Tentamentsskrivning i **Matematisk Statistik med Metoder MVE490**

Tid: 21 december 2017
Examinatorer: Erik Broman.
Hjälpmedel: Typgodkänd miniräknare, egenhändigt skriven formelsamling om två A4 fram och bak (dvs 4 sidor), samt utdelade tabeller.

Tentamen består av 8 frågor om sammanlagt 50 poäng. Preliminära betygsgränser är satta till:
betyg "3": 20 till 29
betyg "4": 30 till 39 poäng
betyg "5": 40 eller fler poäng.

OBS! Alla lösningar skall vara väl redovisade och motiverade. Talen är ej ordnade efter svårighetsgrad

1. Alexandra åker till Nya Zeeland. Väl där blir hon attackerad av en aggressiv Kiwi-fågel. Vi kan anta att antalet anfall innan Kiwi-fågeln ger upp är Poisson-fördelat med parameter $\lambda = 4/3$.

   (a) Vad är sannolikheten att Alexandra blir utsatt för högst tre anfall? (2p)

   (b) Givet att Alexandra blir utsatt för minst två anfall, vad blir nu sannolikheten att hon blir utsatt för högst tre anfall? (2p)

   (c) Antag att Alexandra blir attackerad av 100 Kiwi fåglar och att varje fågel attackerar oberoende av varandra. Antag också att varje Kiwi-fågel anfaller ett Poisson-fördelat antal gånger med $\lambda = 4/3$. Beräkna (approximativt) sannolikheten att Alexandra får utstå minst 150 anfall. (3p)

   Lösning.

   (a) Låt X =antalet anfall. Vi söker då

   $$\mathbb{P}(X \leq 3) = \sum_{k=0}^{3} \mathbb{P}(X = k) = \sum_{k=0}^{3} \frac{(4/3)^k}{k!} e^{-4/3} \approx 0.9534.$$

   (b) Här efterfrågas

   $$\mathbb{P}(X \leq 3 | X \geq 2) = \frac{\mathbb{P}(2 \leq X \leq 3)}{\mathbb{P}(X \geq 2)} = \frac{\mathbb{P}(2 \leq X \leq 3)}{1 - \mathbb{P}(X \leq 1)} = \frac{\mathbb{P}(X = 2) + \mathbb{P}(X = 3)}{1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1)}.$$

   Vi använder att $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ för att få

   $$\mathbb{P}(X \leq 3 | X \geq 2) \approx 0.8792$$

(c) Vi använder normalapproximation. Om X_k är antalet anfall från fågel nummer k och Y = antal anfall sammanlagt har vi att

$$Y = \sum_{k=1}^{100} X_k$$

är approximativt normalfördelat. Parametrarna blir då

$$\mathbb{E}\left[\sum_{k=1}^{100} X_k\right] = 100\mathbb{E}[X_1] = 100\frac{4}{3} = \frac{400}{3},$$

och

$$\mathrm{Var}\left(\sum_{k=1}^{100} X_k\right) = 100\mathrm{Var}(X_1) = \frac{400}{3},$$

där vi har använt oberoendet. Vi har alltså att $Y \sim N(400/3, 400/3)$ och får då att

$$\mathbb{P}(Y \geq 150) = \mathbb{P}\left(\frac{Y - 400/3}{\sqrt{400/3}} \geq \frac{150 - 400/3}{\sqrt{400/3}}\right) \approx \mathbb{P}(Z \geq 1.44) \approx 0.0749$$

enligt tabell.

2. Antag att X är likformigt fördelad på talen $1, 2, \ldots, 10$ (dvs $X \sim U(\{1, 2, \ldots, 10\})$), medans Y är likformigt fördelad på talen $1, 2, \ldots, 6$ (dvs $Y \sim U(\{1, 2, \ldots, 6\})$). Antag även att X och Y är oberoende.

   (a) Vad är sannolikheten att $X + Y$ blir minst 4? (2p)

   (b) Vad blir $\mathbb{E}[2X + Y]$? (2p)

   (c) Låt Z vara likformigt fördelad på talen $101, 102, \ldots, 110$ (dvs $Z \sim U(\{101, 102, \ldots, 110\})$). Ordna varianserna av X, Y och Z i storleksordning. (OBS: Om man löser talet med räkningar får man full poäng. Om man istället löser talet med ett korrekt och fullständigt resonemang får man också full poäng) (2p)

Lösning.

(a) Vi söker $\mathbb{P}(X+Y \geq 4) = 1-\mathbb{P}(X+Y \leq 3) = 1-\mathbb{P}(X+Y = 2)-\mathbb{P}(X+Y = 3)$. Vidare är

$$\mathbb{P}(X + Y = 2) = \mathbb{P}(X = 1, Y = 1) = \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = \frac{1}{10}\frac{1}{6} = \frac{1}{60}$$

och

$$\begin{aligned}\mathbb{P}(X + Y = 3) &= \mathbb{P}(X = 1, Y = 2) + \mathbb{P}(X = 2, Y = 1)\\ &= \mathbb{P}(X = 1)\mathbb{P}(Y = 2) + \mathbb{P}(X = 2)\mathbb{P}(Y = 1) = \frac{1}{60} + \frac{1}{60} = \frac{2}{60}.\end{aligned}$$

Sammanlagt får vi då att

$$\mathbb{P}(X+Y \geq 4) = 1 - \mathbb{P}(X+Y \leq 3) = 1 - \frac{3}{60} = \frac{57}{60}.$$

(b) Vi har att

$$\begin{aligned}\mathbb{E}[2X+Y] &= 2\mathbb{E}[X] + \mathbb{E}[Y] \\ &= 2\sum_{k=1}^{10} k\mathbb{P}(X=k) + \sum_{l=1}^{6} l\mathbb{P}(Y=l) \\ &= 2\sum_{k=1}^{10} \frac{k}{10} + \sum_{l=1}^{6} \frac{l}{6} = 2\frac{11}{2} + \frac{7}{2} = \frac{29}{2}.\end{aligned}$$

(c) Slumpvariabeln Z kan skrivas som $X+100$. Vi ser då att $\text{Var}(Z) = \text{Var}(X+100) = \text{Var}(X)$ då en adderad konstant inte bidrar till variansen. Vidare ser vi att Y måste ha mindre spridning än X, då Y är likformigt fördelad på en delmängd av mängden på vilken X är likformigt fördelad. Vi drar därför slutsatsen att $\text{Var}(Y) < \text{Var}(X) = \text{Var}(Z)$.

3. Låt A, B var två händelser med sannolikheter $\mathbb{P}(A) = 1/4$ och $\mathbb{P}(B) = 1/3$.

(a) Beräkna $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B^c)$ och $\mathbb{P}(A|B)$ om $\mathbb{P}(A \cap B) = 1/5$. (2p)

(b) Beräkna $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B^c)$ och $\mathbb{P}(A|B)$ om A, B är oberoende. (2p)

(c) Beräkna $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B^c)$ och $\mathbb{P}(A|B)$ om A, B är disjunkta. (2p)

Lösning:

(a) Vi får att

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{1}{4} + \frac{1}{3} - \frac{1}{5} = \frac{15+20-12}{60} = \frac{23}{60}.$$

Vidare har vi att

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \frac{1}{4} - \frac{1}{5} = \frac{1}{20},$$

och att

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/5}{1/3} = \frac{3}{5}.$$

(b) Vi får att

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{1}{4} + \frac{1}{3} - \frac{1}{4}\frac{1}{3} = \frac{3+4-1}{12} = \frac{1}{2}.$$

Vidare har vi att

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \frac{1}{4} - \frac{1}{12} = \frac{1}{6},$$

och att

$$\mathbb{P}(A|B) = \mathbb{P}(A) = \frac{1}{4}.$$

(c) Vi får att

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}.$$

Vidare har vi att

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) = \frac{1}{4},$$

och att

$$\mathbb{P}(A|B) = 0.$$

4. Försäkringsbolaget Hans Trygga nishar sig på två typer av försäkringar, nämligen brandförsäkringar och stöldförsäkringar. Bolaget har 12300 försäkringstagare för brandförsäkringarna och 23800 försäkringstagare för stöldförsäkringarna. Utgifterna för dessa försäkringar (dvs ersättningar till kunder som råkat ut för brand respektive stölder) anses vara oberoende och normalfördelade med parametrar $\mu_b = 60$ MSEK (dvs 60 miljoner kronor) och $\sigma_b^2 = 200$ för brandförsäkringarna samt $\mu_s = 23$ MSEK och $\sigma_s^2 = 54$ för stöldförsäkringarna.

(a) Om de sammanlagda försäkringspremierna för brandförsäkringarna uppgår till 70 MSEK, hur stor är sannolikheten att Hans Trygga går med vinst på brandförsäkringarna? (2p)

(b) Hur mycket måste Hans Trygga minst få för stöldförsäkringspremierna för att sannolikheten att de skall gå med vinst skall vara åtminstone 0.9? (2p)

(c) Hans Tryggas sammanlagda tillgångar uppgår till 112 MSEK efter att alla försäkringspremier har betalats in. Om de sammanlagda utbetalningarna för stöld och brand överstiger detta belopp går Hans Trygga i konkurs. Beräkna sannolikheten för att detta inträffar. (2p)

Lösning:

(a) Låt $X_b \sim N(60, 200)$, vi efterfrågar då

$$\mathbb{P}(X_b \leq 70) = \mathbb{P}\left(\frac{X_b - 60}{\sqrt{200}} \leq \frac{70 - 60}{\sqrt{200}}\right) \approx \mathbb{P}(Z \leq 0.7) \approx 0.758.$$

(b) Vi söker B så att

$$0.9 = \mathbb{P}(X_s \leq B) = \mathbb{P}\left(\frac{X_s - 23}{\sqrt{54}} \leq \frac{B - 23}{\sqrt{54}}\right) = \mathbb{P}\left(Z \leq \frac{B - 23}{\sqrt{54}}\right).$$

En titt i tabell ger att $\mathbb{P}(Z \leq 1.28) \approx 0.9$ så att

$$\frac{B - 23}{\sqrt{54}} = 1.28,$$

vilket ger att $B \approx 32.4$ MSEK.

(c) Vi har att $X_b + X_s \sim N(60 + 23, 200 + 54) = N(83, 254)$ då X_b, X_s enligt uppgift är oberoende. Hans Trygga går i konkurs om $X_b + X_s \geq 112$, så vi söker

$$\begin{aligned}&\mathbb{P}(X_b + X_s \geq 112)\\ &= \mathbb{P}\left(\frac{X_b + X_s - 83}{\sqrt{254}} \geq \frac{112 - 83}{\sqrt{254}}\right) \approx \mathbb{P}(Z \geq 1.8196) \approx 0.0344.\end{aligned}$$

5. Tet är intresserad av att undersöka hur stor andel av studenterna på Chalmers som är vegetarianer. Hen gör en undersökning genom att fråga 150 personer på campus.

(a) Tet hoppas att stickprovet ska vara stort nog för att ett 95%-konfidensintervall ska bli högst 0.1 långt (dvs sträcka sig högst 0.05 åt varje håll från punktskattningen). Är detta säkerställt, dvs hur många personer behöver tillfrågas för att det ska vara helt säkert att konfidensintervallets längd blir högst 0.1?

(3p)

(b) När Tet genomför sin undersökning svarar 28 utav de 150 tillfrågade att de är vegetarianer. Konstruera ett 95% konfidensintervall för det sanna andelen vegetarianer på Chalmers utifrån datan Tet samlat in.

(2p)

(c) Vilket/vilka villkor behöver vara uppfyllt för att konfidensintervallet i b) ska vara giltigt? Verkar detta vara uppfyllt?

(1p)

Lösning:

(a) Konfidensintervallet ges av

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

där $\hat{p}$ är andelen vegetarianer av de tillfrågade, n är antalet tillfrågade och $z_{\alpha/2}$ är den önskade kvantilen från normalfördelningen. För ett 95% konfidensintervall är $\alpha = 0.05$, och således $z_{\alpha/2} = 1.96$. Det som ska vara uppfyllt enligt uppgiften är att

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 0.05 \Leftrightarrow n > z^2_{\alpha/2}\frac{\hat{p}(1-\hat{p})}{0.05^2} \qquad (1p)$$

Innan vi gjort undersökningen kan vi omöjligt veta vad $\hat{p}$ är, så för att vara säkra på att få ett så kort konfidensintervall som vi önskar räknar vi med det värde på $\hat{p}$ som ger längsta möjliga konfidensintervall, vilket är $\hat{p} = 0.5$ (1p). Med detta värde får vi att vi måste fråga minst 385 personer (1p).

(b) Konfidensintervallet ges av

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \qquad (1p)$$

Som nämnts ovan blir $z_{\alpha/2} = 1.96$, och n är nu 150. Vidare är $\hat{p} = 28/150 = 0.187$. Detta ger att konfidensintervallet blir $[0.124, 0.249]$ (1p).

(c) För att konfidensintervallet ska vara giltigt ska n vara stort, vilket enligt tumregeln betyder $np > 15$ och $n(1-p) > 15$. Vi vet inte p, utan får använda $\hat{p}$, vilket ger $n\hat{p} = 28$ (och $n(1-\hat{p}) = 122$). Detta är betryggande mycket större än 15, så ja det tycks vara uppfyllt. (1p)

6. Sten-Åke livnär sig på att tillverka tegel, och har precis fått en ny tegelstensmaskin till sin fabrik. En fördel med den nya maskinen är att vikten varierar mindre mellan stenarna – den gamla maskinen tillverkade stenar med en vikt som hade en standardavvikelse på 0.5 kg, medan den för den nya maskinen bara är 0.25 kg. Men Sten-Åke misstänker att den nya maskinen även gör lättare stenar. Han väger därför 35 stenar från den nya maskingen, och 37 som var tillverkade i den gamla. Nedan är en sammanfattning av vikterna. Låt x beteckna stickprovet av stenar från den gamla maskinen och y stickprovet från den nya maskinen.

Gamla	Nya
$n_X = 37$	$n_Y = 35$
$\sum x_i = 119.2$	$\sum y_i = 104.1$
$\sum x_i^2 = 391.2$	$\sum y_i^2 = 311.6$

Sten-Åke tycker inte resultatet är tydligt nog, utan vill ha hjälp av dig med en statistisk analys.

(a) Formulera hypoteser och gör ett hypotestest vid signifikansnivå 1%. Verkar Sten-Åke ha rätt i sina misstankar?

(4p)

(b) Sten-Åke förstår inte vad du menar med "signifikansnivå". Hur skulle du förklara för Sten-Åke vad det innebär att signifikansnivån är 1%?

(2p)

Lösning:

(a) Hypoteserna vi vill testa är (1p)

$$H_0 : \mu_X = \mu_Y \quad \text{mot}$$
$$H_a : \mu_X > \mu_Y$$

Notera att vi vet standardavvikelserna för de två maskinerna, som $\sigma_X = 0.5$ och $\sigma_Y = 0.25$. Teststatistikan som ska användas är därför

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} . (1p)$$

Från tabellen får vi att $\bar{x} = 119.2/37 = 3.22$ och $\bar{y} = 104.1/35 = 2.97$. Med dessa värden instoppade får vi att det observede värdet på teststatistikan är $t = 2.68$. Det kritiska värdet är $z_\alpha = z_{.01} = 2.33(1p)$, eftersom vi gör ett enkelsidigt test. Då värdet på teststatistikan är större än det kritiska värdet förkastar vi H_0. (1p)

(b) Signifikansnivån är sannolikheten att vi förkastar H0 om den är sann/risken för typ I-fel. (2p)

7. Avgör om följande påståenden är sanna eller falska. Svara utan motivering! Varje rätt svar ger 1 poäng men varje fel svar ger -1 poäng. Inget svar ger 0 poäng. Totalt ger uppgiften som sämst 0 poäng. Tänk igenom varje svar ordentligt!

(a) Om $X_1, \ldots, X_n$ är ett stickprov från en fördelning så är $\bar{X}$ en väntevärdesriktig skattare för fördelningens väntevärde.

(b) Om $X_1, \ldots, X_n$ är ett stickprov från en fördelning så är s^2 en väntevärdesriktig skattare för fördelningens standardavvikelse.

(c) Om $X_1, \ldots, X_n$ är ett stickprov från en fördelning och om n är stort nog så säger Centrala gränsvärdessatsen att stickprovet är approximativt normalfördelad.

(d) Ju högre signifikansnivå desto större är den kritiska regionen.

(e) Sannolikheten att göra ett typ-II-fel är alltid större än sannolikheten att göra ett typ-I-fel.

(f) Om vi tar ett stickprov från en $N(0, 1)$-fördelning så kommer ungefär 95% av observationerna att ligga mellan -1.96 och 1.96.

(g) Om p-värdet är mindre än signifikansnivån så förkastar vi nollhypotesen.

(h) Medelvärde, median och typvärde är exempel på spridningsmått.

Lösning:

(a) Sant.

(b) Falskt, s^2 är en väntevärdesriktig skattare för variansen, inte för standardavvikelsen.

(c) Falskt, det är stickprovsmedelvärdet, inte stickprovet, som är approximativt normalfördelat.

(d) Sant. Ju högre signifikansnivå desto mindre extrema observationer krävs för att vi ska förkasta nollhypotesen, dvs desto större är den kritiska regionen.

(e) Falskt. Sannolikheten för ett typ-II-fel beror på det riktiga värdet på parametern och kan mycket väl vara mindre än signifikansnivån som är sannolikheten att göra ett typ-I-fel.

(f) Sant, sannolikheten för en $N(0,1)$-fördelad variabel att hamna mellan -1.96 och 1.96 är precis 95%.

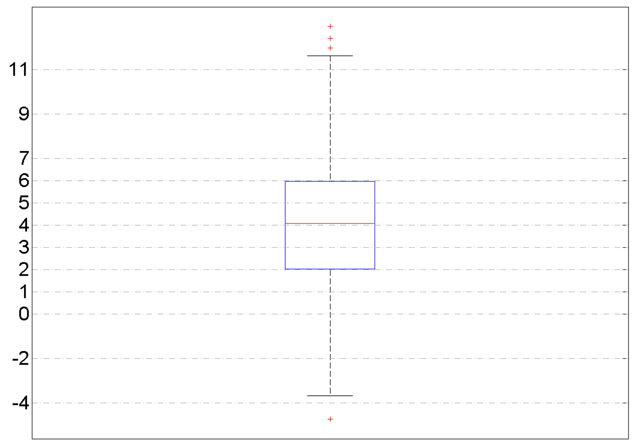
(g) Sant.

(h) Falskt, de är exempel på lägesmått.

8. Anna har samlat in en del data från en intressant studie. Som ett första steg i att analysera stickprovet har hon skapat en boxplot med hjälp av ett lämpligt datorprogram. I bilden nedan kan du se denna boxplot.

   (a) Förklara för Anna hur man kan få fram medianen och kvartilavståndet för ett stickprov genom att titta i boxploten. Ungefär vad är medianen och kvartilavståndet för detta stickprov? (3p)

   (b) Förklara för Anna hur en boxplot för ett skevt stickprov ser ut. Skulle du säga att detta stickprov är skevt? (2p)



Lösning:

(a) Medianen av stickprovet är den mittersta linjen i en boxplot, i detta fall ungefär lika med 4 (1p). Kvartilavståndet ges av $Q_U - Q_L$, dvs den övre kvartilen minus den undre kvartilen. Kvartilerna i en boxplot ges av gränserna på lådan. I detta fall ungefär 6 och 2. Kvartilavståndet är alltså ungefär lika med 6-2 = 4 i detta fall. (2p)

(b) Att stickprovet är skevt innebär att det är asymmetriskt sett från medelvärdet, dvs att det finns mycket fler 'extrema' observationer åt ena hållet än åt andra hållet. Ett skevt stickprov ger upphov till en asymmetrisk boxplot. Till exempel skulle då medianlinjen inte ligga i mitten av lådan. I detta fallet verkar det inte som att stickprovet är skevt.