$$\frac{d^2 l_y}{d\theta^2} = -\frac{nr}{\theta^2} - \frac{\sum(y_i - r)}{(1 - \theta)^2}, \quad (12)$$

avaluada a  $\hat{\theta}_{MV}$  és negativa. La informació de Fisher de la mostra de  $n$  observacions,  $Y$ , és:

$$I_Y(\theta) = E\left[\frac{nr}{\theta^2} + \frac{\sum(y_i - r)}{(1 - \theta)^2}\right] = \frac{nr}{\theta^2(1 - \theta)}. \quad (13)$$

Per tant, quan  $n$  és gran la distribució de  $\hat{\theta}_{MV}$  es pot aproximar per:

$$\text{Normal}(\theta, \text{Var}(\hat{\theta}_{MV}) = \frac{\theta^2(1 - \theta)}{nr}). \quad (14)$$

d). Construeix un interval de confiança aproximat per  $\theta$ , utilitzant l'estimador màxim versemblant i la seva distribució assintòtica.

Un interval de confiança del 95% aproximat seria, per exemple:

$$\theta \in [\hat{\theta}_{MV} \pm 1.96 \sqrt{\frac{\hat{\theta}_{MV}^2(1 - \hat{\theta}_{MV})}{nr}}]. \quad (15)$$

e). Dissenya un procediment UMP amb nivell de significació  $\alpha = 0.05$  per fer la prova d'hipòtesi:

$$\begin{cases} H_0 : \theta \geq .5 \\ H_1 : \theta < .5 \end{cases}$$

Com que el likelihood ratio:

$$\Lambda(Y) = \frac{L_Y(\theta_1)}{L_Y(\theta_0)} = \frac{\theta_1^{nr}(1 - \theta_0)^{nr}}{\theta_0^{nr}(1 - \theta_1)^{nr}} \frac{(1 - \theta_1)^{\sum y_i}}{(1 - \theta_0)^{\sum y_i}}, \quad (16)$$

quan  $\theta_1 < \theta_0$  es monòton creixent amb  $\sum y_i$ , per el teorema de Karlin i Rubin tenim que el test amb regió crítica de rebuig:

$$C = \{Y : \sum y_i \geq C\}, \quad (17)$$

és un test UMP amb nivell  $\alpha = P(\sum y_i \geq C | \theta = 0.5)$ . La constant  $C$  es determina resolent:

$$P(\sum y_i \geq C | \theta = 0.5) \leq 0.05 \quad (18)$$

(La distribució de  $\sum y_i$  és com la de  $y_i$  canviant  $r$  per  $nr$ ).

f). Dissenya un procediment amb nivell de significació  $\alpha = 0.05$  per fer la prova d'hipòtesi:

$$\begin{cases} H_0 : \theta = .5 \\ H_1 : \theta \neq .5 \end{cases}$$

El test de raó de versemblança generalitzat fa servir l'estadístic:

$$\lambda(Y) = \frac{0.5^{\sum y_i}}{(\frac{r}{\bar{y}})^{nr}(1 - \frac{r}{\bar{y}})^{\sum(y_i - r)}} = \frac{(0.5\bar{y})^{\sum y_i}}{r^{nr}(\bar{y} - r)^{\sum(y_i - r)}}, \quad (19)$$

i té per regió crítica, de rebuig:

$$C = \{Y : \lambda(Y) \leq A\}, \quad (20)$$

on  $A$  és tal que  $P(\lambda(Y) \leq A | \theta = .5) \leq 0.05$ . Aquesta  $A$  es pot determinar fent servir el fet que quan  $\theta = .5$ , la distribució de  $-2 \log(\lambda(Y))$  es pot aproximar per una  $\chi_1^2$ . (Com a alternativa, aquesta distribució també permet aproximar el p-valor).

- g). En un estudi sobre el nombre de vegades que el nom dels personatges d'una novel·la apareixen a la novel·la,  $Y$ , s'observa que:

|            |    |    |    |   |   |   |   |   |   |    |    |    |
|------------|----|----|----|---|---|---|---|---|---|----|----|----|
| Y          | 1  | 2  | 3  | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| frequència | 45 | 31 | 20 | 9 | 6 | 5 | 4 | 2 | 1 | 1  | 2  | 1  |

Suposant que  $Y_i$  per  $i = 1, \dots, 127$ , és una mostra i.i.d. amb funció de probabilitat com la d'aquest problema amb  $r = 1$ , calcula el valor de l'estimador màxim versemblant de  $\theta$  i un interval de confiança aproximat del 95% per  $\theta$ .

$$\hat{\theta}_{MV} = \frac{1}{\bar{y}} = \frac{127}{360} = 0.3528 \quad (21)$$

$$\theta \in [0.3528 \pm 1.96 \sqrt{\frac{0.3528^2(1 - 0.3528)}{127}}] = [0.3528 \pm 1.96 * 0.02518] = [0.3034, 0.4022]. \quad (22)$$

- h). Explica com ho faries per decidir si és creïble que la mostra de la pregunta anterior sigui independent i idènticament distribuïda amb funció de probabilitat com la d'aquest problema amb  $r = 1$ . Dona tots els detalls necessaris per demostrar que sabries implementar aquest test.

Veure les transparències de l'assignatura.

#### PROBLEMA 2: (1 punt)

A l'arxiu *Barella0.csv* hi teniu els resultats d'un experiment factorial  $3^3$ , fet als anys 50 del segle passat en una indústria tèxtil del Vallès, en el que la variable resposta,  $y$ , s'observa a tres nivells de cada una de les tres variables explicatives,  $x_1, x_2$  i  $x_3$ .

Les variables explicatives les heu de tractar com a quantitatives, codificades com a  $-1, 0$  i  $1$ , i no com a factors. Es a dir, no les tracteu com a categòriques. Observeu que l'arxiu *Barella0.csv* ja us dona els termes quadràtics i els productes creuats calculats.

- a). Ajusta el model lineal de  $y$  amb només els termes  $x_1, x_2$  i  $x_3$ , i compara'l amb el model lineal de  $y$  amb els termes  $x_1, x_2$  i  $x_3$ , amb les seves tres interaccions, i amb els seus tres termes quadràtics. Quin d'aquests dos model triaries, i perquè?

El residus del model que només inclou els tres termes lineals indiquen que estas ajustant un pla a una relació que no és plana, que la variància no és constant sino que augmenta amb el valor esperat, i que falla la hipòtesi de normalitat. Per això aquest model no és acceptable. Els residus del model amb termes lineals, quadràtics i interaccions no són "ideals", però són més acceptables, i gairebé tots els seus coeficients són significativament diferents de 0.

- b). Repeteix l'exercici anterior modelant primer  $\sqrt{y}$ , després  $1/y$  i finalment  $\log(y)$  en comptes de  $y$ . D'entre tots els vuit models que has ajustat, quin triaries i perquè?

Triaria el model que relaciona  $\log(y)$  amb els tres termes lineals i prou, perquè els seus residus indiquen que es satisfan les suposicions del model lineal, i perquè quan hi afegeixes interaccions i/o termes quadràtics, els coeficients d'aquests termes no són significativament diferents de 0.

Tots els altres set models són més complicats i/o tenen residus que indiquen que no es satisfan les suposicions del model lineal.

#### PROBLEMA 3: (1 punt)

Es vol estudiar si la degradació de la farina,  $W2$ , depèn de la varietat de blat,  $VAR$ , i/o de la presència d'insectes,  $PRES$ . Les dades les teniu a l'arxiu *Farina0.csv*. Observa que el nombre de repeticions és molt diferent en diferents condicions experimentals.

- a). Quin és el model lineal que representa millor la relació de  $W2$  amb  $VAR$  i  $PRES$ ? Justifica perquè esculls el model que esculls, i perquè descartes els que descartes.

El millor model és el de  $W2$  amb  $VAR$ .

Els residus del model amb interaccions son acceptables, i ni les interaccions de  $VAR$  amb  $PRES$ , ni la variable  $PRES$  en el model sense interaccions, tenen coeficients significativament diferents de zero, i per tant es poden eliminar del model.

- b). Escull el gràfic o els gràfics que creguis que resumeixen més be la relació entre  $W2$  i les variables explicatives. Fes servir aquest(s) gràfic(s) per explicar les conclusions a les que arribes sobre l'efecte de la varietat i la presència d'insectes sobre la degradació de la farina.

El millor resum gràfic, és el de les mitjanes en funció de varietat,  $VAR$ , calculat partint del model lineal ajustat amb  $VAR$  i prou, (fent servir la funció `emmip` de la llibreria `emmeans`).

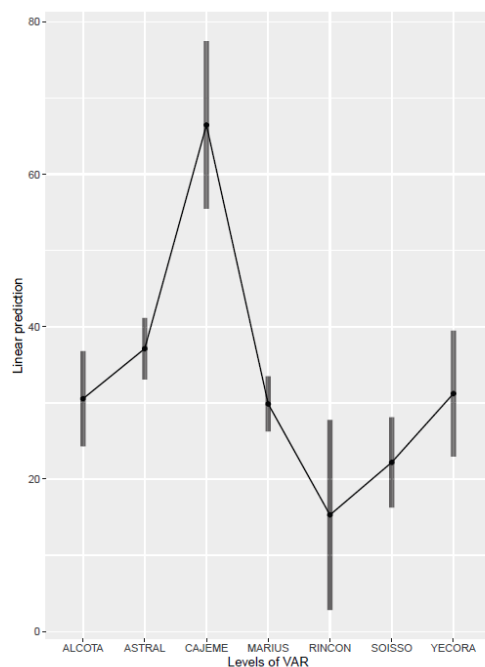


Figura 1: Gràfic de mitjanes en funció de varietat.

La presència d'insectes no afecta la degradació de la farina.

La varietat per la que la farina es degrada més, és la CAJEME. La diferència de la degradació entre les altres varietats no es veu gaire significativa, tot i que al fer les comparacions entre parelles de varietats, la ASTRAL sembla ser lleugerament pitjor que RINCON i SOISSO.