

TMS136: Dataanalys och statistik – Tentamen 2013-08-27

Examinator och jour: Mattias Sunden, tel. 0730 79 29 79

Hjälpmedel: Chalmersgodkänd räknare och formelsamling (formelsamling delas ut med tentan).

Betygsgränser: För betyg **3** krävs 14 poäng, för **4** krävs 20 poäng och för betyg **5** krävs 27 poäng. Totalt finns 35 poäng.

*Om du vill att jag skall mejla dig resultatet när jag rättat klart din tenta kan du skriva din mejladress på det **sista** pappret du lämnar in. Skriv i så fall gärna läsligt. Observera att tentan i detta fall kanske inte blir anonym, så strunta i det om det känns viktigt.*

Fullständiga och välmotiverade lösningar skall ges till varje uppgift!

1. Definiera följande begrepp: *Lösningar*
 - (a) Signifikansnivå. *Sannolikheten att göra ett typ I-fel (förkasta nollhypotesen då nollhypotesen är sann)* (1p)
 - (b) Styrka. *Sannolikheten att inte göra ett typ II-fel (att inte göra ett typ II-fel betyder att man förkastar nollhypotesen då nollhypotesen är falsk)* (1p)
 - (c) Väntevärdesriktighet. *En skattare är väntevärdesriktig om väntevärdet av skattaren är precis den parameter som skattaren avser* (1p)
 - (d) Av två väntevärdesriktiga skattare (för samma parameter) väljer du den med? *Minst varians* (1p)
2. Ett elevråd bestående av tre killar och tre tjejer ska väljas bland 7 killar och 8 tjejer. Den ena killen är syskon med den ena tjejen. Vad är sannolikheten att:
 - (a) Båda syskonen hamnar i elevrådet? (2p)

Lösning: Om båda syskonen ska vara med återstår att välja 2 av 6 killar och 2 av 7 tjejer. Detta kan göras på

$$C_2^6 \cdot C_2^7 = \frac{6!}{2!4!} \frac{7!}{2!5!} = 15 \cdot 21 = 315$$

sätt. Det finns alltså 315 gynnsamma utfall. Antalet möjliga utfall vid val av elevråd blir

$$C_3^7 \cdot C_3^8 = \frac{7!}{3!4!} \frac{8!}{3!5!} = 35 \cdot 56 = 1960.$$

Så sannolikheten att båda syskonen hamnar i elevrådet är

$$\frac{315}{1960} = \frac{9}{56}.$$

(b) Inget av syskonen hamnar i elevrådet? (2p)

Lösning: Om inget av syskonen ska vara med ska man välja 3 av 6 killar och 3 av 7 tjejer. Detta kan göras på

$$C_3^6 \cdot C_3^7 = \frac{6!}{3!3!} \frac{7!}{3!4!} = 20 \cdot 35 = 700$$

sätt. Det finns alltså 700 gynnsamma utfall. Så sannolikheten att båda syskonen hamnar i elevrådet är

$$\frac{700}{1960} = \frac{5}{14}.$$

3. Som ni säkert vet har Franska Ligue 1 precis dragit igång och då måste vi ju ha en fotbollsfråga. Jag har kollat upp lite statistik på Zlatan. Han gör mål i ungefär 65% av matcherna och hans lag PSG vinner 84% av dessa matcher. PSG vinner 52% av matcherna där Zlatan inte gör mål. Om PSG vinner en match, vad är sannolikheten att Zlatan gjorde mål? (4p)

Lösning: Låt V vara händelsen att PSG vinner och Z vara händelsen att Zlatan gör mål. Vi har då att

$$\mathbf{P}(Z) = 0.65, \quad \mathbf{P}(V|Z) = 0.84, \quad \mathbf{P}(V|Z') = 0.52.$$

Den sannolikhet vi ska beräkna är $\mathbf{P}(Z|V)$. Enligt Bayes sats följer att

$$\mathbf{P}(Z|V) = \frac{\mathbf{P}(V|Z)\mathbf{P}(Z)}{\mathbf{P}(V|Z)\mathbf{P}(Z) + \mathbf{P}(V|Z')\mathbf{P}(Z')} = \frac{0.84 \cdot 0.65}{0.84 \cdot 0.65 + 0.52 \cdot 0.35} = 0.75.$$

4. Låt X och Y vara oberoende och (*kontinuerligt*) likformigt fördelade på intervallet $[0, 1]$. Låt C och A vara omkretsen respektive arean av en rektangel med sidorna X och Y . Beräkna kovariansen $Cov(C, A)$. (4p)

Lösning: Vi har att $C = 2X + 2Y$ och $A = XY$. Det följer (bilinjariet hos kovariansoperatorn) att

$$Cov(C, A) = Cov(2X + 2Y, XY) = 2Cov(X, XY) + 2Cov(Y, XY).$$

Eftersom X och Y är oberoende har vi

$$\begin{aligned} Cov(X, XY) &= \mathbf{E}[X^2Y] - (\mathbf{E}[X])^2 \mathbf{E}[Y] \\ &= \int_0^1 \int_0^1 x^2 y dx dy - \left(\int_0^1 x dx \right)^2 \int_0^1 y dy = \frac{1}{6} - \left(\frac{1}{2} \right)^2 \cdot \frac{1}{2} = \frac{1}{24}. \end{aligned}$$

Av symmetriskäl är också $Cov(Y, XY) = \frac{1}{24}$. Det följer att $Cov(C, A) = \frac{1}{6}$.

5. Tiden mellan landningar på en flygplats är exponentialfördelad med väntevärde 2 minuter. Antag att ett plan precis har landat. Vad är sannolikheten att det under nästkommande fem minuter landar åtminstone tre plan? (4p)

Lösning: Under antagande om att tiderna mellan landningar är oberoende följer att antalet plan som landar under en minut är Poissonfördelat med väntevärde 0.5 plan. Under antagande om att landningar från en minut till en annan sker oberoende följer att antalet landningar under fem minuter är Poissonfördelat med väntevärde 2.5 plan. Låt N vara antalet plan som landar under fem minuter. Den sökta sannolikheten blir då

$$\mathbf{P}(N \geq 3) = 1 - \mathbf{P}(N \leq 2) = 1 - e^{-2.5} \left(1 + 2.5 + \frac{2.5^2}{2} \right) \approx 0.46.$$

6. Under en Gyllene Tider-konsert spelas 22 låtar. Den förväntade längden för en Gyllene Tider-låt är tre minuter och 30 sekunder. Standardavvikelsen för en låtlängd är 45 sekunder. Under antagande om att låtlängderna är oberoende av varandra och att låtarna väljs godtyckligt bland oändligt många GT-låtar, vad är sannolikheten att den sammanlagda låtlängden under en konsert är större än en timme och 15 minuter? (3p)

Lösning: Enligt CGS är den sammanlagda låtlängden approximativt normalfördelat med väntevärde $22 \cdot 210 = 4620$ sekunder och standardavvikelse $\sqrt{22} \cdot 45$. Om vi låter X_i vara längden för en låt $i = 1, \dots, 22$ följer att den sökta sannolikheten är (vi räknar i sekunder och låter Z beteckna en $N(0, 1)$ -variabel...)

$$\mathbf{P} \left(\sum_{i=1}^{22} X_i > 4500 \right) \approx \mathbf{P} \left(Z > \frac{4500 - 4620}{\sqrt{22} \cdot 45} \right) \approx \mathbf{P}(Z > -0.57) = \Phi(0.57) \approx 0.72.$$

7. Låt $X_1, \dots, X_n$ vara ett stickprov från en fördelning med täthetsfunktion

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0.$$

Bestäm Maximum likelihood-skattaren för λ . (3p)

Lösning: Med den givna täthetsfunktionen följer att likelihood-funktionen är

$$L(\lambda) = \lambda^n e^{-\lambda \sum X_i}.$$

Detta betyder att log-likelihood-funktionen är

$$l(\lambda) = n \ln \lambda - \lambda \sum X_i.$$

Deriverar vi l (med avseende på λ) och sätter derivatan lika med noll följer att ML-skattaren för λ är $\frac{n}{\sum X_i} = \frac{1}{\bar{X}}$.

8. Punkter (X, Y) på enhetscirkeln genereras slumpmässigt genom att låta $X = \cos Z$ och $Y = \sin Z$ där Z är (kontinuerligt) likformigt fördelat på intervallet $[0, 2\pi)$.

- (a) Vad är väntevärdet för X ? (ingen beräkning krävs) (1p)

Lösning: Väntevärdet är noll eftersom det är lika stor area ovanför x-axeln som under x-axeln i integrationsområdet.

(b) Är X och Y okorrelerade? (2p)

Lösning: Ja, eftersom $Cov(X, Y) = \mathbf{E}[XY]$ här och

$$\mathbf{E}[XY] = \mathbf{E}[\cos Z \sin Z] = \frac{1}{2} \mathbf{E}[\sin(2Z)] = 0.$$

(c) Är X och Y oberoende? (2p)

Lösning: Nej, eftersom $X = \pm\sqrt{1 - Y^2}$.

9. En leverantör av muttrar påstår att antalet defekta muttrar i de lådor om 50 muttrar som han levererar är binomialfördelat. Du testar 100 lådor för att undersöka om leveratörens påstående stämmer. Du finner noll defekta muttrar i 44 lådor, en defekt mutter i 36 lådor, två defekta muttrar i 18 lådor och fyra defekta muttrar i resterande lådor. Ställ upp lämpliga hypoteser och testa leveratörens påstående. (4p)

Lösning: Vi ska använda ett goodness of fit-test med nollhypotesen att antalet defekta muttrar i en låda är binomial fördelat med $n = 50$ och någon proportion p . Vi väljer signifikansnivån $\alpha = 0.05$. Vi har inget värde på p så vi skattar p från våra observerade data. Vi observerar att

$$\hat{p} = \frac{0 \cdot 44 + 1 \cdot 36 + 2 \cdot 18 + 4 \cdot 2}{50 \cdot 100} = \frac{2}{125}.$$

Om vi låter X vara antalet defekta muttrar i en låda har vi alltså under nollhypotesen att

$$\mathbf{P}(X = k) = C_k^{50} \left(\frac{2}{125} \right) \left(\frac{123}{125} \right)^{50-k}.$$

Det följer att

$$\mathbf{P}(X = 0) = \left(\frac{123}{125} \right)^{50} \approx 0.4464$$

$$\mathbf{P}(X = 1) = 50 \frac{2}{125} \left(\frac{123}{125} \right)^{49} \approx 0.3630$$

$$\mathbf{P}(X = 2) = 49 \cdot 25 \left(\frac{2}{125} \right)^2 \left(\frac{123}{125} \right)^{48} \approx 0.1446$$

$$\mathbf{P}(X \geq 3) \approx 0.0460$$

Det ger i sin tur att de förväntade antalen lådor med 0, 1, 2 respektive fler än 3 defekta muttrar är

$$E_0 = 0.4464 \cdot 100 = 44.64, \quad E_1 = 0.3630 \cdot 100 = 36.30,$$

$$E_2 = 0.1446 \cdot 100 = 14.46, \quad E_3 = 0.0460 \cdot 100 = 4.64.$$

Det observerade värdet på den χ^2 med två frihetsgrader-fördelade test-statistikan blir alltså

$$\frac{(44 - 44.64)^2}{44.64} + \frac{(36 - 36.30)^2}{36.30} + \frac{(18 - 14.46)^2}{14.46} + \frac{(2 - 4.64)^2}{4.64} \approx 2.38$$

Det kritiska värdet är 5.99 så vi kan alltså inte förkasta nollhypotesen. Vi har alltså inte fog för att säga att leverantören ljugar.