

Lösningar till tentan ESS011: Matematisk statistik och signalbehandling – Tid: 14.00-18.00, Datum: 2010-05-25

Examinatorer: Ottmar Cronie och Bill Karlström

Jour: Bill Karlström, tel. 070-624 44 88. Ottmar Cronie, tel. 031-772 35 44.

Hjälpmedel: Utdelad formelsamling, Formelsamling "Beta", Räknedosa.

Betygsgränser: För betyg 3 krävs 16 poäng, för betyg 4 krävs 24 poäng, för betyg 5 krävs 32 poäng, max. 40 poäng.

Fullständig och välmotiverad lösning på uppgift krävs för att full poäng skall erhållas.

Om du redan är godkänd på signalbehandlingsdelen från tidigare år kan du hoppa över signalbehandlingsuppgiften (sista uppgiften) men ange då detta på ett separat blad!

1. Låt A och B vara två händelser.

- a) Markera (tydligt) händelsen $(A \cup B)^c = A^c \cap B^c$ i ett Venn-diagram. **(1p)**
- b) Visa att om händelserna A och B är oberoende så gäller även att A^c och B^c är oberoende. **(2p)**

Låt nu $\mathbb{P}(A) = 0.95$, $\mathbb{P}(B) = 0.93$ och $\mathbb{P}(A \cap B) = 0.90$.

- c) Är A och B oberoende? **(1p)**
- d) Beräkna $\mathbb{P}(A|B)$. **(1p)**

Notera att i boken används A' istället för A^c .

Lösningar:

- b) $\mathbb{P}(A^c)\mathbb{P}(B^c) = (1-\mathbb{P}(A))(1-\mathbb{P}(B)) = 1-(\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A)\mathbb{P}(B)) = 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)) = 1 - \mathbb{P}(A \cup B) = \mathbb{P}((A \cup B)^c) = \mathbb{P}(A^c \cap B^c)$.
- c) $\mathbb{P}(A)\mathbb{P}(B) = 0.95 \cdot 0.93 = 0.8835 \neq 0.90 = \mathbb{P}(A \cap B)$. Nej, de är ej oberoende.
- d) $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = 0.90/0.93 = 0.968$.

2. Det har visat sig att 60% av alla printere som används i folks hem fungerar korrekt direkt vid installationstillfället. Resten kräver extra justering. En viss datorbutik säljer 10 printere under en viss månad. Låt X beteckna antalet printere som fungerar korrekt vid installationstillfället bland de av butiken sålda printerna.

- a) Vad för fördelning har X ? (Ange både fördelningstyp och värden på fördelningens parametrar). **(1p)**
- b) Hur många av de sålda printerna förväntas fungera korrekt? **(1p)**
- c) Beräkna sannolikheten att åtminstone 2 printere fungerar. **(1.5p)**

- d) Om butiken hade sålt fler printrar, säg 100, då skulle man kunna approximera X 's fördelning med en annan fördelning. Vilken? Vad är det för resultat som motiverar detta? Förklara varför så är fallet. **(1.5p)**

Lösningar:

- a) $X \sim \text{Bin}(n, p) = \text{Bin}(10, 0.6)$
 b) $\mathbb{E}[X] = np = 6$
 c) $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 1) = 1 - \binom{10}{0}0.8^0 0.2^{10} - \binom{10}{1}0.8^1 0.2^9 = 1 - 1.0486 \cdot 10^{-4} - 0.0016 = 0.9983$.
 d) Centrala gränsvärdeessatsen motiverar att vi kan approximera Binomialfördelningen med Normalfördelningen pga att vi kan skriva $X = \sum_{i=1}^n Y_i$ där $Y_1, \dots, Y_n$ är oberoende $\text{Ber}(p)$ -fördelade.
3. Låt $\{X_t\}_{t=0,1,2,\dots}$ vara så kallad slumpvandring, dvs en stokastisk process sådan att

$$X_0 = 0, \quad X_t = \sum_{i=1}^t Y_i, \quad \text{för } t = 1, 2, \dots,$$

där $Y_1, Y_2, Y_3, \dots$ är oberoende och normalfördelade med väntevärde μ och varians σ^2 .

- a) Illustrera grafiskt realisationen av processens 10 första tider, $(X_0, X_1, \dots, X_{10})$, som ges av $\mathbf{x} = (0, 2.42, 3.11, 0.14, 4.16, 5.03, 9.47, 12.88, 10.83, 6.95, 6.73)$ (Denna realisering är tagen från $Y_i \sim N(0, 9)$). **(1p)**
 b) Vad för fördelning har X_t (för ett godtyckligt $t \geq 1$)? Ange både fördelningstyp (motivera!) samt vad den har för parametrar (räkna ut)! **(1.5p)**
 c) Beräkna kovariansen mellan X_1 och X_2 . **(2.5p)**

Lösningar:

- a) Plotta punkterna $(0, \mathbf{x}_1), \dots, (0, \mathbf{x}_{11})$.
 b) $X_t = \sum_{i=1}^t Y_i$ är en linjärkombination av oberoende normalfördelade stokastiska variabler och är därför normalfördelad, $X_t \sim N(\mu_t, \sigma_t^2)$. Dess väntevärde och varians ges av

$$\mu_t = \mathbb{E}[X_t] = \mathbb{E}\left[\sum_{i=1}^t Y_i\right] = \sum_{i=1}^t \mathbb{E}[Y_i] = t\mu$$

$$\sigma_t^2 = \text{Var}(X_t) = \text{Var}\left(\sum_{i=1}^t Y_i\right) \stackrel{\text{ober.}}{=} \sum_{i=1}^t \text{Var}(Y_i) = t\sigma^2.$$

 c) Eftersom Y_1 och Y_2 är oberoende gäller att $\text{Cov}(Y_1, Y_2) = 0$ och därför får vi

$$\text{Cov}(X_1, X_2) = \text{Cov}(Y_1, Y_1 + Y_2) = \mathbb{E}[Y_1(Y_1 + Y_2)] - \mathbb{E}[Y_1]\mathbb{E}[Y_1 + Y_2] = \mathbb{E}[Y_1^2] - \mathbb{E}[Y_1]^2 + \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] = \text{Var}(Y_1) + \underbrace{\text{Cov}(Y_1, Y_2)}_{=0} = \sigma^2.$$

4. Låt $X_1, \dots, X_n$ vara ett stickprov.

- a) Vad är (den matematiska) definitionen av ett stickprov? **(1p)**

- b) Vad är två önskvärda egenskaper hos en punktskattare $\hat{\theta}$ som skattar parametern θ ? Förklara utförligt! **(1p)**

Låt nu $X_1, \dots, X_n$ komma från en kontinuerlig fördelning med täthet $f(x) = \frac{x}{\theta^2} e^{-x/\theta}$, $x > 0$.

- c) Härled (fullständigt) maximum-likelihood-skattaren för parametern θ . **(2p)**
- d) Skatta θ med hjälp av skattaren i c) och observationerna
2.2 5.5 1.7 1.3 3.5 3.2 0.6 3.8 1.9 **(1p)**

Lösningar:

- a) $X_1, \dots, X_n$ är oberoende med samma fördelning.
- b) Väntevärdesriktighet: $\mathbb{E}[\hat{\theta}] = \theta$. Standardfelet, $\sqrt{\text{Var}(\hat{\theta})}$, skall vara litet.
- c)

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{x_i}{\theta^2} e^{-x_i/\theta} \\ l(\theta) &= \ln(L(\theta)) = \sum_{i=1}^n \ln\left(\frac{x_i}{\theta^2}\right) + \sum_{i=1}^n \ln\left(e^{-x_i/\theta}\right) \\ &= \sum_{i=1}^n \ln(x_i) - 2n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^n x_i \\ \frac{dl(\theta)}{d\theta} &= -\frac{2n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i = 0 \\ \Rightarrow \hat{\theta} &= \frac{1}{2n} \sum_{i=1}^n X_i \end{aligned}$$

- d) Skattningen av θ ges av $\hat{\theta} = \frac{1}{2 \cdot 9}(2.2 + 5.5 + 1.7 + 1.3 + 3.5 + 3.2 + 0.6 + 3.8 + 1.9) = 1.3167$.

5. En studie genomfördes där man ville skatta den genomsnittliga vikten för en viss typ av får. Vikten antas vara normalfördelad med (känd) standardavvikelse 5 (kg). Efter att ha vägt 100 får så erhöll man konfidensintervallet

$$92.6 \leq \mu \leq 95.2$$

för vikten (med enheten kg).

- a) Vilken konfidensgrad (i procent) har intervallet? **(2.5p)**
- b) Hur många observationer borde man ha gjort om man vill ha ett intervall med samma intervalllängd men med 95% konfidensgrad? **(1.5p)**
- c) Vad för tre faktorer påverkar längden på ett intervall av detta slag? **(1p)**

Lösningar:

- a) Intervall-längden ges av $2.6 = 95.2 - 92.6 = \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} - (\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) = 2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = 2z_{1-\alpha/2} \frac{5}{\sqrt{100}} = z_{1-\alpha/2}$. Det gäller för $Z \sim N(0, 1)$ att $1 - \alpha/2 = \mathbb{P}(Z \leq z_{1-\alpha/2}) = \mathbb{P}(Z \leq 2.6) = 0.9953$ och därför får vi $\alpha = 0.0094$ så vi har konfidensgraden $100(1 - \alpha)\% = 99.06\%$.
- b) Eftersom $z_{1-\alpha/2} = z_{1-0.05/2} = 1.96$ får vi $2.6 = 2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = 2 \cdot 1.96 \frac{5}{\sqrt{n}}$ vilket ger oss $n \approx 57$.
- c) Längden ökar då: 1) α minskar, 2) σ ökar och 3) n minskar.
6. En grupp sjuksköterskor vill studera om antalet mödravårdsbesök (före barnets födelse) påverkar födelsevikten av barnet så att mödrar med ett stort antal besök skulle föda tyngre barn än de med ett litet antal besök. Mödrar (tagna slumpmässigt från födelseregistret) delades in i två grupper: 14 mödrar hade haft 5 eller färre mödravårdsbesök före födseln och 14 mödrar hade haft 6 eller fler mödravårdsbesök före födseln. Låt X (5 eller färre) och Y (6 eller fler) beteckna de respektive födelsevikterna hos de nyfödda i dessa två grupper av mödrar.
- a) Formulera noll- och mot-hypotesen för att hjälpa sjuksköterskorna sin frågeställning. Motivera ditt val av mothypotes! **(0.5p)**
- b) Testa hypotesen i a) på signifikansnivån 5% genom att använda viktorna i tabellen nedan (enheten är uns, 1 uns är ungefär 28.35 gram): $(\bar{x} = 100.21, \bar{y} = 114.21, s_X = 24.59, s_Y = 18.15)$ **(2p)**

x	49	108	110	82	93	114	134	114	96	52	101	114	120	116
y	133	108	93	119	119	98	106	131	87	153	116	129	97	110

- c) Vad för statistiskt fel finns det risk att vi begår, baserat på vår utsaga i b)? **(0.5p)**
- d) Skulle du få samma resultat om du använde en annan mothypotes (dvs en ensidig om du använde en tvåsidig i a) eller en tvåsidig om du använde en ensidig i a)? **(1p)**
- e) Vad har du gjort för antaganden i b)? Motivera! **(1p)**

Lösningar:

- a) Låt $\mu_X = \mathbb{E}[X]$ och $\mu_Y = \mathbb{E}[Y]$. Vi testar $H_0 : \mu_X = \mu_Y$ mot $H_1 : \mu_X < \mu_Y$ eftersom de tror att mödrar med fler besök föder tyngre barn.
- b) Man använder ett två-stickprovs T-test där man antar att varianserna är lika.
 Teststatistika: $T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{1/n_X + 1/n_Y}} \sim T_{n_X + n_Y - 2} = T_{26}$ under H_0 -fördelningen.
 Då $\alpha = 0.05$ förkastar vi H_0 om $|T_{obs}| > t_{1-0.05, 26} = 1.706$.
 Vi får $s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} = \frac{(14 - 1)24.59 + (14 - 1)18.15}{14 + 14 - 2} = 466.87$,

$$s_p = 21.61 \text{ och d\u00e4rmed } T_{obs} = \frac{\bar{x} - \bar{y}}{S_p \sqrt{1/n_x + 1/n_y}} = \frac{100.21 - 114.21}{4.6228 \sqrt{1/14 + 1/14}} = -1.7143.$$

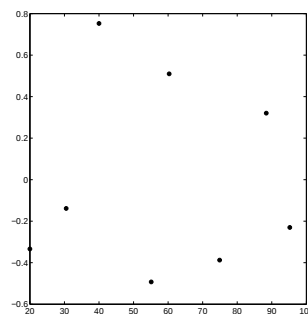
Vi kan f\u00f6rkasta H_0 p\u00e5 signifikansniv\u00e5n 0.05 dvs det verkar som ett \u00f6kat antal bes\u00f6k leder till h\u00f6gre vikt.

- c) Typ I-fel.
 - d) Nej, vi skulle inte f\u00e5 samma resultat eftersom $t_{1-0.05/2, 26} = 2.056 > 1.7143 = |T_{obs}|$ och d\u00e4rmed skulle vi i detta fall inte kunna f\u00f6rkasta H_0 .
 - e) Vi har antagit att vi betraktar tv\u00e5 oberoende stickprov tagna fr\u00e5n $N(\mu_X, \sigma^2)$ respektive $N(\mu_Y, \sigma^2)$. Notera att vi har antagit att varianserna \u00e4r lika.
7. F\u00f6rh\u00e5llandet mellan energikonsumption och hush\u00e5lls inkomst studerades i hush\u00e5ll i USA. Man fick f\u00f6ljande data f\u00f6r inkomst X (\$1000/\u00e5r) och energikonsumption Y (10^8 Btu/\u00e5r).

x	20.0	30.5	40.0	55.1	60.3	74.9	88.4	95.2
y	1.8	3.0	4.8	5.0	6.5	7.0	9.0	9.1

$$(\sum_i x_i = 464.40, \sum_i y_i = 46.20, \sum_i x_i^2 = 32089.96, \sum_i y_i^2 = 315.34, \sum_i x_i y_i = 3173.17)$$

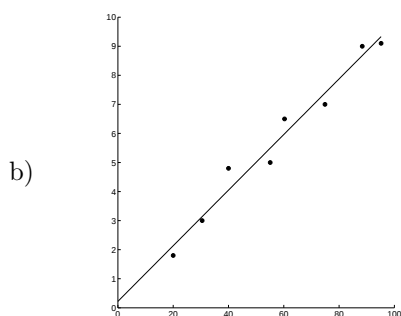
- a) Finn skattningen av den linj\u00e4ra regressionsekvationen $\mu_{Y|x}$ och visa dess slutgiltiga form. Tolka de skattade koefficienterna i ord. **(2p)**
- b) Rita en graf som inneh\u00e5ller punkterna $(x_1, y_1), \dots, (x_8, y_8)$ tillsammans med $\hat{\mu}_{Y|x}$. **(0.5p)**
- c) Detta \u00e4r en s\u00e5 kallad residualplott. Tolka plotten. **(1p)**



- d) Beräkna determinationskoefficienten R^2 och avgör, baserat på denna, om modellen passar data. **(1.5p)**

Lösningar:

$$\begin{aligned} \text{a) } b_1 &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{3173.17 - \frac{1}{8} 464.40 \cdot 46.20}{32089.96 - \frac{1}{8} 464.40^2} = 0.0957 \\ b_0 &= \bar{y} - b_1 \bar{x} = \frac{1}{8} 46.20 - 0.0957 \frac{1}{8} 464.40 = 0.2196 \\ \hat{\mu}_{Y|x} &= b_0 + b_1 x = 0.2196 + 0.0957x \end{aligned}$$



- c) Den talar om för oss att regressionsmodellen verkar rimlig. Residualerna ligger centrerade kring 0 och residualerna sprider sig jämnt över x -axeln.

$$\text{d) } r^2 = \frac{S_{xy}^2}{S_{yy} S_{xx}} = \frac{(\sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i)^2}{(\sum_i x_i^2 - \frac{1}{n} (\sum_i x_i)^2)(\sum_i y_i^2 - \frac{1}{n} (\sum_i y_i)^2)} = 0.9690 \text{ så den linjära regressions modellen förklarar en stor del av variationen i } Y.$$