Tentamentsskrivning i **Matematisk statistik TMA321, 4.5 hp.**

Tid: Fredagen den 29 augusti, 2008 kl 08:30-12:30
Examinator och jour: Erik Broman, tel. 772-3514, mob. 073 7320791, MV-huset rum L3080.
Hjälpmedel: valfri räknare, egenhändigt skriven formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

———————————————

Tentamen består av 8 frågor om sammanlagt 30 poäng. Preliminära betygsgränser är satta till:
betyg "3": 12 till 17 poäng
betyg "4": 18 till 23 poäng
betyg "5": 24 eller fler poäng.

———————————————

OBS! Tydligt skrivna och utörliga svar krävs för poäng.

1. (4 poäng) Slumpvariabeln X är Poisson fördelad med parameter $\lambda > 0$. Antag att oberoende samplingar $X_1, \ldots, X_n$ tas från denna fördelning.

a. Bestäm MME (method of moments estimate) skattaren för λ.

b. Bestäm MLE (maximum likelihood estimate) skattaren för λ.

Lösning
a. Vi har att

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} = \lambda.$$

Vi får att vår MME blir $\hat{\lambda} = \bar{X} = (X_1 + \cdots + X_n)/n$.

b. Vi har att

$$f(X_1, \cdots, X_n|\lambda) = \frac{\lambda^{X_1+\cdots+X_n}e^{-n\lambda}}{\prod_{i=1}^{n} X_i!}.$$

Vi logaritmerar och får

$$l(\lambda) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X_i!.$$

Om vi deriverar detta och sätter resultatet till noll fås:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0,$$

och vi får (igen) att $\hat{\lambda} = \bar{X} = (X_1 + \cdots + X_n)/n$.

2. (5 poäng) En nyöppnad speditionsfirma skall transportera paket från Stockholm till Göteborg. Firman väger de 20 första paketen dom får i uppdrag att transportera. Resultatet blev:

paket:	1	2	3	4	5	6	7	8	9	10
vikt i kilo:	16.5	6.7	21	22.3	10.8	29.5	29.5	19.7	22.6	21.4
paket:	11	12	13	14	15	16	17	18	19	20
vikt i kilo:	18.5	25.8	15.3	37.5	18.9	20.9	28.5	20.5	19.2	13.3

Speditionsfirman använder sig av hypotesen att vikterna är normalfördelade.

a. Skatta medelvikten och variansen.

b. Ange exakta 95% konfidensintervall för medelvikten.

Speditionsfirman bestämmer sig för att ta emot N st paket innan de stuvas in i ett fordon som kan lastas med maximalt 2 ton.

c. Hur skall de välja N så att de N är det maximala antalet paket som inte väger mer än 2 ton tillsammans med 99% sannolikhet?

Lösning:

a. Låt X_i beteckna vikten av paket nr i. Medelvikten är 20.92 kg. Variansen skattas med hjälp av formeln

$$\hat{s}^2 = \frac{1}{19}\sum_{i=1}^{20}(X_i - \bar{X})^2 \approx 49.3$$

b. Vi använder oss av t-fördelning med 19 frihetsgrader. Det relevanta värdet läses ur tabell att vara 2.093. Därför blir ett 95 % CI

$$\bar{X} \pm s_{\bar{X}} 2.093 = 20.92 \pm 2.093 * 1.57 = 20.92 \pm 3.286.$$

Här är $s_{\bar{X}} = \hat{s}/\sqrt{20}$.

c. Vi skall bestämma N så att $X = X_1 + \cdots + X_N \leq 2000$ med sannolikhet 0.99. Vi använder normalapproximation och får att $X \sim N(N\hat{\mu}, N\hat{s}^2)$. Därför är

$$\frac{X - N\hat{\mu}}{\sqrt{N}\hat{s}} \sim N(0,1).$$

Ur tabell fås att $0.99 = \mathbb{P}(Z \leq 2.33)$, dvs vi måste välja N så att

$$\frac{2000 - N\hat{\mu}}{\sqrt{N}\hat{s}} \leq 2.33,$$

dvs

$$2000 \leq N\hat{\mu} + 2.33\sqrt{N}\hat{s}.$$

Sätt $N = M^2$ och lös

$$M^2 + M\frac{\hat{s} * 2.33}{\hat{\mu}} - \frac{2000}{\hat{\mu}} = 0.$$

Vi får att

$$M = -\frac{\hat{s} * 2.33}{2\hat{\mu}} + \sqrt{\left(\frac{\hat{s} * 2.33}{2\hat{\mu}}\right)^2 + \frac{2000}{\hat{\mu}}} \approx 9.39$$

Det vill säga att $N \approx 88.26$, dvs vi väljer $N = 88$.

3. (3 poäng) X är en slumpvariabel med pmf (sannolikhetsfördelning):

$$p_k = \mathbb{P}(X = k) = (1 - x^2)x^{2k} \quad \text{för } k \in \{0, 1, 2, \ldots\}.$$

a. För vilka värden på x är detta verkligen en sannolikhetsfördelning?

b. Beräkna väntevärdet av X för värdena på x ur uppgift a.

Lösning:
a. Vi har att

$$\sum_{k=0}^{\infty} p_k = (1 - x^2)\sum_{k=0}^{\infty} x^{2k} = (1 - x^2)\frac{1}{1 - x^2},$$

om $x^2 < 1$, dvs om $-1 < x < 1$. Dessutom ser vi att p_k är positivt för dessa val av x. Möjligtvis kan $x = 0$ anses vara ett specialfall även om konventionen dikterar att i det fallet är $p_0 = 1$.

b.Vi har att

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k(1 - x^2)x^{2k} = \frac{(1 - x^2)}{2}\sum_{k=0}^{\infty} 2kx^{2k} \\
&= \frac{x(1 - x^2)}{2}\sum_{k=0}^{\infty} 2kx^{2k-1} = \frac{x(1 - x^2)}{2}\frac{d}{dx}\sum_{k=0}^{\infty} x^{2k} \\
&= \frac{x(1 - x^2)}{2}\frac{d}{dx}\frac{1}{1 - x^2} = \frac{x(1 - x^2)}{2}(-2x)\frac{-1}{(1 - x^2)^2} = \frac{x^2}{1 - x^2}.
\end{aligned}$$

4. (3 poäng) Följande data observerades vid ett experiment:

x :	0.9	−4.1	4.9	−6.2	3.7	−6.3	−2.6	2.5	5.6	−8.4
y :	20.2	8.6	18.1	−1.9	15.3	−4.6	5.7	14.5	25.9	1.7

a. Plotta data i en figur.

b. Använd minsta kvadratmetoden för att bestämma regressionslinjen $y = b_0 + b_1 x$. Rita även in denna i din figur.

c. Använd minsta kvadratmetoden för att bestämma regressionslinjen $x = c_0 + c_1 y$. Rita även in denna i din figur.

d. Är linjerna i uppgift b. och c. dom samma. Varför/ Varför inte?

Lösning:
a. Plotten ses i separat figur ihop med regressionslinjerna.

b. Vi använder oss av:

$$b_1 = \frac{10\sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10\sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2}$$

och

$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Vi får att $b_1 \approx 1.78$ och $b_0 \approx 12.13$, se figur.

c. Vi använder oss av:

$$c_1 = \frac{10\sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10\sum_{i=1}^{10} y_i^2 - (\sum_{i=1}^{10} y_i)^2}$$

och

$$c_0 = \bar{X} - c_1 \bar{Y}.$$

Vi får att $c_1 \approx -5.81$ och $c_0 \approx 0.46$, se figur. OBS! Man plottar förstås $y = -c_0/c_1 + x/c_1$.

d. Nej dom är ej samma. Anledningen är att i b. minimeras felet i y-led medan i c. minimeras felet i x-led.

5. (4 poäng) En vanlig kortlek består av 52 kort uppdelade i fyra färger med 13 kort i vardera färgen. En bridgehand består av 13 kort.

a. Vad är sannolikheten att en bridgehand innehåller exakt fyra spader?

b. Vad är sannolikheten att en bridgehand innehåller exakt fyra spader och tre hjärter?

En pokerhand består av fem kort.

c. Hur många pokerhänder finns det där alla fyra färgerna ingår?

d. Herr Lööf spelar poker och får tre spader och två hjärter på given. Han slänger de två hjärterna och får två nya kort. Vad är sannolikheten att bägge dessa är spader?

Lösning: a. Antalet bridgehänder totalt är $\binom{52}{13}$. Vi kan välja 4 spader utav 13 totalt på $\binom{13}{4}$ och resterande 9 kort på $\binom{39}{9}$ sätt. Alltså får vi att sannolikheten blir

$$\frac{\binom{13}{4}\binom{39}{9}}{\binom{52}{13}} \approx 0.2386.$$

b. Igen kan vi välja 4 spader utav 13 totalt på $\binom{13}{4}$ sätt, dessutom kan vi välja 3 hjärter utav 13 totalt på $\binom{13}{3}$ sätt, och resterande 6 kort på $\binom{26}{6}$ sätt. Alltså får vi att sannolikheten blir

$$\frac{\binom{13}{4}\binom{13}{3}\binom{26}{6}}{\binom{52}{13}} \approx 0.0741.$$

c. Antalet händer med fyra kort som innehåller alla fyra olika färger är 13^4. För att få en hand med fem kort så kan vi välja det femte på vilket sätt vi vill. Dock inses att på detta sätt kommer varje hand att räknas två gånger. Det totala antalet händer blir därför $13^4 * 48/2$.

d. Sannolikheten att det första kortet är spader är $9/47$ och givet denna händelse är sannolikheten att även det andra är spader $8/46$. Den totala sannolikheten blir alltså

$$\frac{9 * 8}{47 * 46} \approx 0.0333.$$

6. (3 poäng) Låt X, Y vara likformigt fördelade på $[0,2]$ respektive $[1,3]$ och låt dom även vara oberoende.

a. Ange den gemensamma pdf:en för (X, Y)

Låt $Z = XY$ och $W = X + Y$.

b. Beräkna kovariansen av Z och W.

Lösning:

a.Den gemensamma pdf:en ges av

$$f_{X,Y}(x,y) = \begin{cases} 1/4 & \text{om } (x,y) \in [0,2] \times [1,3] \\ 0 & \text{annars} \end{cases}$$

b. Kovariansen ges av $Cov(Z, W) = \mathbb{E}[(Z - \mathbb{E}[Z])(W - \mathbb{E}[W])]$. Vi har att $\mathbb{E}[Z] = \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 1 * 2 = 2$. Vidare att $\mathbb{E}[W] = \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 1 + 2 = 3$. Därför får vi:

$$\begin{aligned}
\mathbb{E}[(Z - \mathbb{E}[Z])(W - \mathbb{E}[W])] &= \mathbb{E}[(XY - 2)(X + Y - 3)] \\
&= \mathbb{E}[X^2Y + XY^2 - 3XY - 2X - 2Y + 6] \\
&= \mathbb{E}[X^2]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y^2] - 3\mathbb{E}[X]\mathbb{E}[Y] - 2\mathbb{E}[X] - 2\mathbb{E}[Y] + 6 \\
&= \mathbb{E}[X^2] * 2 + \mathbb{E}[Y^2] - 3 * 2 - 2 - 2 * 2 + 6 = \mathbb{E}[X^2] * 2 + \mathbb{E}[Y^2] - 6.
\end{aligned}$$

Dessutom gäller att

$$\mathbb{E}[X^2] = \frac{1}{2}\int_0^2 x^2 dx = \frac{1}{2}\left[\frac{x^3}{3}\right]_0^2 = \frac{4}{3},$$

och att

$$\mathbb{E}[Y^2] = \frac{1}{2}\int_1^3 y^2 dx = \frac{1}{2}\left[\frac{y^3}{3}\right]_1^3 = \frac{13}{3},$$

så att

$$Cov(Z, W) = \frac{8}{3} + \frac{13}{3} - \frac{18}{3} = 1.$$

7. (4 poäng) Kor på bete misstänktes ta upp bly från en närliggande förorenad soptipp genom gräset dom åt. Därför mättes halten av bly i prov tagna från korna. Mätningar utfördes på våren innan dom hade gått ut på bete och på hösten. Resultatet blev som följer (där andra och tredje raden anger mängden bly i proven på våren respektive hösten):

Ko nr:	1	2	3	4	5	6	7	8	9	10
våren:	1.48	1.07	1.01	1.05	0.46	1.08	1.26	1.36	1.92	1.03
hösten:	2.05	1.72	1.73	0.81	0.82	1.46	1.01	1.24	2.71	1.21

a. Beräkna samplingsmedlet och samplingsvariansen av skillnaderna. Förklara vad den senare mäter.

b. Vi antar att skillnaderna är normalfördelade. Låt nollhypotesen vara att $\mu = 0$. Vad är p-värdet vid ett dubbelsidigt respektive enkelsidigt ($\mu > 0$) test?

Lösning:

a. Låt D_i vara skillnaden i mätningarna för Ko nr i. Vi får att $\bar{D} = \sum_{i=1}^{10} D_i/10 \approx 0.304$. Samplingsvariansen s^2 ges av

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(D_i - 0.304)^2 \approx 0.1563,$$

denna ger ett mått på hur mycket skillnaderna varierar runt sitt medel.

b. p-värdet ges av den minsta signifikansnivån på vilken vi skulle förkasta H_0. Under antagandet att skillnaderna är normalfördelade så är

$$T = \frac{\bar{D}}{s/\sqrt{10}} \approx 2.43,$$

t-fördelad med 9 frihetsgrader. Perfekt utläsning av tabell är ej möjlig, men vi ser att det två -sidiga p-värdet ges av ungefär

$$\mathbb{P}(T \leq -2.43 \cup T \geq 2.43) \approx 0.04,$$

medan det ensidiga p-värdet blir exakt hälften så stort.

8. (4 poäng) En ny metod för att diagnostisera en viss vanlig form av cancer har utvecklats. Enligt de som utvecklat metoden så är sannolikheten att en patient med cancer får rätt diagnos 90% medan sannolikheten att en friskt patient får rätt diagnos 95%. I en population vet man att sannolikheten att ha denna typ av cancer är 1 %.

a. Per får diagnosen cancer, vad är sannolikheten att han faktiskt har cancer?

b. Sara får diagnosen "ej cancer", vad är sannolikheten att hon faktiskt har cancer?

c. Vad kan vi dra för slutsatser ur svaren från a. och b. om detta test?

Lösning: a. Låt D vara en slumpvariabel som är 0 om diagnosen är "ej cancer" och 1 om diagnosen är "cancer". Låt C vara händelsen att Per verkligen har cancer. Vi vill beräkna $\mathbb{P}(C|D = 1)$. Först observerar vi att

$$\mathbb{P}(D = 1) = \mathbb{P}(D = 1|C)\mathbb{P}(C) + \mathbb{P}(D = 1|C^c)\mathbb{P}(C^c) = 0.9*0.01 + 0.05*0.99 = 0.0585$$

Därför får vi att,

$$\mathbb{P}(C|D = 1) = \frac{\mathbb{P}(C \cap D = 1)}{\mathbb{P}(D = 1)} = \frac{\mathbb{P}(D = 1|C)\mathbb{P}(C)}{\mathbb{P}(D = 1)} = \frac{0.9 * 0.01}{0.0585} \approx 0.154,$$

dvs sannolikheten att Per faktiskt har cancer är cirka 15.4%.

b. Vi vill nu beräkna $\mathbb{P}(C|D = 0)$. Vi får att $\mathbb{P}(D = 0) = 1 - \mathbb{P}(D = 1) = 0.9415$. Därför blir

$$\mathbb{P}(C|D = 0) = \frac{\mathbb{P}(C \cap D = 0)}{\mathbb{P}(D = 0)} = \frac{\mathbb{P}(D = 0|C)\mathbb{P}(C)}{\mathbb{P}(D = 0)} = \frac{0.1 * 0.01}{0.9415} \approx 0.00106,$$

dvs sannolikheten att Sara verkligen har cancer är cirka 0.1 %.

c. Blir diagnosen "ej cancer" så kan man känna sig ganska så säker. Får man diagnosen cancer skall man vara "försiktigt optimistisk". Uppenbarligen skall man då gå vidare med andra tester.

Tabeller:
1. Normalfördelning
2. t-fördelning

Lycka till!