STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik

MT 5003
TENTAMEN
30 maj 2013

Kortfattade lösningar till tentamen i Statistisk inferensteori
30 maj 2013, kl. 9-14

Examinator: Martin Sköld, tel. 16 45 62, mskold@math.su.se.
Tillåtna hjälpmedel: Miniräknare. Formelsamling på tentamens sista sidor.
Återlämning: Meddelas på kurshemsida och via e-post.

Resonemang skall vara tydliga och lätta att följa. Eventuella regularitetsvillkor kan antas vara uppfyllda och behöver ej specificeras närmare. Varje korrekt och fullständigt löst uppgift ger 10 poäng. För betyg A-E krävs 25 poäng på Del 1 sammanräknat med eventuella bonuspoäng, samt att följande gränser uppnås på Del 2:

A	B	C	D	E
25	19	13	7	0

Del 1

Om en stokastisk variabel Z är normalfördelad med väntevärde 0, $N(0,\sigma^2)$, så har dess absolutbelopp $|Z|$ en s.k. Halv-normalfördelning, $HN(\sigma)$ (se formelsamling).

Låt $x = (x_1, \ldots, x_n)$ vara en realisering av $X = (X_1, \ldots, X_n)$, en vektor av n oberoende $HN(\sigma)$-fördelade stokastiska variabler.

Uppgift 1

a) Visa att familjen av fördelningar för X, $\sigma > 0$, utgör en exponentialfamilj.
Lösning: Då X är en vektor oberoende stokastiska variabler ges tätheten av

$$\begin{aligned} p(x;\sigma) &= \prod_{i=1}^{n} \frac{2}{\sqrt{2\pi}\sigma} \exp(-\frac{x_i^2}{2\sigma^2}) = \big(\frac{2}{\sqrt{2\pi}\sigma}\big)^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2) \\ &= A(\sigma)\exp(\zeta(\sigma)T(x)), \end{aligned}$$

med $\zeta(\sigma) = -1/(2\sigma^2)$ och $T(x) = \sum_{i=1}^{n} x_i^2$. En familj som kan skrivas på denna form utgör en Exponentialfamilj.

b) Bestäm moment- (baserat på första momentet $\bar{x}$) och maximum-likelihood skattarna av σ.
Lösning: Momentskattaren $\hat{\sigma}_{MM}$ ges av lösningen till

$$\bar{x} = E_\sigma(X) = \sigma\sqrt{2/\pi} \Rightarrow \hat{\sigma}_{MM} = \bar{x}\sqrt{\pi/2}.$$

Maximum likelihood skattaren $\hat{\sigma}_{ML}$ ges av lösningen till

$$0 = V(\sigma; x) = \frac{d}{d\sigma}\log(p(x;\sigma)) = \frac{d}{d\sigma}(-n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2)$$
$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\sigma}_{ML} = \sqrt{\sum_{i=1}^{n} x_i^2/n}.$$

c) Avgör om skattarna i b) är tillräckliga (sufficient) för σ.
Lösning: Enligt faktoriseringssatsen är T tillräcklig om och endast om $p(x;\sigma)$ kan skrivas som $p(x;\sigma) = A(\sigma)h(x)g(T(x),\sigma)$. I det aktuella fallet gäller att $\exp(-\sum_{i=1}^{n} x_i^2/(2\sigma^2))$ kan bestämmas givet $(\hat{\sigma}_{ML}, \sigma)$ men inte $(\hat{\sigma}_{MM}, \sigma)$ då $\sum x_i^2$ ej kan bestämmas ur $\sum x_i$ (undantaget fallet $n = 1$).

Uppgift 2

En så kallad naturlig parametrisering av fördelningen ges av $\eta = 1/(2\sigma^2)$ med täthet $p(x_i;\eta) = 2\sqrt{\eta/\pi}\exp(-\eta x_i^2)$.

a) Bestäm en konjungerande familj av aprioriförderlningar i den naturliga parametriseringen.
Lösning: Tätheten $p(x,\eta)$ skrivs på exponentialfamiljform som $p(x;\eta) = A(\eta)h(x)\exp(\zeta(\eta)T(x))$ med $\zeta(\eta) = -\eta$ och $A(\eta) = \sqrt{\eta}$. En konjungerande familj ges då av $p(\eta;a,b) \propto A(\eta)^a \exp(\zeta(\eta)b) = \eta^{a/2}\exp(-\eta b)$, vilket känns igen som en $Gamma(a/2+1, b)$ fördelning.

b) Bestäm aposteriorifördelningen $p(\eta|x)$ givet en "improper" aprioriförderlning $p(\eta) \propto 1$, $\eta > 0$.
Lösning: Aposterioriförderlningen ges av

$$p(\eta|x) \propto p(x|\eta)p(\eta) \propto \eta^{n/2}\exp(-\eta\sum_{i=1}^{n} x_i^2),$$

vilket känns igen som en $Gamma(n/2+1, \sum_{i=1}^{n} x_i^2)$-fördelning.

Uppgift 3

a) Visa att momentskattaren är väntevärdesriktig. Antar dess varians Cramér-Raos undre gräns $1/I(\sigma)$?
Lösning: Momentskattaren är väntevärdesriktig ty

$$E(\hat{\sigma}_{MM}) = E(\bar{X})\sqrt{\pi/2} = \sigma\sqrt{2/\pi}\sqrt{\pi/2} = \sigma.$$

Fisher-informationen ges av

$$I(\sigma) = -E(V''(\sigma, X)) = -E(n/\sigma^2 - 3\sum X_i^2/\sigma^4) = 2n/\sigma^2,$$

och då $Var(\hat{\sigma}_{MM}) = (\pi/2)Var(\bar{X}) = \sigma^2(\pi/2-1)/n > \sigma^2/(2n)$ är likhet ej uppfylld.

b) Givet att $\hat{\sigma}_{ML} = 1.2$ och $n = 100$, vilket kan antas stort nog för att asymptotiska resultat skall gälla med god noggrannhet, utför ett likelihood-kvot test av hypotesen $H_0 : \sigma = 1$ mot $H_1 : \sigma \neq 1$ på nivån 5%.
Lösning: Likelihood-kvotstatistikan ges av

$$T(x) = \frac{L(1,x)}{L(\hat{\sigma}_{ML},x)} = \frac{\exp(-\sum_{i=1}^{n} x_i^2/2)}{\exp(-\sum_{i=1}^{n} x_i^2/(2\hat{\sigma}_{ML}^2))/\hat{\sigma}_{ML}^n} = \frac{\exp(-12^2/2)}{\exp(-100/2)/1.2^{100}}.$$

För stora n gäller att $-2\log(T) = -2(-12^2/2+100/2+100\log(1.2)) = 7.53\ldots$ kan betraktas som en dragning från $\chi^2(1)$ under H_0. Då detta är större än 3.84 kan vi förkasta hypotesen.

Del 2

Uppgift 4

Antalet bakterier i en odling dag i kan beskrivas av en stokastisk variabel $X \sim$Poisson(λi). Ett laboratorium förbereder fem odlingar. Varje dag räknar man antalet bakterier i en av skålarna varefter den förstörs. Experimentet utfaller således med ett värde per dag under fem dagar, $x_1, \ldots, x_5$, där x_i kan ses som en realisering av X oberoende av övriga.

a) Bestäm maximum-likelihood skattaren av λ och visa att den har lägsta variansen bland alla väntevärdesriktiga skattare.
Lösning: Loglikelihoodfunktionen ges av

$$l(\lambda; x) = \sum_{i=1}^{5}(-\lambda i + x_i \log(\lambda) + x_i \log(i) + \log(x_i!))$$

och maximeras för $\hat{\lambda}_{ML} = \sum x_i/15$. Den är väntevärdesriktig, ty $E(\hat{\lambda}_{ML}) = \sum \lambda i/15 = \lambda$ och har varians $\sum \lambda i/15^2 = \lambda/15$. Den lägsta möjliga variansen ges vidare av $1/I(\lambda)$ enligt Cramér-Raos olikhet, där

$$I(\lambda) = -E(l''(\lambda; X)) = -E(-\sum(X_i/\lambda^2)) = \sum i/\lambda = 15/\lambda.$$

Eftersom $Var(\hat{\lambda}_{ML}) = \lambda/15 = 1/I(\lambda)$ har den minst varians bland alla väntevärdesriktiga skattare.

b) En alternativ skattare ges av

$$\tilde{\lambda}(x) = \frac{1}{5}\sum_{i=1}^{5} \frac{x_i}{i},$$

bestäm medelkvadratfelet (mean squared error) för denna.
Lösning: Medelkvadratfelet ges av

$$MSE = E((\tilde{\lambda}(X) - \lambda)^2) = Var(\tilde{\lambda}(X)) + Bias(\tilde{\lambda})^2,$$

där $Bias = 0$ då $E(\tilde{\lambda}(X)) = (\sum E(X_i)/i)/5 = \lambda$ och

$Var(\tilde{\lambda}(X)) = \sum Var(X_i)/(25i^2) = \sum \lambda/(25i) = \lambda(1.95 + 1/3)/25$

c) Antag att vi börjar med n odlingar och utför motsvarande experiment över n dagar. Visa att

$$\tilde{\lambda}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{i}$$

är konsistent då $n \to \infty$.
Lösning: Skattaren är konsistent om den konvergerar i sannolikhet mot det sanna värdet, eftersom konvergens i kvadratiskt medel implicerar konvergens i sannolikhet räcker det att visa att MSE konvergerar mot 0. Som ovan har vi

$$MSE = Var(\tilde{\lambda}(X)) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)/i^2 = \frac{\lambda}{n^2}\sum_{i=1}^{n}\frac{1}{i} \leq \frac{\lambda}{n} \to 0$$

då $n \to 0$. Alltså är skattaren konsistent.

Uppgift 5

Antag en modell för linjär regression, där vi för enkelhetens skull låter residualerna vara normalfördelade med varians 1. Vi observerar således par $(y_i, x_i), i = 1, \ldots, n$, där x_i antas fixa och y_i observationer av $Y_i = \alpha + \beta x_i + \epsilon_i$ för oberoende residualer $\epsilon_i \sim N(0,1)$, $i = 1, \ldots, n$.

a) Bestäm scorefunktionen (vektorn) $V(\alpha, \beta; y)$ och Fishers informationsmatris $I(\alpha, \beta)$.
Lösning: Scorevektorn ges av

$$\begin{aligned} V(\alpha,\beta;y) &= (\frac{d}{d\alpha}\log(L(\alpha,\beta;y)), \frac{d}{d\beta}\log(L(\alpha,\beta;y))) \\ &= (\sum(y_i - \alpha - \beta x_i), \sum x_i(y_i - \alpha - \beta x_i)) \end{aligned}$$

och Fishers informationsmatris

$$\begin{aligned} I(\alpha,\beta) = Var(V) &= \begin{pmatrix} nVar(Y) & Cov(\sum Y_i, \sum x_i Y_i) \\ Cov(\sum Y_i, \sum x_i Y_i) & Var(\sum x_i Y_i) \end{pmatrix} \\ &= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \end{aligned}$$

b) Vi vill testa $H_0 : \beta = 0$ med hjälp av ett score-test, d.v.s. teststatistikan $T(y) = V(\hat{\alpha}, 0; y)^T I(\hat{\alpha}, 0)^{-1} V(\hat{\alpha}, 0; y)$. Visa att denna är (exakt) χ^2-fördelad med en frihetsgrad under H_0.
Hjälp med matrisinvertering: För reella tal a, b, c sådana att $ac \neq b^2$ gäller

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} = \frac{1}{ac - b^2}\begin{pmatrix} c & -b \\ -b & a \end{pmatrix}.$$

Lösning: Under $H_0 : \beta = 0$ är $Y_i \sim N(\alpha, 1)$ vilket ger $\hat{a} = \bar{y}$, vidare

$$V(\hat{\alpha}, 0; y) = (\sum(y_i - \bar{y}), \sum x_i(y_i - \bar{y})) = (0, \sum x_i(y_i - \bar{y}))$$

där $W = \sum x_i(Y_i - \bar{Y})$ är normalfördelad med väntevärde 0 och varians $\sum(x_i - \bar{x})^2$. Med beteckningar som ovan för informationsmatrisens invers blir

$$T(Y) = \frac{a}{ac - b^2}(\sum x_i(Y_i - \bar{Y}))^2 = \frac{n}{n\sum x_i^2 - (\sum x_i)^2}W^2$$
$$= \frac{n}{n\sum x_i^2 - (\sum x_i)^2}\sum(x_i - \bar{x})^2 Z^2 = Z^2,$$

där $Z = W/\sqrt{\sum(x_i - \bar{x})^2} \sim N(0,1)$. Alltså är $T(Y) \sim \chi^2(1)$ (givet att $\sum(x_i - \bar{x})^2 > 0$).

Uppgift 6

Antag att vi har en observation x från en likformig fördelning på $[-\theta, \theta]$, $\theta > 0$. Vi vill testa hypotesen $H_0 : \theta = 1$ mot $H_1 : \theta > 1$ och väljer mellan två test; det som förkastar H_0 då $x \in R_1 = \{x; x \geq 0.8\}$ och det som förkastar då $x \in R_2 = \{x; |x| \geq 0.9\}$.

a) Visa att båda testen har signifikansnivå 0.1.
Lösning: Signifikansnivån ges av $P(X \in R|H_0)$. Under H_0 är X likformig på $[-1, 1]$, varför $P(X \in R_1|H_0) = P(X \geq 0.8|H_0) = \int_{0.8}^{1} 1/2dx = 0.1$ och på motsvarande sätt $P(X \in R_2|H_0) = 2P(X \geq 0.9|H_0) = 0.1$.

b) Bestäm styrkefunktionerna för båda testen för $\theta \geq 1$. Vilket är starkast?
Lösning: Styrkefunktionen är $\beta(\theta) = P_\theta(X \in R)$, vi får

$$P_\theta(X \in R_1) = P_\theta(X > 0.8) = \int_{0.8}^{\theta} 1/(2\theta)\, dx = (\theta - 0.8)/(2\theta)$$

för $\theta \geq 1$ och

$$P_\theta(X \in R_2) = 2P_\theta(X > 0.9) = 2\int_{0.9}^{\theta} 1/(2\theta)\, dx = (\theta - 0.9)/\theta$$

$\theta \geq 1$. Eftersom $P_\theta(X \in R_2) \geq P_\theta(X \in R_1)$ är det andra testet starkast.

c) Antag vi observerar $x = -1.2$, bestäm P-värden för båda testen.
Lösning: P-värden ges av $P(X \geq x|H_0) = P(X \geq -1.2|H_0) = 1$ för test 1 och $P(|X| \geq |x||H_0) = P(|X| \geq 1.2|H_0) = 0$.

Lycka till!

Användbara fördelningar

Normalfördelningen $X \sim N(\mu, \sigma^2), \quad -\infty < \mu < \infty, 0 < \sigma < \infty.$
Täthetsfunktion:

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad -\infty < x < \infty.$$

$E(X) = \mu$, $V(X) = \sigma^2$.
Några approximativa kvantiler för $N(0,1)$:
$P(X > 2.58) = 0.005, P(X > 2.33) = 0.01, P(X > 1.96) = 0.025, P(X > 1.64) = 0.05.$

Halv-normalfördelningen $X \sim HN(\sigma), \quad 0 < \sigma < \infty.$
Täthetsfunktion:

$$p(x; \sigma) = \frac{2}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big), \quad x \geq 0.$$

$E(X) = \sigma\sqrt{2/\pi}$, $V(X) = \sigma^2(1 - 2/\pi)$.

Gammafördelningen $X \sim Gamma(\alpha, \beta), \quad \alpha > 0, \beta > 0.$
Täthetsfunktion:

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x \geq 0.$$

$E(X) = \alpha/\beta$, $V(X) = \alpha/\beta^2$.

χ^2-fördelningen $X \sim \chi^2(k), \quad k = 1, 2, 3, \ldots.$
Täthetsfunktion:

$$p(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\Big(-\frac{x}{2}\Big), \quad x \geq 0.$$

$E(X) = k$, $V(X) = 2k$.
Några approximativa kvantiler:

$k = 1;$	$P(X > 3.84) = 0.05$
$k = 10;$	$P(X > 18.3) = 0.05$
$k = 100;$	$P(X > 124) = 0.05$

Binomialfördelningen $X \sim Binomial(n, p), \quad 0 \leq p \leq 1, n = 1, 2, \ldots.$
Sannolikhetsfunktion:

$$p(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

$E(X) = np$, $V(X) = np(1-p)$.