STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik

MT 5003
TENTAMEN
25 augusti 2020

Tentamen i Statistisk inferensteori
25 augusti 2020, kl. 9-14

Examinator: Chun-Biu Li, cbli@math.su.se.
Återlämning: Tillkännages via e-post.

Resonemang skall vara tydliga och lätta att följa. Varje korrekt och fullständigt löst uppgift ger 10 poäng. För betyg A-E krävs 20 poäng på Del 1, samt att följande gränser uppnås på Del 2:

A	B	C	D	E
25	19	13	7	0

Del 1

Låt $X_i,\ i = 1, \ldots, n$ beteckna oberoende stokastiska variabler med täthetsfunktioner $f(x_i|\theta) = 2\theta x_i \exp(-\theta x_i^2),\ x_i > 0$, där $\theta \in (0, \infty)$. Vi observerar $x = (x_1, \ldots, x_n)$, en realisering av $X = (X_1, \ldots, X_n)$ för något okänt värde på θ. För fördelningen gäller att $E(X_i) = \sqrt{\pi/(4\theta)}$ och $E(X_i^2) = 1/\theta$.

Uppgift 1

a) Vad kallas en stickprovsvariabel t med egenskapen att fördelningen för $X|t(X) = t(x)$ inte beror på θ? (2p)

Lösning: En tillräcklig (sufficient) stickprovsvariabel.

b) Visa att $t(x) = \sum_{i=1}^n x_i^2$ besitter egenskapen i a). (4p)

Lösning: Enligt faktoriseringskriteriet är t tillräcklig (och besitter då egenskapen i a)) om tätheten kan faktoriseras som $f(x|\theta) = g(t(x), \theta)h(x)$. Genom att utnyttja att variablerna är obeoende får vi

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^n f(x_i|\theta) = (2\theta)^n \prod_{i=1}^n x_i \exp(-\theta x_i^2) \\ &= (2\theta)^n \exp(-\theta t(x)) \times \prod_{i=1}^n x_i = g(t(x), \theta) \times h(x), \end{aligned}$$

och drar slutsatsen att t är tillräcklig.

c) Bestäm maximum-likelihood skattaren av $\phi = 1/\theta$ och visa att den är väntevärdesriktig. (4p)

Lösning: Log-likelihood funktionen

$$l(\theta) = n\log(2\theta) - \theta \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \log(x_i),$$

maximeras av $\hat{\theta} = n/\sum_{i=1}^{n} x_i^2$. Maximum-likelihood skattaren av $\phi = 1/\theta$ ges därför av $\hat{\phi} = 1/\hat{\theta} = \sum_{i=1}^{n} x_i^2/n$. Då $E(\hat{\phi}) = E(X_1^2) = 1/\theta = \phi$ är den väntevärdesriktig.

Uppgift 2

a) Låt apriorifördelningen för θ vara en $Exp(\lambda)$ fördelning med täthetsfunktion $f(\theta|\lambda) = \lambda\exp(-\lambda\theta)$ för ett givet värde på λ. Visa att aposteriorifördelningen för θ kan skrivas på formen $f(\theta|x) \propto \theta^a \exp(-b\theta)$ och bestäm a och b. (5p)

Lösning: Aposteriorifördelningen

$$f(\theta|x) \propto f(x|\theta) \times f(\theta|\lambda) \propto \theta^n \exp(-\theta \sum_{i=1}^{n} x_i^2) \times \exp(-\lambda\theta)$$

$$= \theta^n \exp(-(\sum_{i=1}^{n} x_i^2 + \lambda)\theta),$$

är på den efterfrågade formen med $a = n$ och $b = \sum_{i=1}^{n} x_i^2 + \lambda$.

b) Beskriv hur man kan använda parametrisk Bootstrap för att konstruera ett 95% konfidensintervall för θ. Din beskrivning skall vara så detaljerad att en programmerare kan följa den utan att göra egna härledningar av resultat. Du kan anta att programmeringsspråket kan simulera oberoende slumptal från fördelningen för X_i givet ett värde på θ. (5p)

Lösning: Flera möjligheter, t.ex.

- Bestäm värdet på $\hat{\theta} = n/\sum_{i=1}^{n} x_i^2$.
- Simulera $N = 10000$ vektorer $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})$, $j = 1, \ldots, N$, innehållande oberoende slumptal från $f(x|\hat{\theta})$.
- För varje vektor $x^{(j)}$ i föregående steg, bestäm $\hat{\theta}_j = n/\sum_{i=1}^{n}(x_i^{(j)})^2$.
- Sortera värdena $\hat{\theta}_1, \ldots, \hat{\theta}_N$ i växande ordning, ett konfidensintervall ges med undre gräns i värde nr 250 och övre i värde nr 9750 i storleksordning.

Uppgift 3

a) Låt $\hat{\phi}$ vara maximum-likelihood skattaren i 1c), bestäm gränsvärdet $\lim_{n\to\infty} nVar(\hat{\phi})$. (5p)

Lösning:
Vi har att $\hat{\phi} = g(\hat{\theta}) = 1/\hat{\theta}$. Ur maximum-likelihood skattarens asymptotik fås att $\lim_{n\to\infty} nVar(\hat{\theta}) = 1/J_1(\theta)$, där $J_1(\theta)$ är förväntad Fisherinformation i en observation. Vi har genom att derivera loglikehoodfunktionen två gånger att

$$J_1(\theta) = -E(d^2 \log(f(X_1|\theta))/d\theta^2) = 1/\theta^2.$$

Den efterfrågade variansen fås nu med hjälp deltametoden som

$$g'(\theta)^2 J_1(\theta)^{-1} = 1/\theta^2.$$

b) Bestäm ett 95% konfidensintervall för θ baserat på score-funktionen. (5p)

Lösning:
Ett score-baserat intervall fås som

$$\{\theta; -1.96 < S(\theta)/\sqrt{J(\theta)} < 1.96\} = \{\theta; -1.96 < (n - \theta \sum_{i=1}^{n} x_i^2)/\sqrt{n} < 1.96\},$$

efter omflyttning av termer blir detta

$$((n - 1.96\sqrt{n})/\sum_{i=1}^{n} x_i^2, (n + 1.96\sqrt{n})/\sum_{i=1}^{n} x_i^2).$$

Del 2

Uppgift 4

Låt $f(x|\theta)$ betäckna täthetsfunktionen till exponentialfördelningen med parameter θ, där $f(x|\theta) = \frac{1}{\theta}\exp[-x/\theta]$ om $x \geq 0$ och $f(x|\theta) = 0$ annars. Vi har två mängder med realiseringar av oberoende slumpvariabler, $\{X_1, X_2, \cdots, X_n\}$ där varje element är ett stickprov från en exponentialfördelning med parameter θ_1, och $\{X'_1, X'_2, \cdots, X'_m\}$ där varje element är ett stickprov från en exponentialfördelning med parameter θ_2. Anta att $n = 55$ och att $m = 80$. Anta vidare att stickprovens medelvärden är $\bar{X} = 0.88$ respektive $\bar{X}' = 1.07$.
Utför ett statistiskt hypotestest med nollhypotesen H_0: $\theta_1 = \theta_2$ och alternativhypotesen H_{A}: $\theta_1 \neq \theta_2$, använd likelihood ratio testet med signifikansnivån 5%. (10p)

Lösning:
Likelihood: $f(x,x'|\theta_1,\theta_2) = \prod_{i=1}^n \frac{1}{\theta_1}\exp(-x_i/\theta_1)\prod_{j=1}^m \frac{1}{\theta_2}\exp(-x'_j/\theta_2) = \frac{1}{\theta_1^n\theta_2^m}\exp(-\sum_i x_i/\theta_1 - \sum_j x'_j/\theta_2)$. Log-likelihood: $l(\theta_1,\theta_2) = -n\ln\theta_1 - m\ln\theta_2 - \sum_i x_i/\theta_1 - \sum_j x'_j/\theta_2$.
ML-estimation $\hat\theta_1$ and $\hat\theta_2$: $\frac{\partial}{\partial\theta_1}l(\theta_1,\theta_2) = 0 \Rightarrow \hat\theta_1 = \bar X$ and $\frac{\partial}{\partial\theta_2}l(\theta_1,\theta_2) = 0 \Rightarrow \hat\theta_2 = \bar X'$.
Under hypothesis $H_0 : \theta_1 = \theta_2 = \theta$, $l(\theta_1=\theta,\theta_2=\theta) = -(n+m)\ln\theta - \frac{1}{\theta}(\sum_i x_i + \sum_j x'_j)$. $\frac{\partial}{\partial\theta}l(\theta_1=\theta,\theta_2=\theta) = -(n+m)/\theta + \frac{1}{\theta^2}(\sum_i x_i + \sum_j x'_j) = 0 \Rightarrow \hat\theta = \frac{n\bar X + m\bar X'}{n+m}$.
Likelihood ratio: $\Lambda = \frac{f(x,x'|\hat\theta_1,\hat\theta_2)}{f(x,x'|\hat\theta,\hat\theta)} = \left(\frac{n\bar X + m\bar X'}{n+m}\right)^{n+m}/(\bar X^n \bar X'^m)$.
Likelihood ratio test: Reject H_0 if $\Lambda \geq K \Leftrightarrow$ reject H_0 if $2\ln\Lambda \geq K'$, where K' is determined by $P(2\ln\Lambda \geq K') = \alpha = 0.05 \Rightarrow K' = \chi^2_{1-\alpha}(1) = \chi^2_{0.95}(1) = 3.84$.
With $n = 55$, $m = 80$, $\bar X = 0.88$ and $\bar X' = 1.07$, $2\ln\Lambda = 1.23 < 3.84 \Rightarrow$ not rejecting H_0.

Uppgift 5

a) Låt θ beteckna sannolikheten att en fabriksprodukt är defekt, där $0 \leq \theta \leq 1$. I ett slumpmässigt urval av 30 produkter visade det sig att 3 var defekta. Anta att sannolikheten att en produkt är defekt är oberoende av de andra produkterna. Använd Jeffrey's prior för att hitta posteriorfördelningen av θ. (5p)

Lösning: Denote X the number of defected products, likelihood function: $X|\theta \sim \mathrm{Bin}(n,\theta)$. Jeffrey's prior $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta) = -E_{X|\theta}[\frac{d}{d\theta}S(\theta)]$ is expected Fisher Information, $S(\theta) = \frac{d}{d\theta}l(\theta)$ is the Score function, $l(\theta) = x\log(\theta) + (n-x)\log(1-\theta)$ is the log-likelihood. So $S(\theta) = \frac{d}{d\theta}l(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$ and $I(\theta) \propto \frac{1}{\theta} + \frac{1}{1-\theta} = \theta^{-1}(1-\theta)^{-1}$, implies Jeffrey's prior $\pi(\theta) = \theta^{-1/2}(1-\theta)^{-1/2} \propto \mathrm{Be}(1/2,1/2)$. Finally posterior distribution of θ is $\theta|x \sim \mathrm{Be}(x+1/2, n-x+1/2)$. With $x = 3$ and $n = 30$, we have $\theta|x \sim \mathrm{Be}(7/2, 55/2)$.

b) I uppgift 5a) var storleken på urvalet givet. Vi använder nu en annan urvalsmetod där vi forsätter ta produkter tills vi har sett 3 defekta. Det visar sig att den 30e produkten vi tar är den tredje och sista defekta. Använd återigen Jeffrey's prior för att hitta posteriorfördelningen av θ. (5p)

Lösning: Likelihood function: $X|\theta \sim \mathrm{NBin}(r,\theta)$. Log-likelihood $l(\theta) = \log[\theta^r(1-\theta)^x] = r\log(\theta) + x\log(1-\theta)$, score function $S(\theta) = \frac{r}{\theta} - \frac{x}{1-\theta}$, expected Fisher Information $I(\theta) \propto \theta^{-2}(1-\theta)^{-1}$ and so Jeffrey's prior $\pi(\theta) = \theta^{-1}(1-\theta)^{-1/2} \propto \mathrm{Be}(0,1/2)$. Finally posterior distribution of θ is $\theta|x \sim \mathrm{Be}(r, x+1/2)$. With $x = 27$ and $r = 3$, we have $\theta|x \sim \mathrm{Be}(3, 55/2)$.

Uppgift 6

Låt $\mathbf{X}_{1:n}$ bestå av n oberoende observationer från en bivariat normalfördelning med väntevärdesvektor $\mu = 0$ och kovariansmatris $\mathbf{\Sigma} = \begin{pmatrix} 4\alpha & \beta \\ \beta & \alpha \end{pmatrix}$, där $\alpha > 0$.

a) Bestäm ML-skattarna $\hat{\alpha}_{\mathrm{ML}}$ och $\hat{\beta}_{\mathrm{ML}}$. (5p)

Lösning: First $\Sigma^{-1} = \frac{1}{4\alpha^2-\beta^2}\begin{pmatrix}\alpha & -\beta \\ -\beta & 4\alpha\end{pmatrix}$. Log-likelihood $l(\alpha,\beta) = -\sum_{i=1}^n \frac{1}{2}\left(\log|\mathbf{\Sigma}| + (x_i,y_i)\mathbf{\Sigma}^{-1}\begin{pmatrix}x_i \\ y_i\end{pmatrix}\right) = -\frac{n}{2}\log(4\alpha^2-\beta^2) - \frac{Q}{4\alpha^2-\beta^2}$ with $Q = \sum_{i=1}^n(\alpha x_i^2 - 2\beta x_i y_i + 4\alpha y_i^2)$. Components of the score vector $\frac{d}{\alpha}l(\alpha,\beta) = \frac{8\alpha Q}{(4\alpha^2-\beta^2)^2} - \frac{1}{4\alpha^2-\beta^2}[4\alpha n + \sum_{i=1}^n(x_i^2+4y_i^2)]$ and $\frac{d}{\beta}l(\alpha,\beta) = \frac{-2\beta Q}{(4\alpha^2-\beta^2)^2} + \frac{1}{4\alpha^2-\beta^2}[\beta n + 2\sum_{i=1}^n x_i y_i])$. Set score vector to zero results in $\frac{8\alpha Q}{4\alpha^2-\beta^2} - 4n\alpha - A = 0$ and $\frac{-2\beta Q}{4\alpha^2-\beta^2} + n\beta + B = 0$ with $A = \sum_{i=1}^n(x_i^2+4y_i^2)$ and $B = 2\sum_{i=1}^n x_i y_i$ independent of α and β. Note that $Q = \alpha A - \beta B$. After some algebra, we have $\hat{\alpha}_{\mathrm{ML}} = \frac{A}{4n}$ and $\hat{\beta}_{\mathrm{ML}} = \frac{B}{n}$.

b) Bestäm profil-likelihood-funktionen $L_p(\beta)$ och estimated-likelihood-funktionen $L_e(\beta)$. (4p)

Lösning: $L_e(\beta) = L(\hat{\alpha}_{\mathrm{ML}},\beta) \propto (\frac{A^2}{4n^2} - \beta^2)^{n/2}\exp\left[-\frac{A^2-4n\beta B}{A^2-4n^2\beta^2}\right]$.
$L_p(\beta) = L(\hat{\alpha}_{\mathrm{ML}}(\beta),\beta)$. $\hat{\alpha}_{\mathrm{ML}}(\beta)$ is solved by $\frac{d}{d\alpha}l(\alpha,\beta) = 0$ with fixed β, implying $16n\hat{\alpha}_{\mathrm{ML}}(\beta)^3 - 4A\hat{\alpha}_{\mathrm{ML}}(\beta)^2 + (8\beta B - 4\beta^2 n)\hat{\alpha}_{\mathrm{ML}}(\beta) - \beta^2 A = 0$ which is a cubic equation in $\hat{\alpha}_{\mathrm{ML}}(\beta)$. Full mark will be given if students arrive this step and explicit expression of $\hat{\alpha}_{\mathrm{ML}}(\beta)$ is not needed.

c) Beskriv, i termer av resultaten från a) och b), hur man kan utföra ett generalized-likelihood-ratio-test av hypotesen H_0: $\beta = 0$ på nivån 5%. (1p)

Lösning: See p.148 of the course book.

Lycka till!