STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik

MT 5003
TENTAMEN
15 mars 2013

Kortfattade lösningar till tentamen i Statistisk inferensteori
15 mars 2013, kl. 9-14

Examinator: Martin Sköld, tel. 16 45 62, mskold@math.su.se.
Tillåtna hjälpmedel: Miniräknare. Formelsamling på tentamens sista sidor.

Resonemang skall vara tydliga och lätta att följa. Eventuella regularitetsvillkor kan antas vara uppfyllda och begöver 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

Rayleigh-fördelningen används bland annat för att modellera vindhastigheter och svarar mot en uppdelning av vinden i två ortogonala riktningskomponenter som är normalfördelade med väntevärde 0 och samma varians.

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

Uppgift 1

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

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

med $A(\theta) = \theta^{-n}$, $h(x) = \prod_{i=1}^{n} x_i$, $\zeta(\theta) = -1/(2\theta)$ och $T(x) = \sum_{i=1}^{n} x_i^2$. En familj som kan skrivas på denna form är en exponentialfamilj.

b) Bestäm moment- (baserat på första momentet $\bar{x}$) och Maximum-Likelihood skattarna av θ.
Lösning: Momentskattaren $\hat{\theta}_{MM}$ ges av lösningen till

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

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

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

c) Avgör om skattarna i b) är tillräckliga (sufficient) för θ.
Lösning: Enligt faktoreringssatsen är T tillräcklig om $p(x;\theta)$ kan skrivas $p(x;\theta) = A(\theta)h(x)g(\theta, T(x))$. Med A och h som i a), får vi $g(\theta, T(x)) = \exp(-\sum x_i^2/(2\theta))$. Detta är en funktion av $\hat{\theta}_{ML}$, men ej av $\hat{\theta}_{MM}$ eftersom $\sum x_i^2$ ej kan bestämmas ur $\bar{x}^2$ (undantaget fallet $n = 1$). Därför är endast $\hat{\theta}_{ML}$ tillräcklig (båda i fallet $n = 1$).

Uppgift 2

a) Bestäm Fisherinformationen $I(\theta)$ i X.
Lösning: Då X består av oberoende likafördelade variabler är $I(\theta) = nI_{X_1}(\theta)$, där

$$I_{X_1}(\theta) = -E_\theta(\frac{d^2}{d\theta^2}\log(p(X_1;\theta))) = -\frac{1}{\theta^2} + \frac{1}{\theta^3}E_\theta(X_1^2)$$
$$= -\frac{1}{\theta^2} + \frac{1}{\theta^3}((4-\pi)\theta/2 + \theta\pi/2) = \frac{1}{\theta^2}.$$

b) Bestäm aprioriförde lningen för θ enligt Jeffreys princip (Jeffrey's prior) och visa att motsvarande aposteriorifördelning är en invers Gammafördelning.
Lösning: Jeffrey's prior ges av $p(\theta) \propto \sqrt{I(\theta)} \propto 1/\theta$ enligt ovan. Aposteriorifördelningen ges av

$$p(\theta|x) \propto p(x;\theta) \times p(\theta) \propto \frac{1}{\theta^n}\exp\Big(\frac{-1}{2\theta}\sum_{i=1}^{n} x_i^2\Big) \times \frac{1}{\theta}$$
$$\propto \theta^{-\alpha-1}\exp\Big(-\frac{\beta}{\theta}\Big)$$

d.v.s. en invers gammafördelning med $\alpha = n$ och $\beta = \sum_{i=1}^{n} x_i^2/2$.

Uppgift 3

a) Bestäm den asymptotiska fördelningen för ML-skattaren.
Lösning: Under regularitetsvillkor ges denna av $N(\theta, 1/I(\theta)) = N(\theta, \theta^2/n)$.

b) Antag att $n = 100$, vilket kan anses tillräckligt stort för att asymptotiska resultat gäller med god noggrannhet. Testa hypotesen $H_0 : \theta = 1$ mot $H_1 : \theta \neq 1$ på nivån 5% med lämpligt (valfritt) asymptotiskt test givet att $\hat{\theta} = 1.3$.
Lösning: Här finns flera alternativ, enklast är kanske ett Wald-test som bygger på den asymptotiska fördelningen i a). Test-statistikan ges av

$$T_W(x) = (\hat{\theta} - 1)^2 I(\hat{\theta}) = 0.3^2 \times 100/1.3^2 \approx 5.3$$

och har en asymptotisk $\chi^2(1)$-fördelning under noll-hypotesen. Om vi förkastar då T_W är större än 3.84 får vi ett test på nivån 5%.

c) Bestäm den asymptotiska fördelningen för moment-skattaren.
Lösning: Enligt CGS är $\sqrt{n}(\bar{x}-\mu)$ asymptotiskt $N(0,\sigma^2)$ där $\mu = \sqrt{\theta\pi/2}$ och $\sigma^2 = (4-\pi)\theta/2$. Enligt Delta-metoden är då den asymptotiska fördelningen för $\sqrt{n}(g(\bar{x}) - g(\mu))$ $N(0, \sigma^2 g'(\mu)^2)$. Med $g(\mu) = 2\mu^2/\pi$ får vi att $\hat{\theta}_{MM}$ är asymptotiskt $N(g(\mu), \sigma^2\mu^2(4/\pi)^2/n) = N(\theta, 4\theta^2(4-\pi)/(n\pi))$.

Del 2

Uppgift 4

Livslängden i timmar hos en typ av elektroniska komponenter kan antas beskrivas av en stokastisk variabel X. Komponenterna inspekteras dagligen och byts ut då de är trasiga. För att skatta $p = P(X > 24)$, sannolikheten att en nyinstallerad komponent överlever till nästa inspektion, mäts livslängden hos 50 oberoende komponenter. Dessa livslängder betecknas $x_1, \ldots, x_{50}$.

a) Låt $\hat{p}$ beteckna andelen observerade livslängder som är mindre än 24, d.v.s.

$$\hat{p} = \frac{1}{50}\sum_{i=1}^{50} \mathbf{1}\{x_i > 24\}, \text{ där } \mathbf{1}\{x > 24\} = \begin{cases} 1 & \text{om } x > 24 \\ 0 & \text{om } x \leq 24. \end{cases}$$

Bestäm medelkvadratfelet (mean squared error) för skattaren $\hat{p}$.
Lösning: $50\hat{p}$ är $Binomial(50, p)$-fördelad med väntevärde $50p$ och varians $50p(1-p)$. Följer att $\hat{p}$ har väntevärde p (väntevärdesriktig) och varians $p(1-p)/50$. Eftersom $\hat{p}$ är väntevärdesriktig ges medelkvadratfelet av variansen, $MSE(p) = p(1-p)/50$.

b) Antag nu att X är $Exponential(\theta)$-fördelad och bestäm maximum-likelihood skattaren $\tilde{p}$ av $p = P_\theta(X > 24)$.
Lösning: Maximum-likelihood skattaren av p ges av $\tilde{p} = P_{\hat{\theta}}(X > 24) = \exp(-24/\hat{\theta})$, där $\hat{\theta}$ är ML skattaren av θ vilken ges av lösningen till

$$0 = V(\theta; x) = -\frac{50}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{50} x_i \Rightarrow \hat{\theta} = \bar{x}.$$

Vi får alltså ML-skattaren $\tilde{p} = \exp(-24/\bar{x})$.

c) Jämför de båda skattarnas varianser då $\theta = 24$ givet att stickprovsstorleken är stor och att antagandet om exponentialfördelning är korrekt.
Lösning: Givet att exponentialfördelningsantagandet är korrekt är $Var(\hat{p}) = \exp(-24/\theta)(1 - \exp(-24/\theta))/n$. Den asymptotiska variansen för $\tilde{p}$ fås genom Deltametoden, vi har att $Var(\hat{\theta}) = \theta^2/n$ och $Var(\tilde{p}) \approx Var(\hat{\theta})g'(\theta)^2$ där $g(\theta) = \exp(-24/\theta)$. Insatt ger detta $Var(\tilde{p}) \approx \theta^2(24/\theta^2)^2\exp(-48/\theta)/n$. För $\theta = 24$ och $n = 50$ får vi $Var(\hat{p}) = (e^{-1} - e^{-2})/50 \approx 0.0047$ och $Var(\tilde{p}) = e^{-2}/n \approx 0.0027$.

Uppgift 5

Ett mer flexibelt alternativ till att modellera livslängder utgörs av Weibullfördelningen. Fördelningen bär namn efter Wallodi Weibull som var professor i teknisk fysik på KTH.

a) För vilka av följande tre situationer, där θ betecknar den okända parametern, utgör familjen av Weibull-fördelningar en exponentialfamilj: $\theta = (\lambda, k)$, $\theta = \lambda$ (k känd/fix), $\theta = k$ (λ känd/fix)?
Lösning: Familjen utgör en exponentialfamilj om

$$p(x;\theta) = A(\theta)h(x)\exp(\sum_{i=1}^{m}\zeta_i(\theta)T_i(x)).$$

Tätheten ges av

$$p(x;\theta) = \frac{k}{\lambda}\Big(\frac{x}{\lambda}\Big)^{k-1}\exp\Big(-\Big(\frac{x}{\lambda}\Big)^{k}\Big)$$

och innehåller faktorn $\exp(-(x/\lambda)^k)$. Här kan $(x/\lambda)^k$ ej faktoriseras som $\zeta(\theta)T(x)$ då k är del av θ. I dessa två fall utgör familjen inte en exponentialfamilj. Då k är känd och $\theta = \lambda$ kan vi välja $A(\theta) = (k/\lambda^k)$, $h(x) = x^{k-1}$, $\zeta(\theta) = -1/\lambda^k$ och $T(x) = x^k$, d.v.s. i detta fall har vi en exponentialfamilj.

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

b) Bestäm en tillräcklig (sufficient) och fullständig (complete) statistika för λ i fallet där k känt.
Lösning: I detta fallet utgör $p(x_1, \ldots, x_n; \theta)$ en exponentialfamilj med $A_n(\theta) = A(\theta)^n$, $h_n(x) = \prod_{i=1}^n x_i^k$, $\zeta_n(\theta) = \zeta(\theta)$ och $T_n(x) = \sum_{i=1}^n T(x_i) = \sum_{i=1}^n x_i^k$ (A, h, ζ och T som i a)). Följer av sats (Theorem 4.6) att $T_n(x)$ är tillräcklig och fullständig.

c) Beskriv hur man kan konstruera ett konfidensintervall för λ med approximativ täckningsgrad 95% i fallet där båda parametrarna är okända. Det finns inget explicit uttryck för maximum-likelihood skattaren av $\theta = (\lambda, k)$, men du kan anta att du har en numerisk lösning $\hat{\theta} = (\hat{\lambda}, \hat{k})$ till score-ekvationen $V(\theta, x) = 0$. Stickprovsstorleken kan antas vara stor.
Lösning: Vi vet att $(\hat{\lambda}, \hat{k})$ asymptotiskt är fördelad som $N(\theta, I(\theta)^{-1})$. Detta betyder att $\hat{\lambda}$ är asymptotiskt fördelad som $N(\lambda, \sigma^2)$ där σ^2 utgörs av elementet i position $(1,1)$ av matrisen $I(\theta)^{-1}$. Om vi kan skatta detta element ges ett konfidensintervall av $(\hat{\lambda} - 1.96\hat{\sigma}, \hat{\lambda} + 1.96\hat{\sigma})$. Här kan man tänka sig att $\hat{\sigma}^2$ ges av motsvarande element i $I(\hat{\theta})^{-1}$, men Fisherinformationsmatrisen är inte helt lätt att bestämma så ett enklare altenativ är att använda motsvarande element i observerade informationsmatrisen $J(\hat{\theta}, x)^{-1}$.

Uppgift 6

Genotypfrekvenserna av (`AA`,`Aa`,`aa`) i en population fördelar sig som $p = (p_1, p_2, p_3)$, där $(p_1, p_2) \in [0,1]^2$ och $p_3 = 1 - p_1 - p_2$. Då en population befinner sig i genetisk

jämvikt, gäller enligt Hardy-Weinbergs lag att $p = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$ för något $\theta \in [0,1]$.

I en stor population undersöks 1000 slumpmässigt valda individer med resultatet att antalen (AA,Aa,aa) fördelar sig som (x_1, x_2, x_3). Detta kan ses som en observation av $(X_1, X_2, X_3) \sim Multinomial(1000, p_1, p_2, p_3)$ med (p_1, p_2, p_3) som ovan. För att avgöra om populationen befinner sig i genetisk jämvikt vill vi testa $H_0 : (p_1, p_2) \in \Theta_0$, där $\Theta_0 = \{(\theta^2, 2\theta(1-\theta)),\ 0 < \theta < 1\}$ mot $H_1 : (p_1, p_2) \in [0,1]^2, (p_1, p_2) \notin \Theta_0$.

a) Bestäm likelihood-kvot statistikan för att testa H_0 mot H_1.
Lösning: Likelihood-kvot statistikan ges av

$$T_{LR} = \frac{\sup_{(p_1,p_2)\in\Theta_0} p(x;p)}{\sup_{(p_1,p_2)\in[0,1]^2} p(x;p)} = \frac{\sup_{\theta\in[0,1]} (\theta^2)^{x_1}(2\theta(1-\theta))^{x_2}((1-\theta)^2)^{x_3}}{\sup_{(p_1,p_2)\in[0,1]^2} p_1^{x_1} p_1^{x_2}(1-p_1-p_2)^{x_3}}.$$

ML-skattaren under H_0 ges av lösningen $\hat{\theta}_{ML} = (2x_1 + x_2)/(2n)$ till

$$0 = V(\theta, x) = \frac{2x_1}{\theta} + \frac{x_2}{\theta} - \frac{x_2}{1-\theta} - \frac{2x_3}{1-\theta} = \frac{2x_1 + x_2}{\theta} - \frac{x_2 + 2x_3}{1-\theta}.$$

ML-skattaren under H_1 ges på liknande sätt av $p = (x_1/n, x_2/n, x_3/n)$. Insatt ovan ger detta

$$T_{LR} = \frac{2^{x_2}((2x_1 + x_2)/(2n))^{2x_1+x_2}(1 - (2x_1 + x_2)/(2n))^{2x_3+x_2}}{(x_1/n)^{x_1}(x_2/n)^{x_2}(x_3/n)^{x_3}}.$$

b) Antag vi observerar $(x_1, x_2, x_3) = (430, 334, 236)$. Testa hypotesen på nivån 5% med ett (asymptotiskt) likelihood-kvot test.
Lösning: För stora n är $-2\log(T_{LR})$ approximativt $\chi^2(k)$, där k är antalet fria parametrar i H_1 minus antalet i H_0, d.v.s. $k = 2 - 1 = 1$ i det aktuella fallet. Insättning ger $-2\log(T_{LR}) \approx 94 > 3.84$. Vi kan alltså förkasta med goda marginaler.

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 $N(0,1)$:
$P(X > 2.58) = 0.005, P(X > 2.33) = 0.01, P(X > 1.96) = 0.025, P(X > 1.64) = 0.05$.

Rayleighfördelningen $X \sim Rayleigh(\theta)$, $\quad 0 < \theta < \infty$.
Täthetsfunktion:

$$p(x;\theta) = \frac{x}{\theta} \exp\Big(-\frac{x^2}{2\theta}\Big), \quad 0 < x < \infty.$$

$E(X) = \sqrt{\theta\pi/2}$, $V(X) = (4-\pi)\theta/2$.

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

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

$E(X) = \beta/(\alpha - 1)$ då $(\alpha > 1)$, $V(X) = \beta^2/((\alpha-1)^2(\alpha-2))$ då $\alpha > 2$.

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:

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

Exponentialfördelningen $X \sim Exponential(\theta), \quad \theta > 0.$
Täthetsfunktion:

$$p(x; \theta) = \frac{1}{\theta} \exp\Big(-\frac{x}{\theta}\Big), \quad x \geq 0.$$

$E(X) = \theta$, $V(X) = \theta^2$.

Weibullfördelningen $X \sim Weibull(\lambda, k), \quad \lambda > 0, k > 0.$
Täthetsfunktion:

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

$E(X) = \lambda\Gamma(1 + 1/k)$, $V(X) = \lambda^2(\Gamma(1 + 2/k) - \Gamma(1 + 1/k)^2)$.

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

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

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

Multinomialfördelningen $(X_1, \ldots, X_k) \sim Multinomial(n, p_1, \ldots, p_k)$,
$0 \leq p_i \leq 1, i = 1, \ldots, k, \sum_{i=1}^k p_i = 1, n = 1, 2, \ldots$.
Sannolikhetsfunktion:

$$p(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

$x_i = 0, 1, \ldots, n, i = 1, \ldots, k, \sum_{i=1}^k x_i = n.$

$E(X_i) = np_i$, $V(X_i) = np_i(1 - p_i), i = 1, \ldots, k.$