Tentamentsskrivning i **Matematisk Statistik med Metoder MVE490**

Tid: 21 oktober 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. I ett lotteri finns det $N$ lotter, där $N$ är ett mycket stort tal. Sannolikheten att en given lott ger vinst är $p$.

   (a) Sara köper 12 lotter och låter $X$ vara antalet vinster av dessa 12. Hon antar att $X$ är binomialfördelad. Vilka antaganden måste Sara göra för att detta skall stämma? Är dessa antaganden rimliga? Vilka är parametrarna? (2p)

   (b) Om $N = 20000$ och $p = 0.1$, beräkna $\mathbb{P}(X = 3)$ där $X$ är som i uppgift $(a)$. (2p)

   (c) I ett annat lotteri är $p$ istället 0.005. Torkel köper $n = 1000$ lotter. Om $Y = $#vinster av dessa 1000 lotter, beräkna $\mathbb{P}(Y \leq 3)$ och ange svaret med tre värdesiffror. (2p)

   **Lösning.**

   (a) För att antalet vinster skall vara binomialfördelat så måste vinstchanserna vara samma för alla lotter och dessutom vara oberoende. Oberoende är alltså det extra antagandet som Sara har gjort. Detta är antagligen inte sant då antalet vinster typiskt är fixt i ett lotteri. Dock gör det faktum att $N$ är stort att det är approximativt ok. Parametrarna är $n = 12$ och $p$ som är okänt.

   (b) Storleken på $N$ är inte relevant. Vi har att

   $$\mathbb{P}(X = 3) = \binom{12}{3} 0.1^3 (1 - 0.3)^9 \approx 0.0089$$

(c) Då $p$ är mycket litet och $n$ är stort använder vi Poisson approximation. Dvs att $Y \sim$Poi(5). Vi får då att

$$\begin{aligned}\mathbb{P}(Y \leq 3) &= \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) + \mathbb{P}(Y = 2) + \mathbb{P}(Y = 3)\\ &= e^{-5}\left(\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!}\right) \approx 0.265.\end{aligned}$$

2. Kalle har två vanliga sexsidiga tärningar och kastar båda.

(a) Låt $A$ vara händelsen att summan av resultaten är högst fem. Vad blir $\mathbb{P}(A)$? (2p)

(b) Låt $B$ vara händelsen att summan av resultaten är ett udda tal. Vad blir $\mathbb{P}(B)$? (2p)

(c) Är $A, B$ oberoende? (2p)

**Lösning:**

(a) Det enklaste torde vara att använd sig av divisionsregeln. Vi har att

$$A = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\}$$

och därmed blir

$$\mathbb{P}(A) = \frac{|A|}{|S|} = \frac{10}{36}.$$

(b) Låt $X_i$ =resultatet av kast $i$ där $i = 1, 2$. Vi får då att

$$\begin{aligned}&\mathbb{P}(X_1 + X_2 \text{ udda})\\ &\quad= \mathbb{P}(X_1 + X_2 \text{ udda } | X_1 \text{ jämnt })\mathbb{P}(X_1 \text{ jämnt })\\ &\qquad+\mathbb{P}(X_1 + X_2 \text{ udda} | X_1 \text{ udda })\mathbb{P}(X_1 \text{ udda })\\ &\quad= \mathbb{P}(X_2 \text{ är udda })\mathbb{P}(X_1 \text{ jämnt }) + \mathbb{P}(X_2 \text{ jämnt })\mathbb{P}(X_1 \text{ udda })\\ &\quad= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.\end{aligned}$$

(c) Vi har att

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

och att

$$\mathbb{P}(A \cap B) = \frac{|A \cap B|}{|S|} = \frac{6}{36}.$$

Därmed är

$$\mathbb{P}(A|B) = \frac{|A \cap B|}{|S|}\frac{1}{1/2} = 2\frac{6}{36} = \frac{12}{36} \neq \mathbb{P}(A).$$

Vi drar slutsatsen att $A$ och $B$ ej är oberoende.

3. Hummerpremiären brukar ge höga priset vid årets första hummerauktion. Hummrarna säljs vid detta första tillfälle individuellt, och kilopriset för hummer kan anses vara normalfördelat med parametrar $\mu = 1000$ och $\sigma^2 = 2500$.

   (a) Vad är sannolikheten att hummerpriset överstiger 1150 kronor? (2p)

   (b) Jöns-Harald har fiskat upp en hummer som väger 6.3 kilo, vad är sannolikheten att Jöns-Harald får mer än 6000 kronor för sin hummer? (2p)

   (c) Mr Bakos har en hummer som väger 4.6 kilo och en exklusiv bläckfisk som väger 5.4 kilo. Kilopriset för bläckfisken kan antas följa en normalfördelning med väntevärde 600 och varians på 1100 kronor (även en exklusiv bläckfisk är mindre värd än en hummer). Om priserna är oberoende, vilken fördelning har den totala summan pengar som Mr Bakos får vid försäljning av hummern och bläckfisken? (2p)

   **Lösning**

   (a) Låt $X \sim N(1000, 1500)$, vi har att

$$\begin{aligned}\mathbb{P}(X \geq 1150) &= \mathbb{E}\left(\frac{X - 1000}{\sqrt{2500}} \geq \frac{1150 - 1000}{\sqrt{2500}}\right)\\ &= \mathbb{P}(Z \geq 3) = 1 - \mathbb{P}(Z \leq 3) \approx 1 - 0.9987 = 0.0013.\end{aligned}$$

   (b) Låt $Y$ vara priset för Jöns-Haralds hummer. Då gäller att $Y = 6.3X \sim N(6300, 99225)$. Vi får då att

$$\begin{aligned}\mathbb{P}(Y \geq 6000) &= \mathbb{E}\left(\frac{Y - 6300}{\sqrt{99225}} \geq \frac{6000 - 6300}{\sqrt{99225}}\right)\\ &= \mathbb{P}(Z \geq -0.9524) = \mathbb{P}(Z \leq 0.9524) \approx 0.8296.\end{aligned}$$

   (c) Låt $H, B$ vara priserna så att $H \sim N(4600, 52900)$ och $B \sim N(3240, 32076)$. Då priserna enligt uppgift är oberoende får vi att $H+B \sim N(7840, 84976)$.

4. Låt $X$ vara en slumpvariabel som kan anta värdena $-1, 3, 4$ och $6$ där sannolikheterna anges i följande tabell.

| $k$ | $-1$ | $3$ | $4$ | $6$ |
|---|---|---|---|---|
| $\mathbb{P}(X = k)$ | 0.2 | 0.3 | 0.4 | 0.1 |

   (a) Beräkna väntevärde och varians av $X$. (2p)

   (b) Låt $X_1, X_2, X_3, X_4$ och $X_5$ vara oberoende slumpvariabler som alla har samma fördelning som $X$. Beräkna väntevärde och varians för summan $X_1 + X_2 + X_3 + X_4 + X_5$. Beräkna även väntevärde och varians för $X_1 - X_2$. (3p)

(c) Låt nu $X_1, \dots, X_{30}$ vara oberoende slumpvariabler som alla har samma fördelning som $X$. Beräkna $\mathbb{P}\left(\sum_{k=1}^{30} X_k \geq 90\right)$. (2p)

**Lösning**

(a) Vi har att

$$\mathbb{E}[X] = \sum k\mathbb{P}(X = k) = -1 * 0.2 + 3 * 0.3 + 4 * 0.4 + 6 * 0.1 = 2.9.$$

Dessutom gäller att

$$\mathbb{E}[X^2] = \sum k^2\mathbb{P}(X = k) = (-1)^2 * 0.2 + 3^2 * 0.3 + 4^2 * 0.4 + 6^2 * 0.1 = 12.9,$$

så att

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 12.9 - 2.9^2 = 4.49.$$

(b) Räkningen blir lätt om vi använder linearitet. Vi får att

$$\mathbb{E}[X_1 + X_2 + X_3 + X_4 + X_5] = \sum_{k=1}^{5} \mathbb{E}[X_k] = 5 * 2.9 = 14.5$$

och att

$$\mathrm{Var}(X_1 + X_2 + X_3 + X_4 + X_5) = \sum_{k=1}^{5} \mathrm{Var}[X_k] = 5 * 4.49 = 22.45.$$

Vidare blir

$$\mathbb{E}[X_1 - X_2] = \mathbb{E}[X_1] - \mathbb{E}[X_2] = 0$$

och

$$\mathrm{Var}[X_1 - X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] = 2.4.49 = 8.98.$$

(c) Låt nu $X_1, \dots, X_{30}$ vara oberoende slumpvariabler som alla har samma fördelning som $X$. Vi får utnyttja centrala gränsvärdessatsen som ger att

$$Y = \sum_{k=1}^{30} X_k \sim N(30 * \mathbb{E}[X], 30 * \mathrm{Var}(X)) = N(87, 134.7).$$

Vi får då att

$$\begin{aligned}\mathbb{P}(Y \geq 90) &= \mathbb{P}\left(\frac{X - 90}{\sqrt{134.7}} \geq \frac{87 - 90}{\sqrt{134.7}}\right) \\ &\approx P(Z \geq -0.2585) = \approx P(Z \leq 0.2585) \approx 0.602.\end{aligned}$$

5. 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.

(a) Om p-värdet är mindre än det kritiska värdet så förkastar man nollhypotesen.

(b) Om $X_1, \dots, X_n$ är ett stickprov från en fördelning med varians $\sigma^2$ så är variansen av medelvärdet lika med $\sigma^2/n$.

(c) Avståndet från den nedre kvartilen till medianen är alltid samma som avståndet från medianen till den övre kvartilen.

(d) Kvalitativ data kan illustreras med en boxplot.

(e) Om $X$ är $N(0,1)$-fördelad och $Y$ är $t$-fördelad med 10 frihetsgrader så gäller att $\mathbb{E}[X] = \mathbb{E}[Y]$.

(f) P-värdet är sannolikheten att nollhypotesen är sann.

(g) Ju högre signifikansnivå desto större är sannolikheten att vi förkastar nollhypotesen.

(h) Om vi skapar ett konfidensintervall för ett väntevärde så är det sanna väntevärdet alltid i mitten av intervallet.

**Lösning:**

(a) Falskt. p-värdet ska jämföras med signifikansnivån. Om man ska jämföra något med det kritiska värdet är det värdet på teststatistikan.

(b) Sant: $Var(\bar{X}) = Var(1/n \sum X_i) = 1/n^2 \sum Var(X_i) = \sigma^2/n$.

(c) Falskt. Det är alltid en lika stor andel (25%) mellan nedre kvartilen och medianen som mellan medianen och övre kvartilen, men avstånden är olika om fördelningen är skev.

(d) Falskt. En box plot illustrerar numeriska värden som medianen och övre/nedre kvartilerna. Dessa saknar mening för kvalitativ data.

(e) Sant. Båda är 0.

(f) Falskt. p-värdet är sannolikheten att observera något lika eller ännu mer extremt än vad testet gav OM nollhypotesen är sann. Påståenden angående 'sannolikheten att nollhypotesen är sann' är meningslösa, då den antingen är sann eller falsk – vi vet bara inte vilket.

(g) Sant. Signifikansnivån är precis sannolikheten att förkasta nollhypotesen om den är sann.

(h) Falskt. Det observerade medelvärdet är i mitten av konfidensintervallet, det sanna väntevärdet är okänt.

6. Egon och Lisa har två stickprov, $X_1, \dots, X_n$ respektive $Y_1, \dots, Y_n$, av samma storlek från en fördelning med väntevärde $\mu$ och varians $\sigma^2$, och är intresserade av att skatta $\mu$. Egon föreslår att de bara ska ta hans stickprov och använda $\hat{\mu} = \bar{X}$ som skattare. Lisa, däremot, föreslår att dom ska mötas på hälften och istället använda $\hat{\mu} = \frac{\bar{X}+\bar{Y}}{2}$.

(a) Visa att Lisas förslag också ger en väntevärdesriktig skattare. (3p)

(b) Vilken skattare skulle du säga är bäst? Motivera. (3p)

**Lösning:**

(a) Att en skattare är väntevärdesriktig innebär att väntevärdet av skattaren är lika med parametern vi vill skatta. I detta fall vill vi alltså visa att

$$\mathbb{E}\left[\frac{\bar{X}+\bar{Y}}{2}\right] = \mu.$$

På grund av lineariteten hos väntevärde får vi att

$$\mathbb{E}\left[\frac{\bar{X}+\bar{Y}}{2}\right] = \frac{\mathbb{E}[\bar{X}]+\mathbb{E}[\bar{Y}]}{2} = \frac{\mu+\mu}{2} = \mu$$

(b) Vi vill ha liten varians för vår punktskattare så vi borde jämföra varianserna. Vi vet att $\mathrm{Var}(\bar{X}) = \sigma^2/n$. Om vi antar att stickproven är oberoende (vilket är rimligt) har vi även att

$$\mathrm{Var}\left(\frac{\bar{X}+\bar{Y}}{2}\right) = \frac{\mathrm{Var}(\bar{X})+\mathrm{Var}(\bar{Y})}{2^2} = \frac{\sigma^2/n+\sigma^2/n}{4} = \frac{\sigma^2}{2n}$$

som är mindre än $\sigma^2/n$. Lisas skattare har alltså mindre varians och anses därför vara bättre.

7. I Sverige genomförs varje månad opinionsundersökningar av ett antal institut. I dessa undersökningar intervjuar man personer och ställer frågan "Vilket parti skulle du rösta på om det var riksdagsval idag?". I senaste riksdagsvaletvalet (2014) fick Vänsterpartiet (V) 5.72 % av rösterna, men i Aftonbladets senaste väljarbarometer svarade 134 av 2066 tillfrågade personer att de skulle rösta på (V).

(a) Ställ upp hypoteser för att testa om andelen som skulle rösta på (V) är högre idag än vid riksdagsvalet 2014, beräkna teststatistikan och $p$-värdet för testet. (Vi antar att valresultatet 2014 gav den sanna andelen av befolknigen som röstade på (V) vid den tidpunkten). (3p)

(b) Testa hypotesen i (a) vid signifikansnivå $\alpha = 0.1$. Vilken blir slutsatsen? (2p)

(c) Vilken typ av fel riskerar du att göra när du drar slutsatsen i (b)? (1p)

**Lösning**

(a) Hypoteser:

$$H_0 : p = 0.0572, \quad \text{mot } H_1 : p > 0.0572$$

Notera att vi vill testa om andelen ökat, därav enkelsidig $H_1$. (1p)

Teststatistikan: $z = \frac{\hat{p}-p_0}{\sqrt{(p_0(1-p_0)/n}} = 1.50$, där $p_0 = 0.0572$, $\hat{p} = 134/2066 = 0.065$ och $n = 2066$. (1p)

$p$-värdet $= P(Z > 1.5) = 0.0668$. (1p)

(b) P-värdet $= 0.0668$ är mindre än signifikansnivån $\alpha = 0.1$, alltså förkastar vi $H_0$ på denna signifikansnivå. (1p)

Det verkar som att andelen som skulle rösta på (V) har ökat. (1p)

(c) Vi förkastar $H_0$, så vi riskerar att göra typ I-fel, dvs förkasta $H_0$ när den är sann. (1p)

8. Simon vill jämföra väntevärdena hos två normalfördelade populationer med lika stora varianser. Till sin hjälp har han två oberoende stickprov (ett från vardera population) av storlek $n_1 = 12$ och $n_2 = 15$. Han har också räknat ut punktskattningarna för väntevärdet och variansen från de två stickproven: $\bar{x} = 24.3$, $\bar{y} = 26.4$, $s_X^2 = 2.3$ och $s_Y^2 = 2.9$.

(a) Hjälp Simon att göra ett 95%-igt konfidensintervall för skillnaden mellan populationernas sanna väntevärden. (4p)

(b) Simon är inte riktigt nöjd med konfidensintervallet du tar fram. Han säger "bättre att ta det säkra före det osäkra" och vill sedan göra ett 99.99%-igt konfidensintervall istället. "Då kan vi ju vara nästan helt säkra på att skillnaden mellan de sanna väntevärdena ligger i det nya konfidensintrevallet!" Förklara för Simon varför detta kanske inte är en bra idé. (1p)

**Lösning:**

(a) Eftersom $n_1 = 12$ och $n_2 = 15$ är för små så kan vi inte använda oss av normalfördelningen utan vi måste använda $t$-fördelningen istället. Formeln för konfidensintervallet kommer att ges av

$$L = (\bar{X} - \bar{Y}) - t_{\alpha/2, n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$
$$U = (\bar{X} - \bar{Y}) + t_{\alpha/2, n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

där $s_p^2$ är den poolade stickprovsvariansen som ges av

$$s_p^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}.$$

Allt vi behöver för att ta fram $s_p^2$ är givet i uppgiften: $s_p^2 = 2.636$. Från $t$-tabellen med $n_1 + n_2 - 2 = 25$ frihetsgrader tar vi fram värdet

$t_{0.025,25} = \pm 2.06$. Nu kan vi lätt ta fram det observerade konfidensintervallet

$$\ell = (24.3 - 26.4) - 2.06 * \sqrt{2.636 * \left(\frac{1}{12} + \frac{1}{15}\right)} = -3.40$$

$$u = (24.3 - 26.4) + 2.06 * \sqrt{2.636 * \left(\frac{1}{12} + \frac{1}{15}\right)} = -0.80.$$

Vårt observerade 95%-iga konfidensintervall för skillnaden $\mu_X - \mu_Y$ blev alltså $[-3.4; -0.8]$.

(b) Om man använde så låg signifikansnivå så kommer konfidensintervallet att blir alldeles för stort. Man kommer helt enkelt inte kunna dra några vettiga slutsatser från konfidensintervallet.