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

Tid: Måndagen den 11 januari, 2010 kl 08:30-12:30
Examinator och jour: Erik Broman, tel. 772-3514,
Hjälpmedel: valfri räknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

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

OBS! För full poäng krävs fullständiga och noggranna lösningar.

1. (4 poäng) Betrakta följande pdf (sannolikhetsfördelning):

$$f(x) = C\cos(x)e^{-\sin(x)}, \quad \text{för } x \in [0, \pi/2]$$

för slumpvariabeln X.

a. Hitta C så att detta blir en pdf (sannolikhetsfördelning).

b. Låt $Y = \sin(X)$, hitta pdf:en for Y.

c. Beräkna $\mathbb{E}[\sin(X)]$.

d. Beräkna $\mathrm{Var}(\sin(X))$.

Lösning: a. Vi har att

$$\int_0^{\pi/2} C\cos(x)e^{-\sin(x)}dx = C\int_0^1 e^{-y}dy = C[-e^{-y}]_0^1 = C(1-e^{-1}),$$

så att $C = 1/(1-e^{-1})$.

b.Vi har att

$$F(z) = \mathbb{P}(X \leq z) = \int_0^z C\cos(x)e^{-\sin(x)}dx = C\int_0^{\sin(z)} e^{-y}dy = C(1-e^{-\sin(z)}),$$

därför blir

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sin(X) \leq y) = \mathbb{P}(X \leq \arcsin(y)) = C(1-e^{-\sin(\arcsin(y))}) = C(1-e^{-y}).$$

Genom att derivera kan vi då bestämma pdf:en för Y till

$$f_Y(y) = Ce^{-y}, \quad \text{för } 0 \leq y \leq 1.$$

c.

$$\mathbb{E}[\sin(X)] = \mathbb{E}[Y] = C\int_0^1 ye^{-y}dy = C\left([-ye^{-y}]_0^1 + \int_0^1 e^{-y}dy\right) = C(-e^{-1}+C^{-1}).$$

d. $\text{Var}(\sin(X)) = \text{Var}(Y)$ och

$$\begin{aligned}\mathbb{E}[Y^2] &= C\int_0^1 y^2e^{-y}dy = C\left([-y^2e^{-y}]_0^1 + \int_0^1 2ye^{-y}dy\right)\\ &= C\left(-e^{-1} + [-2ye^{-y}]_0^1 + 2\int_0^1 e^{-y}dy\right) = C(-3e^{-1} + 2C^{-1}).\end{aligned}$$

Ur detta kan vi skriva variansen som

$$\text{Var}(Y) = C(-3e^{-1} + 2C^{-1}) - C^2(-e^{-1} + C^{-1})^2.$$

2. (5 poäng) På snuslagret i Göteborg försvinner det ovanligt många dosor. Ledningen misstänker stöld. I hela landet så har man sett att det i genomsnitt försvinner 207 dosor snus varje vecka. Detta tillskrivs felaktig hantering, kassering av undermåliga produkter osv. Innan man går till polisen med en formell anmälan så vill man göra ett test för att se om hypotesen är rimlig. Man räknade därför antalet försvunna dosor under tolv veckor och fick följande resultat.

vecka:	1	2	3	4	5	6	7	8	9	10	11	12
dosor:	242	242	203	222	196	198	220	251	205	227	215	242

a. Sätt upp ett hypotestest för att testa ledningens misstankar.

b. Låt signifikansnivån $\alpha = 0.01$. Använd normalapproximation för att testa hypoteserna du satte upp i uppgift a. Skall du förkasta nollhypotesen på den givna signifikansnivån?

c. Låt signifikansnivån igen vara $\alpha = 0.01$. Använd den exakta nollfördelningen för att testa hypoteserna du satte upp i uppgift a. Skall du förkasta nollhypotesen på den givna signifikansnivån?

d. Är det någon skillnad mellan b. och c. Är det viktigt vilken metod som används? Vilket är p-värdet av testen i b. och c.? Motivera.

2. Lösning:

a. Använd $H_0 : \mu = 207$ och $H_1 : \mu > 207$.

b. Vi använder att under H_0 och med normalapproximation är

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1).$$

Vidare blir $\bar{X} \approx 222$ och $s \approx 19.1$. Därför blir

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx 2.7$$

Tabell ger

$$\mathbb{P}(\frac{\bar{X}-\mu}{s/\sqrt{n}} \geq 2.7) \approx 1-0.9965 = 0.0035.$$

Vi kan därför förkasta H_0 på signifikansnivån 0.01.

c. Den exakta nollfördelningen är:

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{11}.$$

Vi får ur tabell:

$$\mathbb{P}(\frac{\bar{X}-\mu}{s/\sqrt{n}} \geq 2.7) \approx 1-0.989 = 0.011,$$

vi kan alltså ej förkasta H_0 på signifikansnivån 0.01.

d. p-värdena är 0.0035 och 0.011 (cirka). Vi ser att med vald signifikansnivå är det en stor skillnad (vi tar ju olika beslut). Lämpligtvis går man vidare och gör en ny mätning under en längre period.

3. (4 poäng) Antag att $X_1, \ldots, X_n$ är oberoende mätdata från en likformig fördelning på intervallet $[1, \theta]$ där $\theta > 1$.

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

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

Lösning:

a. Vi har att $\mathbb{E}[X] = (\theta - 1)/2$, därför blir då MME:en $\bar{X} = (\hat{\theta} - 1)/2$, så att $\hat{\theta} = 2\bar{X} + 1$.

b. Vi har att

$$\begin{aligned} L(\theta) &= f(X_1, \ldots, X_n \Big| \theta) \\ &= \prod_{i=1}^{n} f(X_i \Big| \theta) = \frac{1}{(\theta-1)^n} \prod_{i=1}^{n} I_{1 \leq X_i \leq \theta} \\ &= \frac{1}{(\theta-1)^n} I_{1 \leq \max(X_1, \cdots, X_n) \leq \theta}. \end{aligned}$$

Det är enkelt att se att denna funktion maximeras av $\hat{\theta} = \max(X_1, \cdots, X_n)$ vilket alltså är MLE:en.

3. (4 poäng) Storkar anländer till skorstenarna i en stad enligt exponentialfördelade väntetider med parameter λ. Dvs tiden mellan två ankomster är $Exp(\lambda)$ fördelad.

a. Om $\lambda = 2$, vad är sannolikheten att vi vid tiden $t = 4$ har fått $5, 6, 7, 8$ eller 9 ankomster?

b. Låt T_1 vara tiden det tar tills dess att den första storken anländer. Beräkna $\mathbb{P}(T_1 \leq s + t \big| T_1 \geq s)$

c. Låt T_i vara tiden mellan att storkar nummer $i - 1$ och i anländer. Vad är fördelningen för $\sum_{i=1}^{n} T_i$? Endast svar krävs.

Lösning.
a. Antalet är Poissonfördelade med parameter $\rho = \lambda t = 2 * 4 = 8$. Därför blir sannolikheten

$$\sum_{k=5}^{9} \frac{8^k}{k!} e^{-8} \approx 0.617$$

b. Vi har att

$$\mathbb{P}(T_1 \leq s) = \int_0^s 2e^{-2t} dt = 1 - e^{-2s},$$

därför blir

$$\mathbb{P}(T_1 \geq s + t \big| T_1 \geq s) = \frac{e^{-2(t+s)}}{e^{-2s}} = e^{-2t},$$

så att

$$\mathbb{P}(T_1 \leq s + t \big| T_1 \geq s) = 1 - e^{-2t}.$$

c. $\sum_{i=1}^{n} T_i$ är $Gamma(n, \lambda)$-fördelad.

5. (3 poäng) Maya är 3.5 år gammal och uppskattar julen väldigt mycket. Speciellt presenterna. Hon får tolv julklappar. Nedan kan man se deras pris och ungefär hur glad hon blev över att få dem. Glädjen är svår att mäta exakt men uppskattas (som ett tal mellan 1 och 100) med hjälp av omedelbar reaktion samt hur mycket tid som spenderas med respektive present de närmaste dygnen efter jul. Resultatet blev som följer.

present nr:	1	2	3	4	5	6	7	8	9	10	11	12
glädje:	17	25	37	12	25	34	26	16	23	33	23	13
pris:	59	159	349	49	119	295	195	35	139	299	139	62

a. Plotta glädjen som funktion av pris i en graf.

b. Bestäm regressionslinjen $y = b_0 + b_1 x$.

c. Är ett linjärt samband rimligt? Bedöm från resultatet av din figur

Lösning.
b. Vi använder oss av:

$$b_1 = \frac{12\sum_{i=1}^{12} x_i y_i - \sum_{i=1}^{12} x_i \sum_{i=1}^{12} y_i}{12\sum_{i=1}^{12} x_i^2 - (\sum_{i=1}^{12} x_i)^2}$$

och

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

Vi får att $b_1 \approx 0.08$ och $b_0 \approx 11$.

c. Ett linjärt samband är inte rimligt. Alla mätpunkter ligger under linjen för små och stora priser medans punkterna ligger över för medelstora priser. Snarare är sambandet konkavt t.ex. $y = b0 + b1\sqrt{x}$ eller så.

6. (4 poäng) Halten bly i varmvatten uppmättes vid 14 tillfällen. Resultatet blev som följer.

test nr:	1	2	3	4	5	6	7	8
resultat (ppm):	16.2	13.0	12.9	22.4	19.1	19.0	27.1	21.5
test nr:	9	10	11	12	13	14		
resultat (ppm):	21.0	27.9	16.0	23.5	24.2	18.8		

Dom uppmätta halterna antas vara normalfördelade.

a. Skatta μ och σ.

b. Visa att din skattning av μ är väntevärdesriktig.

c. Ge ett 95 % CI (konfidensintervall) för din skattning av μ. Enligt EU så är det tillåtna gränsvärdet 18.5 ppm, vad säger ditt CI om det?

Lösning:

a. μ skattas med hjälp av $\bar{X} \approx 20.2$ och σ skattas med hjälp av s^2, där

$$s^2 = \frac{1}{13}\sum_{i=1}^{14}(X_i - \bar{X})^2 \approx 21.9.$$

b. $\mathbb{E}[X_i] = \mu$ och

$$\mathbb{E}[(\sum_{i=1}^{n} X_i)/n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu.$$

c. Vi har att

$$T = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{13}.$$

14 prover är för få för att kunna använda normalapproximation. Därför blir vårt 95 % CI för μ:

$$[\bar{X} - \frac{s}{\sqrt{14}}t_{\alpha/2,13}, \bar{X} + \frac{s}{\sqrt{14}}t_{\alpha/2,13}]$$
$$\approx [20.2 - 1.25 * 2.16, 20.2 + 1.25 * 2.16] \approx [17.5, 22.9].$$

7. (3 poäng)
Du är en del av en cell och har därför fått tolv nukleotider. Tre nukleotider vardera av sorterna A,C,T, och G. Din uppgift i cellen är att länka samman dessa till en DNA-sekvens. Du har inte fått något recept och är därför tvungen att sätta ihop dom på måfå.

a. Hur många olika sekvenser kan du bilda av de tolv nukleotiderna?

b. Antag att alla konfigurationer i (a) är lika troliga. Vad blir då sannolikheten att du får en sekvens där alla nukleotider av samma sort ligger bredvid varandra som t.e.x. AAACCCGGGTTT?

Lösning:
a. Om nukleotiderna hette $A1, A2, A3, C1, \ldots$ är det enkelt att se att vi skulle få sammanlagt 12! olika sekvenser. För en given sekvens av A, C, T, G kan $A1, A2, A3$ sorteras på 3! olika sätt. Vi får därför att det existerar

$$\frac{12!}{(3!)^4},$$

olika sekvenser.

b. Det finns exakt 4! olika sätt att generera en sådan sekvens. Alltså blir sannolikheten

$$\frac{4!(3!)^4}{12!}$$

8. (4 poäng) Låt (X, Y) vara likformigt fördelade på området $\Omega = \{|x| + |y| \leq 1\}$.

a. Bestäm pdf:en $f(x,y)$ för (X,Y).

b. Bestäm den betingade pdf:en $f(x\big|y)$.

c. Bestäm $\mathbb{E}[X]$ och $\mathbb{E}[Y]$

d. Är X, Y oberoende? Motivera.

Lösning:

a. Vi har att
$$f(x,y) = \frac{1}{2}, \text{ för } |x| + |y| \leq 1.$$

b. Vi beräknar marginalfördelningen, antag först att $y \geq 0$.
$$f(y) = \int_{y-1}^{1-y} \frac{1}{2} dx = \frac{1}{2}(1 - y - (y-1)) = 1 - y.$$
Om istället $y \leq 0$ får vi att
$$f(y) = \int_{-y-1}^{1+y} \frac{1}{2} dx = \frac{1}{2}(1 + y - (-y-1)) = 1 + y.$$
Dvs att $f(y) = 1 - |y|$. Vi får att
$$f(x\big|y) = \frac{f(x,y)}{f(y)} = \frac{1}{2(1 - |y|)}$$

c. Dessa är 0 av symmetriskäl eller räkning.

d. Nej dom är inte oberoende. T.ex. om för $0 < \epsilon < 1/3$ $X \in [1-\epsilon, 1]$ så innebär det att $Y \in [-\epsilon, \epsilon]$.