

TMS136: Dataanalys och statistik – Tentamen 2013-10-26 med lösningar

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:

- (a) Bias. (1p)

Lösning: Om $\hat{\theta}$ är en skattare (inte skattning) för parametern θ ges skattarens bias av $E(\hat{\theta}) - \theta$. Bias är alltså ett mått hur stor en eventuell systematisk avvikelse är i medeltal.

- (b) Styrka. (1p)

Lösning: Sannolikheten att förkasta en falsk nollhypotes. Ett mått på ett hypotestests förmåga att upptäcka avvikelser från nollhypotesen.

- (c) Typ I-fel. (1p)

Lösning: Att förkasta en sann nollhypotes.

- (d) Konfidensgrad för ett tvåsidigt konfidensintervall. (1p)

Lösning: Om ett tvåsidigt konfidensintervall har konfidensgraden $(1 - \alpha)100\%$ kommer upprepade observationer konfidensintervallet täcka den aktuella populationsparametern i $(1 - \alpha)100\%$ av fallen.

2. I en mörk garderob ligger n olika par skor huller om buller. Du sträcker in handen och tar två skor på måfå. Lös följande både för fixt n och då $n \rightarrow \infty$:

- (a) Sannolikheten att få ett sammanhörande par? (2p)

Lösning: Att välja två skor bland $2n$ skor kan göras på $\binom{2n}{2}$ sätt så det finns $\binom{2n}{2}$ möjliga utfall. Det finns precis n sammanhörande par så ett sammanhörande par kan väljas på precis n sätt. Alltså är antalet gynnsamma utfall n . Det följer att sannolikheten för att få ett sammanhörande par är

$$\frac{n}{\binom{2n}{2}} = \frac{n}{\frac{(2n)!}{2!(2n-2)!}} = \frac{n}{\frac{2n(2n-1)}{2}} = \frac{1}{2n-1} \rightarrow 0, n \rightarrow \infty.$$

Så för ett fixt n är sannolikheten $\frac{1}{2n-1}$ och sannolikheten går mot 0 då n växer.

- (b) Sannolikheten att få en högersko och en vänstersko? (2p)

Det finns n vänsterskor och n högerskor så antalet möjligheter att välja en av varje är n^2 . Det finns alltså n^2 gynnsamma utfall. Det följer att den sökta sannolikheten är

$$\frac{n^2}{\binom{2n}{2}} = \frac{n^2}{\frac{(2n)!}{2!(2n-2)!}} = \frac{n^2}{\frac{2n(2n-1)}{2}} = \frac{n}{2n-1} = \frac{1}{2 - \frac{1}{n}} \rightarrow \frac{1}{2}, n \rightarrow \infty.$$

Så för ett fixt n är sannolikheten $\frac{1}{2 - \frac{1}{n}}$ och sannolikheten går mot $\frac{1}{2}$ då n växer.

3. Före en opinionsundersökning kring hur man skulle rösta om det vore val idag förmodar man att 36% skulle rösta på Socialdemokraterna, 25% skulle rösta på Moderaterna, 12% skulle rösta på Folkpartiet och resterande skulle rösta på de mindre partierna eller blankt. Det visade sig bland 10000 tillfrågade att 3727 personer skulle rösta på Socialdemokraterna, 2418 skulle rösta på Moderaterna, 1223 skulle rösta på Folkpartiet och att resterande skulle rösta på de mindre partierna eller blankt. Finns det anledning att ändra den förmodade fördelningen? (4p)

Lösning: Här är det lämpligt att göra ett Goodness-of-fit-test då det handlar om hur många enheter som hamnar i olika kategorier. Då det är 10000 som tillfrågats följer att de förväntade frekvenserna är 3600, 2500, 1200 och 2700 för Socialdemokraterna, Moderaterna, Folkpartiet respektive övriga partier eller blankt. Vi får då att den observerade testfunktionen

$$\sum \frac{(O - E)^2}{E} = \frac{(3727 - 3600)^2}{3600} + \frac{(2418 - 2500)^2}{2500} + \frac{(1223 - 1200)^2}{1200} + \frac{(2632 - 2700)^2}{2700} \approx 9.32.$$

Eftersom vi inte skattat några parametrar är antalet frihetsgrader 3. Enligt chi-två-tabell är det kritiska värdet på 0.05-nivå 7.815 så på den nivån skulle vi förkasta nollhypotesen om att den förmodade fördelningen stämmer och har således skäl att ändra den förmodade fördelningen. Skulle vi exempelvis välja att göra testet på signifikansnivån 0.01 skulle det kritiska värdet vara 11.345 och då skulle man inte förkasta nollhypotesen.

4. När man överför information i form av ettor och nollor på en brusig kanal händer ibland att skickade nollor tas emot som ettor och skickade ettor tas emot som nollor. Sannolikheten att en skickad nolla tas emot som en etta är δ och sannolikheten att en skickad etta tas emot som en nolla är ϵ . Man skickar ungefär lika många nollor som ettor. Låt X vara det som sänds och Y vara det som tas emot.

(a) Bestäm massfunktionen för Y (3p)

Lösning: Ur texten fås $P(Y = 1|X = 0) = \delta$, $P(Y = 0|X = 1) = \epsilon$ och $P(X = 1) = P(X = 0) = \frac{1}{2}$. Enligt definitionen av betingad sannolikhet följer att

$$P(Y = 1) = P(Y = 1|X = 0)P(X = 0) + P(Y = 1|X = 1)P(X = 1) = \delta \frac{1}{2} + (1 - \epsilon) \frac{1}{2} = \frac{1 - \epsilon + \delta}{2}.$$

Vi får då att

$$P(Y = 0) = 1 - P(Y = 1) = \frac{1 - \delta + \epsilon}{2},$$

och det följer att massfunktionen för Y är

$$f(y) = \left(\frac{1 - \epsilon + \delta}{2} \right)^y \left(\frac{1 - \delta + \epsilon}{2} \right)^{1-y}$$

(b) Bestäm korrelationen mellan X och Y (3p)

Lösning: Korrelationen mellan X och Y ges av

$$\frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}.$$

Vi har att

$$E(XY) = P(X = 1, Y = 1) = P(Y = 1|X = 1)P(X = 1) = \frac{1 - \epsilon}{2},$$

$$E(X) = \frac{1}{2}, \quad E(Y) = \frac{1 - \epsilon + \delta}{2},$$

$$V(X) = E(X^2) - (E(X))^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

och att

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{1 - \epsilon + \delta}{2} - \left(\frac{1 - \epsilon + \delta}{2} \right)^2.$$

Det följer att korrelationen mellan X och Y är

$$\frac{\frac{1-\epsilon}{2} - \frac{1}{2} \frac{1-\epsilon+\delta}{2}}{\sqrt{\frac{1}{4} \left(\frac{1-\epsilon+\delta}{2} - \left(\frac{1-\epsilon+\delta}{2} \right)^2 \right)}} = \frac{1 - \epsilon - \delta}{\sqrt{1 - (\epsilon - \delta)^2}}$$

5. Tiderna mellan kunders ankomster till kassa 8 på ICA Maxi är oberoende och exponentialfördelade med väntevärde 45 sekunder. Antag att en kund just kom till kassan. Vad är sannolikheten att det under under nästkommande fyra timmar kommer åtminstone 330 kunder till kassan? (4p)

Lösning: Att tiderna mellan kundernas ankomster är oberoende och exponentialfördelade med väntevärde 45 sekunder betyder att antalet kunder som kommer till kassan under 45 sekunder är Poissonfördelade med väntevärde $\lambda = 1$. Det följer att antalet kunder som kommer till kassan under fyra timmar är Poissonfördelat med väntevärde $\lambda = 4 \cdot 3600/45 = 320$. Vi har då med hjälp av CGS, om X är antalet kunder under fyra timmar, att

$$P(X \geq 330) \approx P\left(Z \geq \frac{330 - 320}{\sqrt{320}}\right) \approx 1 - \Phi(0.56) \approx 0.29.$$

Sannolikheten att åtminstone 330 kunder kommer till kassan under fyra timmar är alltså ungefär 29 procent.

6. Från två oberoende grupper av bilar har bränsleförbrukningen i liter per mil mätts upp med följande stickprovsresultat
- (a) 0.92, 1.01, 0.86, 0.95, 1.03, 1.12, 0.91, 0.86
- (b) 0.71, 0.67, 0.72, 0.63, 0.58, 0.65, 0.59, 0.71, 0.57

Är detta tillräckligt för att på signifikansnivån 0.05 och under antagande om lika populationsvarianser säga att den normalfördelade grupp(b) har lägre förväntad bränsleförbrukning än den normalfördelade grupp(a)? (4p)

Lösning: Vi låter indexet 1 beteckna gruppen från vilken stickprovet av storlek $n_1 = 8$ kommer och 2 beteckna gruppen från vilken stickprovet av storlek $n_2 = 9$ kommer. Vi ska testa hypoteserna

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0.$$

Då vi kan anta att populationsvarianserna är lika använder vi testfunktionen

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

där $\bar{X}_1$ är medelvärdet för stickrpovet från grupp 1, $\bar{X}_2$ är medelvärdet för stickrpovet från grupp 2, Δ_0 är nollhypotesens värde på skillnaden i populationsmedelvärden $\mu_1 - \mu_2$ och S_p är den poolade stickprovsstandardavvikelsen som ges av

$$S_p = \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}}.$$

Vi observerar $\bar{x}_1 = 0.96$, $\bar{x}_2 = 0.65$, $s_1^2 = 0.0082$ och $s_2^2 = 0.0035$ så

$$s_p = \sqrt{\frac{7 \cdot 0.0082 + 8 \cdot 0.0035}{15}} = 0.075,$$

vilket, med $\Delta_0 = 0$ ger den observerade testfunktionen

$$\frac{0.96 - 0.65}{0.075\sqrt{\frac{1}{8} + \frac{1}{9}}} = 8.50$$

De kritiska värdet är $t_{0.05,15} = 1.753$ så vi kan förkasta nollhypotesen. Så bilgruppen varifrån stickprovet av storlek nio kommer har alltså en signifikant lägre medelförbrukning än bilgruppen varifrån stickprovet av storlek åtta kommer.

7. I VLE-uppgifterna har vi sett att man givet ett stickprov $X_1, \dots, X_n$ från en likformig fördelning på intervallet $[0, b]$ kan skatta ett okänt b med exempelvis ML-skattaren $\max(X_1, \dots, X_n)$ eller momentskattaren $2\bar{X}$. Finn ett argument som talar för att använda momentskattaren över ML-skattaren. Finn minst två argument som talar för att använda ML-skattaren över momentskattaren. Argumenten måste motiveras väl. (5p)

Lösning: Momentskattaren är unbiased eftersom

$$E(2\bar{X}) = 2E(\bar{X}) = \frac{2}{n} \sum_{i=1}^n E(X_i) = \frac{2}{n} n \frac{b}{2} = b.$$

ML-skattaren är inte unbiased eftersom

$$E(\max(X_1, \dots, X_n)) = \frac{n}{b^n} \int_0^b x^{n-1} = \frac{nb}{n+1}$$

där vi fått tätheten, $f(x) = \frac{nx^{n-1}}{b^n}$ för $0 < x < b$ och 0 annars, för ML-skattaren som derivatan m.a.p. x av dess fördelningsfunktion

$$F(x) = P(\max(X_1, \dots, X_n) \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = (P(X_i \leq x))^n = \left(\frac{x}{b}\right)^n, 0 < x < b$$

och $F(x) = 0$ annars. Så ur "biassynpunkt" är momentskattaren bättre än ML-skattaren. ML-skattaren är bättre än momentskattaren eftersom den aldrig kan ge den orimliga skattningen att b skulle vara mindre än en av stickprovsmedlemmarna. ML-skattaren är bättre än momentskattaren eftersom MSE för ML-skattaren är

$$E\left(\max(X_1, \dots, X_n) - b\right)^2 = E\left((\max(X_1, \dots, X_n))^2\right) - 2b \frac{nb}{n+1} + b^2,$$

där

$$E\left((\max(X_1, \dots, X_n))^2\right) = \frac{n}{b^n} \int_0^b x^{2n-1} = \frac{nb^2}{n+2},$$

så MSE för ML-skattaren blir

$$\frac{nb^2}{n+2} - \frac{2nb^2}{n+1} + b^2 = \frac{2b^2}{(n+2)(n+1)},$$

medan MSE för momentskattaren är

$$E\left((2\bar{X} - b)^2\right) = 4E\left((\bar{X})^2\right) - 2bE(2\bar{X}) + b^2 =$$

$$4V(\bar{X}) + 4(E(\bar{X}))^2 - 2b^2 + b^2 = 4\frac{b^2}{12n} + 4\frac{b^2}{4} - b^2 = \frac{b^2}{3n}$$

så kvoten av MSE för ML-skattaren och MSE för momentskattaren blir

$$\frac{6n}{(n+2)(n+1)} \leq 1, n = 1, 2, \dots$$

med likhet för $n = 1$, $n = 2$ och olikhet för $n > 2$ så i alla praktiskt relevanta situationer har ML-skattaren lägre MSE än momentskattaren.

8. Låt $X_1, \dots, X_n$ vara ett stickprov från en fördelning med täthetsfunktionen

$$f(x) = \frac{x}{\theta} e^{-x^2/(2\theta)}, x \geq 0, \theta > 0.$$

- (a) Bestäm ML-skattaren för θ . (2p)

Lösning: Likelihood-funktionen ges av

$$L(\theta) = \prod_{i=1}^n \frac{x_i}{\theta} e^{-x_i^2/(2\theta)} = \frac{\prod_{i=1}^n x_i}{\theta^n} e^{-\frac{1}{2\theta} \sum_{i=1}^n x_i^2}.$$

Det är behändigare att maximera log-likelihood-funktionen istället så vi tar fram den;

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^n \ln x_i - n \ln \theta - \frac{1}{2\theta} \sum_{i=1}^n x_i^2.$$

Vi deriverar $l(\theta)$ och får ekvationen

$$-\frac{n}{\theta} + \frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 = 0,$$

och det följer att ML-skattaren är $\hat{\theta} = \frac{1}{2n} \sum_{i=1}^n X_i^2$.

- (b) Man kan visa att $E(X_i) = \sqrt{\frac{\theta\pi}{2}}$ och $V(X_i) = \frac{4-\pi}{2}\theta$. Vad blir ML-skattarens bias? (2p)

Lösning: Vi har att

$$E\left(\frac{1}{2n} \sum_{i=1}^n X_i^2\right) = \frac{1}{2n} \sum_{i=1}^n E(X_i^2) = \frac{1}{2} E(X_i^2) =$$

$$\frac{1}{2} \left(V(X_i) + (E(X_i))^2 \right) = \frac{1}{2} \left(\frac{4-\pi}{2}\theta + \frac{\theta\pi}{2} \right) = \theta,$$

så ML-skattaren är unbiased.

Lycka till!