Tentamentsskrivning i **Matematisk Statistik med Metoder MVE490**

Tid: den 16 augusti, 2017
Examinatorer: Kerstin Wiklander och Erik Broman.
Jour: Claes Andersson (tel 0734031540).
Hjälpmedel: miniräknare, egenhändigt skriven formelsamling om två A4 fram och bak (dvs 4 sidor) samt tabeller (delas ut på plats).

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. Låt A, B vara händelser så att $\mathbb{P}(A) = 0.4$ och $\mathbb{P}(B) = 0.3$. (6p)

   (a) (2p) Om $\mathbb{P}(A \cap B) = 0.2$, vad blir då $\mathbb{P}(A \cup B)$?

   (b) (2p) Om A, B är oberoende, vad blir då $\mathbb{P}(A^c \cap B^c)$?

   (c) (2p) Antag igen att $\mathbb{P}(A \cap B) = 0.2$ och låt C vara sådan att $\mathbb{P}(C) = 0.6$ och A, C är disjunkta. Beräkna $\mathbb{P}(B \cap C)$

   Lösning:

   (a) Vi har att
   $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.4 + 0.3 - 0.2 = 0.5.$$

   (b) Oberoende ger att
   $$\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\mathbb{P}(B^c) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = 0.6 * 0.7 = 0.42.$$

   (c) Vi inser att $C = A^c$ så att
   $$\mathbb{P}(B \cap C) = \mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(B \cap A) = 0.3 - 0.2 = 0.1.$$

2. Ceclia har fått diverse frågor att besvara. Hjälp henne med detta: (7p)

   (a) (1p) Förklara kort vad det är för skillnad mellan en variabel och en parameter.

   (b) (1p) Ange vad man har för krav på p-värdet för att kunna förkasta en nollhypotes när signifikansnivån är vald till 1%.

(c) (1p) Man har samlat in en datamängd från en variabel som skall ha målvärdet 8.10. Värdena blev: 7.99 8.02 7.95 8.37 8.42. Beräkna en skattning av det systematiska felet (bias).

(d) (1p) Vad betyder det att styrkan i ett test är 80%?

(e) (3p) Ange för varje påstående nedan om det är fel eller om det är rätt.

e1) En signifikansnivå är detsamma som ett minus styrkan.
e2) Gör man två mätningar på samma individ kan man inte säga att värdena är oberoende.
e3) Man skall inte alltid illustrera kvantitativa (numeriska) data med ett histogram.
e4) Om man använder en kvantitativ (numerisk) variabel kan man använda normalfördelning som approximativ modell för medelvärdet om stickprovsstorleken är stor.
e5) Om man använder en kvalitativ (kategorisk) variabel kan man inte göra ett t-test.
e6) Längden på ett konfidensintervall blir kortare när stickprovsstorleken ökar.

Lösning:

a) Parameter; en karakteristika i en population (teoretiskt mått i en fördelning). Variabel; värde som varierar för olika objekt/individer (men kan beskrivas med hjälp av en fördelning).
b) Det skall vara ≤ 0.01.
c) Medelvärdet är 8.15. Det systematiska felet uppskattas till $8.15 - 8.10 = 0.05$. Det går lika bra att först beräkna avvikelserna från målvärdet och sedan medelvärdet av dessa värden.
d) Sannolikheten är 80% att korrekt förkasta en nollhypotes (sannolikheten att förkasta nollhypotesen givet att den är falsk).
e) Alla utom e1 är korrekta.

3. En slumpvariabel X kan anta värdena $1, 2, 4$ och 5 med respektive sannolikheter $\mathbb{P}(X = 1) = 0.1, \mathbb{P}(X = 2) = 0.2, \mathbb{P}(X = 4) = 0.3$ och $\mathbb{P}(X = 5) = 0.4$. (8p)

(a) (2p) Beräkna väntevärde och varians för X.

(b) (2p) Beräkna väntevärde och varians för $1/X$.

(c) (2p) Låt X_1, X_2 vara oberoende slumpvariabler med samma fördelning som X. Vad är sannolikheten att $X_1 + X_2$ blir ett jämnt tal?

(d) (2p) Med samma förutsättningar som i (c), vad är sannolikheten att $X_1 X_2 + X_2$ blir ett jämnt tal?

Lösning:

(a) Vi har att

$$\begin{aligned}\mathbb{E}[X] &= 1\mathbb{P}(X=1) + 2\mathbb{P}(X=2) + 4\mathbb{P}(X=4) + 5\mathbb{P}(X=5) \\ &= 0.1 + 2*0.2 + 4*0.3 + 5*0.4 = 0.1 + 0.4 + 1.2 + 2 = 3.7,\end{aligned}$$

och vidare är

$$\begin{aligned}\mathbb{E}[X^2] &= 1\mathbb{P}(X=1) + 2^2\mathbb{P}(X=2) + 4^2\mathbb{P}(X=4) + 5^2\mathbb{P}(X=5) \\ &= 0.1 + 4*0.2 + 16*0.3 + 25*0.4 = 0.1 + 0.8 + 4.8 + 10 = 15.7\end{aligned}$$

så att $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 15.7 - 3.7^2 = 2.01$.

(b) Vi har att

$$\begin{aligned}\mathbb{E}[1/X] &= 1\mathbb{P}(X=1) + \mathbb{P}(X=2)/2 + \mathbb{P}(X=4)/4 + \mathbb{P}(X=5)/5 \\ &= 0.1 + 0.2/2 + 0.3/4 + 0.4/5 = 0.355,\end{aligned}$$

och vidare är

$$\begin{aligned}\mathbb{E}[(1/X)^2] &= 1\mathbb{P}(X=1) + \mathbb{P}(X=2)/4 + \mathbb{P}(X=4)/16 + \mathbb{P}(X=5)/25 \\ &= 0.1 + 0.2/4 + 0.3/16 + 0.4/25 = \frac{739}{4000}\end{aligned}$$

så att

$$\mathrm{Var}(1/X) = \mathbb{E}[(1/X)^2] - \mathbb{E}[1/X]^2 = \frac{739}{4000} - (71/400)^2 = \frac{596}{10149} \approx 0.058725.$$

(c) Vi har att $X_1 + X_2$ blir ett jämnt tal om antingen båda är jämna eller båda är udda. Vi får då att

$$\begin{aligned}&\mathbb{P}(X_1 + X_2 \text{ jämnt }) \\ &\quad = \mathbb{P}(X_1 \text{ udda }, X_2 \text{ udda }) + \mathbb{P}(X_1 \text{ jämnt }, X_2 \text{ jämnt }) \\ &\quad = \mathbb{P}(X_1 \text{ udda })^2 + \mathbb{P}(X_1 \text{ jämnt })^2 = 0.5^2 + 0.5^2 = 0.5.\end{aligned}$$

Alternativt

$$\begin{aligned}&\mathbb{P}(X_1 + X_2 \text{ jämnt }) \\ &\quad = \mathbb{P}(X_1 = 1, X_2 = 1) + \mathbb{P}(X_1 = 1, X_2 = 5) + \mathbb{P}(X_1 = 5, X_2 = 1) + \mathbb{P}(X_1 = 5, X_2 = 5) \\ &\qquad + \mathbb{P}(X_1 = 2, X_2 = 2) + \mathbb{P}(X_1 = 2, X_2 = 4) + \mathbb{P}(X_1 = 4, X_2 = 2) + \mathbb{P}(X_1 = 4, X_2 = 4) \\ &\quad = 0.5^2 + 0.5^2 = 0.5.\end{aligned}$$

(d) Vi har att $X_1X_2 + X_2 = X_2(1 + X_1)$ är jämnt om antingen X_2 är jämnt eller X_1 är udda. Vi får att

$$\begin{aligned}&\mathbb{P}(X_2(1 + X_1) \text{ jämnt}) \\ &\quad = 1 - \mathbb{P}(X_2(1 + X_1) \text{ udda}) = 1 - \mathbb{P}(X_2 \text{ udda})\mathbb{P}(X_1 \text{ jämnt}) = 1 - 0.5^2 = 0.75.\end{aligned}$$

4. David har en normalfördelningsmodell och vill utföra ett test. (3p)

   (a) (1p) Mothypotesen formulerar han som $H_a : \mu \neq \mu_0$ där μ_0 är ett värde enligt nollhypotesen. Om han använder en teststatistika där σ^2 är känt, vad blir gränsen/gränserna till förkastelseområdet då signifikansnivån är vald till 5%?

   (b) (1p) Om mothypotesen istället är $H_a : \mu > \mu_0$, vad blir gränsen/gränserna till förkastelseområdet då?

   (c) (1p) Om David inte känner till σ^2 och utnyttjar data med stickprovsstorleken $n = 5$, vad blir gränserna till förkastelseområdet då med mothypotes som i (a)-uppgiften?

   Lösning:

   a) Gränserna är -1.96 och 1.96.
   b) Gränsen är 1.645.
   c) Gränserna är -2.776 och 2.776.

5. En vanlig kortlek består av 52 kort, och 13 av dessa väljs ut till en bridgehand. (5p)

   (a) (1p) Hur många bridgehänder finns det?

   (b) (2p) Hur många bridgehänder finns det som innehåller exakt fem spader och två klöver?

   (c) (2p) Vad är sannolikheten att handen innehåller exakt två klöver givet att handen innehåller exakt fem spader?

   Lösning:

   (a) Det finns $\binom{52}{13}$ händer.

   (b) Vi kan välja de fem spaderna på $\binom{13}{5}$ sätt, och sedan kan vi välja de två klöverna på $\binom{13}{2}$ sätt och till sist de resterande 6 korten på $\binom{26}{6}$ sätt. Sammanlagt får vi då

   $$\binom{13}{5}\binom{13}{2}\binom{26}{6}.$$

   (c) Låt $S = \#\text{spader}$ och $K = \#\text{klöver}$. Vi får att

   $$\begin{aligned}\mathbb{P}(K = 2 | S = 5) &= \frac{\mathbb{P}(S = 5, K = 2)}{\mathbb{P}(S = 5)} \\ &= \frac{\binom{13}{5}\binom{13}{2}\binom{26}{6}}{\binom{52}{13}} \frac{\binom{52}{13}}{\binom{13}{5}\binom{39}{8}} = \frac{\binom{13}{2}\binom{26}{6}}{\binom{39}{8}} \approx 0.29189.\end{aligned}$$

6. Anna har gjort en fuktstudie av spånskivor från en tillverkare genom att mäta ånggenomsläppligheten. Värdena (dygnflödet) från mätningarna blev: 8. 9 8.4 9.5 8.6 9.4 8.8 9.1 9.2
   Anna har tidigare haft leveranser från en tillverkare som hon varit nöjd med. Nu används medelvärdet från den, som var 8.5 g/m^2, som ett målvärde för den nya tillverkarens väntevärde. Normalfördelning anses vara en bra modell för variabeln. (8p)

   (a) (5p) Beräkna ett 95%-igt konfidensintervall för väntevärdet. Om syftet var att undersöka om den nya tillverkarens väntevärde avviker från förra, dra en slutsats enbart med hjälp av det uträknade intervallet. Ange också vilken signifikansnivå detta test har.

   (b) (3p) Anna tycker att en avvikelse på som mest 10% från den gamla tillverkarens medelvärde är acceptabelt. Visa hur hon kan ta ett beslut om nya tillverkaren med hjälp av konfidensintervallet. Vad blir din slutsats?

   Lösning:

   a) Låt Y = dygnflödet. Formeln för ett konfidensintervall:

   $$\bar{y} \pm t_{n-1,\alpha/2}\, s/\sqrt{n}$$

   Med de givna värdena blir det (avrundat) $8.91 \pm 2.365 \times 0.35/\sqrt{8}$ dvs $(8.6\,,\,9.2)$. Eftersom målvärdet 8.5 inte ingår i det 95%-iga intervallet kan $H_0 : \mu = 8.5$ förkastas på signifikansnivå 5%.
   b) Acceptabelt är inom intervallet $(7.65, 9.35)$. Eftersom hela konfidensintervallet ryms inom detta intervall anses avvikelsen vara acceptabel och inte signifikant ur ett mer praktiskt perspektiv. Den nya tillverkarens väntevärde i fuktstudien kan anses duga.

7. Charlie vill studera om folk upplever störningar i sitt boende. Lägenheter från bostadsområden i två olika delar av Göteborg har därför valts ut slumpmässigt. I dessa intervjuade han de boende bl a med frågan om de stördes av trafiken. Bland de 60 st i område 1 svarade 15 ja på detta och bland de 65 i område 2 var det 24 st som svarade ja. Syftet är att undersöka om proportionerna skiljer sig åt mellan bostadsområdena. (7p)

   (a) (5p) Charlie vill att du formulerar hypoteserna och utför testet.

   (b) (2p) Charlie vill också att du beräknar ett konfidensintervall för skillnaden mellan bostadsområdenas proportioner.

   Lösning:

   Låt p_1 och p_2 vara parametrarna i två oberoende binomialfördelningar och beteckna andelarna för de som störs av trafik.
   a) $H_0 : p_1 = p_2$ och $H_a : p_1 \neq p_2$. Skattningarna: $\hat{p}_1 = x_1/n_1 = 15/60 =$

0.25 och $\hat{p}_2 = x_2/n_2 = 24/65 \approx 0.37$. I testet med $H_0 : p_1 = p_2$ skattas det gemensamma $p_1 = p_2 = p$ med $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$. Teststatistikan

$$\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

(approximativt normalfördelad) fick värdet -1.44. Jämför detta med förkastelsegränserna $\pm z_{0.025} = \pm 1.96$. Eftersom teststatistikans värde inte har passerat någon av förkastelsegränserna kan H_0 inte förkastas. Man kan inte i denna undersökning påvisa skillnad i andelarna som störs av trafiken.

b) Till skillnad från fallet beräknat under förutsättningen $H_0 : p_1 = p_2 = p$ måste de två proportionerna skattas var för sig. Det gör man med relativa frekvenserna och standardavvikelsen har därför formeln

$$\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}$$

Ett 95%-igt konfidensintervall för $p_2 - p_1$ blir $(-0.04, 0.28)$.

8. Ett sågverk tillverkar brädor. Brädorna är av varierande kvalitet och graderas antingen som A-klass,B-klass eller C-klass.

   Sågverket antar att sannolikheten för att en given bräda blir A-klass är 0.2, B-klass 0.5 och C-klass 0.3. Låt X_A vara antalet brädor som blir A-klass om sågverket en dag tillverkar 2300 brädor. (6p)

   (a) (2p) Hitta en lämplig fördelning för X_A. Vilka antaganden måste du göra? Är dessa antaganden realistiska?

   (b) (2p) Beräkna sannolikheten att X_A blir 420 eller mindre.

   (c) (2p) En fjärde klassificering införs som kallas A^+. Sannolikheten att en given bräda skall få denna klassificering är 0.005. Låt X^+ vara antalet brädor som får denna klassificering av de 2300 som tillverkas. Ange en lämplig approximativ fördelning för X^+ och använd denna för att beräkna sannolikheten att $X^+ = 15$.

   Lösning:

   (a) Vi har att $X_A \sim$Bin$(2300, 0.2)$ under antagandet att alla brädor klassificeras som A oberoende av varandra. Detta antaganden är rimligt, även om man kan tänka sig att eventuella defekter delas av andra brädor som är gjorda av samma träd. Då antalet tillverkade brädor är så stort kommer felet från vårt antagande vara litet.

   (b) Approximativt får vi att

   $$X_A \approx N\left(\mathbb{E}[X], \mathrm{Var}(X)\right) = N(np, np(1-p)) = N(460, 368)$$

   så att

   $$\begin{aligned}\mathbb{P}(X_A \leq 420) &= \mathbb{P}\left(\frac{X_A - 460}{\sqrt{368}} \leq \frac{420-460}{\sqrt{368}}\right) \approx \mathbb{P}(Z \leq -2.0851)\\ &= \mathbb{P}(Z \geq 2.0851) = 1 - \mathbb{P}(Z \leq 2.0851) \approx 1 - 0.981 \approx 0.019.\end{aligned}$$

(c) Vi använder att $X^+ \sim$Poi(λ) där $\lambda = 2300 * 0.005 = 11.5$ vilket är lämpligt då p är väldigt litet. Med hjälp av detta får vi att

$$\mathbb{P}(X^+ = 15) = \frac{11.5^{15}}{15!} e^{-15} \approx 0.0019.$$