STOCKHOLMS UNIVERSITET
Statistiska institutionen
Ellinor Fackle-Fornius

TENTAMEN I STATISTISK TEORI MED TILLÄMPNINGAR II
2017-04-18

Skrivtid: 10.00-15.00
Godkända hjälpmedel: Miniräknare, språklexikon.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
Genomgång av tentamen sker 2017-05-09 kl. 15 i sal D207.

OBS! Glöm inte att ange nödvändiga antaganden överallt.

Uppgift 1. (20 poäng)

Avgör om vart och ett av följande påståenden är sanna eller falska samt motivera ditt svar.

i) Om p-värdet vid ett hypotestest är 0.028 kan nollhypotesen förkastas på signifikansnivån $\alpha = 0.01$.

ii) Styrkan hos ett hypotestest beräknas alltid givet att nollhypotesen är sann.

iii) Signifikansnivån α är sannolikheten att nollhypotesen är falsk.

iv) Om p-värdet är väldigt litet vid ett hypotestest för att jämföra två populationsmedelvärden måste skillnaden mellan medelvärdena vara stor.

v) Vid ett likelihoodkvottest är det möjligt att $L(\widehat{\Omega}_0) > L(\widehat{\Omega})$.

Uppgift 2. (20 poäng)

En stor livsmedelskedja är intresserad av andelen kunder som är positivt inställda till att använda självscanning när de handlar. I ett slumpmässigt urval av 120 kunder svarade 27 st att de var positivt inställda till att använda självscanning. Av de 120 kunderna i urvalet var 84 st medlemmar i kedjans kundklubb, varav 20 st var positivt inställda. Av de resterande 36 kunderna i urvalet (som alltså var icke-medlemmar) var 7 st positivt inställda.

a) Beräkna ett 95 %-igt konfidensintervall för andelen kunder som är positivt inställda till att använda självscanning när de handlar.

b) Hur stort urval skulle krävas om man vill att längden av ett 95 %-igt konfidensintervall för andelen kunder som är positivt inställda till att använda självscanning när de handlar inte ska överstiga 0.1.

c) Beräkna ett 95 %-igt konfidensintervall för skillnaden mellan andelen kundklubbsmedlemmar och andelen icke-medlemmar som är positivt inställda till att använda självscanning när de handlar.

Uppgift 3. (20 poäng)

I en fabrik utvärderas två typer av maskiner, A och B, genom ett experiment. Tiden det tar att utföra en viss uppgift kontrolleras för varje typ av maskin. Åtta maskintekniker är involverade i experimentet, varje tekniker använder maskin A och maskin B i slumpmässig ordning. Tiden i sekunder som det tog att slutföra uppgiften i respektive maskin visas i följande tabell.

Tekniker	Maskin A	Maskin B
1	32	30
2	40	39
3	42	42
4	26	23
5	35	36
6	29	27
7	45	41
8	22	21

a) Testa med ett lämpligt icke-parametriskt test om det går att påvisa någon skillnad mellan tiden det tar för maskin A och maskin B att utföra uppgiften.

b) Testa med ett lämpligt parametriskt test om det går att påvisa någon skillnad mellan tiden det tar för maskin A och maskin B att utföra uppgiften.

c) I det här experimentet användes samma maskintekniker för båda maskiner, är det en fördel eller nackdel?

Uppgift 4. (20 poäng)

Antal gånger som en student på en grundkurs i statistik behöver tentera för att få ett godkänt betyg på kursen antas följa den geometriska fördelningen

$$p(y) = p(1-p)^{y-1}, \quad y = 1, 2, \ldots$$

där y är antal försök och p är sannolikheten att få ett godkänt betyg. I ett slumpmässigt urval av 8 studenter blev antalet gånger som var och en behövde tentera:

$$1, 1, 1, 2, 3, 1, 1, 1$$

a) Härled momentskattningen $\widehat{p}_{MOM}$ och beräkna skattningen för urvalet.

b) Härled maximumlikelihood-skattningen $\widehat{p}_{ML}$ och beräkna skattningen för urvalet.

c) Vad blir maximumlikelihood-skattningen för väntevärdet för antal gånger som man behöver tentera för att få ett godkänt betyg.

Uppgift 5. (20 poäng)

Låt Y_1, Y_2, Y_3 vara ett slumpmässigt urval från

$$f(y) = \begin{cases} \frac{2y}{\theta^2}, & 0 \leq y \leq \theta \\ 0, & \text{annars.} \end{cases}$$

$\widehat{\theta} = \frac{Y_1+Y_2+Y_3}{2}$ är en estimator för θ.

a) Bestäm väntevärdet för $\widehat{\theta}$. Är $\widehat{\theta}$ väntevärdesriktig?

b) Bestäm variansen för $\widehat{\theta}$.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 18/4-2017

Sal: Brunnsvikssalen

Tenta: Statistisk teori med tillämpningar II

Kurs: Statistisk teori med tillämpningar

ANONYMKOD: STM-6032

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					7 8
Lär.ant. 20	20	19	19	20					

98 + 8 bonus

POÄNG	BETYG	Lärarens sign.
106	A	

SU, STATISTIK

Skrivsal: Brunnsviksalen **Anonymkod:** STM-0032 **Blad nr:** 1

(1) i.)

Falskt. P-värdet är sannolikheten att erhålla det observerade värdet eller ett mer extremt värde. P-värdet är den lägsta signifikansnivå som nollhypotesen kan förkastas på. Eftersom $0,01 < 0,029$ kan nollhypotesen inte förkastas på signifikansnivån $\alpha = 0,01$. Däremot kan nollhypotesen förkastas på signifikansnivå 0,05 tex. 4

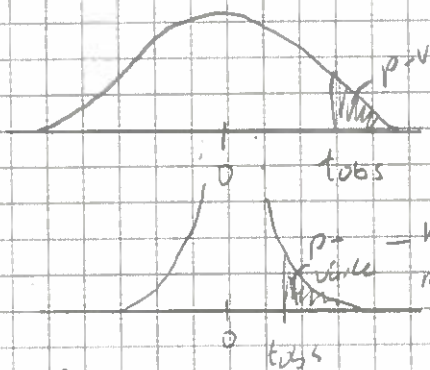
ii.) Falskt. Styrkan = Power(Θ) är $1-\beta$. β är sannolikheten för att ej förkasta H_0 givet att H_a är sann. $\beta = P(\text{ej förkasta } H_0 \mid H_a \text{ sann})$. Styrkan är sannolikheten att förkasta H_0 om H_a är sann. $\text{Power}(\Theta) = P(\text{förkasta } H_0 \mid H_a \text{ sann})$. För att beräkna styrkan måste man veta H_a och förkastelseområdet. 4

iii.) Falskt. Signifikansnivån α är sannolikheten att förkasta en sann nollhypotes: $P(\text{förkasta } H_0 \mid H_0 \text{ sann})$ och är samma sak som ett Typ I-fel. 4

iv. Falskt. Ett litet p-värde kan även fås om man har många observationer eller liten standardavvikelse. Med exempelvis ett t-test: $T = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ Stor skillnad i medelvärden ger ett stort T-värde,

med även en liten standardavvikelse och/eller stora urval ger ett stort T-värde och därmed mindre p-värde.

ex.



- stora standardavvikelse – större skillnad i medelvärdena
- mindre standardavvikelse – mindre skillnad mellan medelvärdena

4

v) Falskt:

$$\lambda_{LR} = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})}$$

$L(\hat{\Omega}_0)$ är likelihooden för H_0 $\theta \in \hat{\Omega}_0$

$L(\hat{\Omega})$ är likelihooden för den bästa av alla modeller. $0 < \lambda_{LR} < 1$ Om λ_{LR} är nära 1 är $L(\hat{\Omega}_0)$ ungefär lika stor som $L(\hat{\Omega})$ och vi kan ej förkasta H_0, då H_0 är den bästa modellen. Om λ_{LR} är liten betyder det att likelihooden för $L(\hat{\Omega})$ är större och det betyder en bättre modell än H_0. Då förkastas H_0.

4

20

Bra!

(2) p- "andelen kunder som är positivt inställda till självskanning"

$n = 120$ kunder

Totalt var 27 kunder positiva

84 st var medlemmar i kundklubben varav 20 positiva.

Av de 36 icke-medlemmarna var 7 positiva.

a) Beräkna 95%-igt konfidensintervall för andelen kunder som är positivt inställda till självscanning.

$$p = \frac{\text{antalet positiva}}{\text{totalt antal}} = \frac{27}{120} = 0{,}225$$

K.I.: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Vi kan anta att andelen är normalfördelad pga det stora urvalet enligt CGS. Därmed kan vi användas $z_{\alpha/2}$ till konfidensintervallet. (Tumregel $np > 5$, $n(1-p) > 5$) OK!

Vi antar att urvalet består av oberoende observationer.

$z_{\alpha/2} = 1{,}96 \Rightarrow$

$$0{,}225 \pm 1{,}96\sqrt{\frac{0{,}225(1-0{,}225)}{120}} \Leftrightarrow 0{,}225 \pm 0{,}0747$$

Svar: Ett 95%-igt konfidensintervall är då $0{,}225 \pm 0{,}0747$ vilket $= [0{,}1503 ; 0{,}2997]$ för andelen kunder som är positivt inställda.

→

b) Hur stort urval skulle krävas om längden på det 95 %-iga k.i $\leq 0,1$

$$\hat{p} \pm \underbrace{z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}}_{B}$$ R

Längden på konfidensintervallet är $2B = 0,1$

$\Rightarrow B = 0,05$

$\Rightarrow z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0,05$ R

Om vi har anledning att tro att p ska vara ett särskilt värde kan vi använda det, men för säkerhets skull använder vi $p = 0,5$ för att uppskatta variansen, då detta ger det största värdet på variansen och därmed kan vi vara säkra på att erhålla ett värde på B som stämmer med det givna villkoret för det n vi räknar ut. Bra

$z_{\alpha/2} = z_{0,025} = 1,96$ $\quad p = 0,5$

$$1,96 \cdot \sqrt{\frac{0,5 \cdot 0,5}{n}} = 0,05$$

$$\sqrt{\frac{0,25}{n}} = \frac{0,05}{1,96} \qquad \sqrt{n} = \frac{1,96}{0,05}\sqrt{0,25} = 19,6$$

$\Rightarrow n = 384,16$

Svar: Det skulle krävas 385 personer i urvalet.

R 7

SU, DEPARTMENT OF STATISTICS

Room: Brunsviksvulen Anonymous code: STM-0032 Sheet number: 3

(2) c) Andelen kundklubbsmedlemmar som är positivt inställda = $p_1 = \frac{20}{84} = 0,2381$ $n_1 = 84$

Andelen icke-medlemmar som är positivt inställda:

$p_2 = \frac{7}{36} = 0,1944$ $n_2 = 36$

Samma antaganden som i a)

Skillnaden i andelen positiva: $p_1 - p_2 =$

$0,2381 - 0,1944 = 0,04365$

95 %-ligt konfidensintervall för skillnaden

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\, \sigma(\hat{p}_1 - \hat{p}_2) \Rightarrow \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$ R

$$0,0436 \pm 1,96 \sqrt{\frac{0,238(1-0,238)}{84} + \frac{0,1944(1-0,1944)}{36}}$$

$0,0436 \pm 1,96 \cdot 0,0806$

$0,0436 \pm 0,1581$

Svar: Ett 95 %-igt konfidensintervall för skillnaden i andelen som är positivt inställda i de båda grupperna ges av $0,0436 \pm 0,158$ eller $[-0,114; 0,202]$ R

7

20

(3) Maskinerna A och B utvärderas med ett experiment. Man mäter tiden det tar att sköta en viss uppgift. Åtta maskintekniker utför ett test på maskin A och B.

Tekniker	Maskin A	Maskin B	D_i	Rang
1	32	30	2	4,5
2	40	39	1	2
3	42	42	0	tie
4	26	23	3	6
5	35	36	−1	(2)
6	29	27	2	4,5
7	45	41	4	7
8	22	21	1	2

n=7 pga en tie

a) Testa med icke-parametriskt test, om det går att påvisa någon skillnad mellan tiden med A eller B.

Jag väljer Wilcoxon Signed rank test.

Antaganden: Parvisa observationer, Seroende observationer mellan varje försök som utförs av samma tekniker. Oberoende observationer mellan de olika teknikernas försök. Antas symmetriska fördelningar. ~~fixir~~

Hypoteser:

H_0: Fördelningarna är lika

H_1: Fördelningarna skiljer sig i läge

Teststatistika $T_{min}(T^-, T^+)$

Jag väljer signifikansnivå $\alpha = 0{,}05$

Förkasta H_0 om $T = \min(T^-, T^+) \leq T_0 = 2$ (tabell 9)

[Dubbelsidigt test, $p = 0{,}05$ $n = 7$ $= T_0 = 2$] R

$T^- = 2$

$T^+ = 26$ R

$T = \min(2, 26] = 2$

$T^- = 2 \leq 2 = T_0$

Vi kan precis förkasta H_0 på signifikansnivå 0,05.

Svar: Vi kan förkasta H_0 på signifikansnivå 0,05.
Fördelningarna skiljer sig åt i läge, dvs det är skillnad mellan de olika maskinerna i tid att utföra uppgiften. 8

5) Utöver antagandena i a) antar vi att felen är normalfördelade och utför ett t-test då σ är okänd och det är i ett urval (CGS gälle inte)
Parvisa observationer.

Hypoteser: $H_0: \underline{D} = 0$ $\alpha = 0{,}05$ V
$H_1: \underline{D} \neq 0$ μ_0

Teststatistika: $T = \frac{\bar{D} - 0}{S/\sqrt{n}} \sim t(n-1) = t(7)$ R

$\bar{D} = 1{,}5$ R

$S^2 = \frac{\sum_{i=1}^{8} D_i^2 - n\bar{D}^2}{n-1} = \frac{36 - 8 \cdot 1{,}5^2}{7} = 2{,}5714$ R

Förkasta H_0 om $|T_{obs}| > T_{krit} = 2{,}36$ (tabell 3, $f_5 = 7$, $\alpha = 0{,}025$) R

forts. på nästa sida

(3) 6 forts.)

$$T_{obs} = \frac{1,5}{\sqrt{257/8}} = 2,64597$$

$T_{obs} = 2,65 > 2,36 = T_{krit}$

$\Rightarrow$ Förkasta H_0

Svar H_0 förkastas på signifikansnivå 0,05
Det finns belägg för att det är skillnad mellan tiden det tar att utföra uppgiften med de olika maskinerna.

7

c)
För att utföra ett parvis test behöver det vara samma person som utför testen. Olika personer arbetar på olika snabbt. Eftersom det skiljer sig genom mycket åt mellan de olika teknikerna, hur lång tid processen tar, från 22 till 45, så påverkar teknikerna väldigt mycket.
Genom att samma tekniker gör processen på både maskin A och B är det endast skillnaden i maskinerna som utgör skillnaden i tiden för varje parvis test. Man skulle också kunna göra ett test där samma tekniker gör många test på maskin A och B och göra ett Mann Whitney U-test / olika tekniker

4

(4)

Y - "antal gånger en student behöver tentera för att få godkänt"

$$p(y) = p(1-p)^{y-1} \qquad y = 1, 2, \ldots$$

Detta är en geometrisk fördelning

p - sannolikheten att få godkänt

Väntevärde för Y i en geometrisk fördelning: $\frac{1}{p}$

$n = 8$ studenter Resultat $Y: 1, 1, 1, 2, 3, 1, 1, 1$

a) Härled momentskattningen $\hat{p}_{mom}$ och beräkna skattningen för urvalet

Momentskattningen görs genom att populationsmomentet μ_1' antas kunna uppskattas med urvalsmomentet m_1'

$$\mu_1' = \frac{1}{p} \qquad m_1' = \frac{\sum Y_i}{n} = \bar{Y} \quad R \qquad R$$

$$\mu_1' = m_1' = \frac{\sum Y_i}{n} = \frac{1}{p} \Rightarrow p = \frac{n}{\sum Y_i} = \frac{1}{\bar{Y}}$$

$$\hat{p}_{mom} = \frac{1}{\bar{Y}} \quad R \qquad \bar{Y} = \frac{1+1+1+2+3+1+1+1}{8} = 1{,}375$$

$$\hat{p}_{mom} = \frac{1}{1{,}375} = 0{,}727 \quad R$$

Sue $\hat{p}_{mom} = \frac{1}{\bar{Y}}$ och får 0,727 med detta urval.

b) Härled maximum likelihood-skattningen $\hat{p}_{ML}$

$$p(y) = p(1-p)^{y-1}$$

$$L(p) = \prod_{i=1}^{n} p(1-p)^{y_i-1} = \{\text{pga oberoende}\} =$$

$$p^n \cdot (1-p)^{\sum y_i - n} \qquad \left(\sum y_i = \sum_{i=1}^{n} y_i\right) \quad R$$

logaritmera:

$$l(p) = \ln\left(p^n (1-p)^{\sum y_i - n}\right) = n \cdot \ln p + \left(\sum y_i - n\right)\ln(1-p)$$

$$= n \ln p + \sum y_i \ln(1-p) - n \ln(1-p) \quad R$$

Derivera med avseende på p:

$$\frac{dl(p)}{dp} = \frac{n}{p} + \frac{\sum y_i}{1-p}(-1) - \frac{n}{(1-p)}(-1) = \frac{n}{p} - \frac{\sum y_i}{1-p} + \frac{n}{1-p} \quad R$$

Sätt derivatan $= 0$ för att hitta maxvärdet

$$\frac{n}{p} - \frac{\sum y_i}{1-p} + \frac{n}{1-p} = 0 \qquad \frac{n(1-p) - \sum y_i (p) + n \cdot p = 0}{p(1-p)}$$

$$n(1-p) - \sum y_i \cdot p + n \cdot p = 0 \qquad n - np - \sum y_i p + np = 0$$

$$n = \sum y_i \, p \quad \Rightarrow \quad p = \frac{n}{\sum y_i} = \frac{1}{\bar{y}} \quad R$$

$\hat{p}_{ML} = \frac{1}{\bar{y}}$ Med det givna urvalet blir skattningen

$$\hat{p}_{ML} = 0{,}727 \quad R \qquad 12$$

Svar $\hat{p}_{ML} = \frac{1}{\bar{y}} = 0{,}727$

c) $\hat{E}(Y) = \frac{1}{\hat{p}_{ML}} = \frac{1}{0{,}727} = 1{,}3755.$

Svar 1,376 ggr Skövar 4

man testen i genomsnitt

(19)

SU, DEPARTMENT OF STATISTICS

Room: Brunnsuksalen Anonymous code: STM-0032 Sheet number: 7

(5)

a) $$f(y) = \begin{cases} \frac{2y}{\theta^2} & 0 \le y \le \theta \\ 0 & \text{annars} \end{cases}$$

Y_1, Y_2, Y_3 s.u. från $f(y)$ y, kont. variabel.
Antag observerade observationer.

$$\hat{\theta} = \frac{Y_1 + Y_2 + Y_3}{2}$$

$$E(\hat{\theta}) = E\left(\frac{Y_1 + Y_2 + Y_3}{2}\right) = \frac{E(Y_1) + E(Y_2) + E(Y_3)}{2}$$ R

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f(y)\, dy = \int_0^{\theta} y \cdot \frac{2y}{\theta^2}\, dy = \int_0^{\theta} \frac{2y^2}{\theta^2}\, dy =$$

$$\left[\frac{2y^3}{3\theta^2}\right]_0^{\theta} = \frac{2\theta^3}{3\theta^2} = \frac{2\theta}{3}$$ R

$$E(\hat{\theta}) = \frac{\frac{2\theta}{3} + \frac{2\theta}{3} + \frac{2\theta}{3}}{2} = \frac{\frac{6\theta}{3}}{2} = \frac{6\theta}{6} = \theta$$ R

Svar $E(\hat{\theta}) = \theta$ Ja, θ är väntevärdesriktig. R

8

b) $V(Y) = \int_{-\infty}^{\infty} y^2 f(y)\,dy - [E(Y)]^2$

$\int_{-\infty}^{\infty} \frac{y^2 \cdot 2y}{\theta^2}\,dy = \int_0^{\theta} \frac{2y^3}{\theta^2}\,dy = \left[\frac{2y^4}{4\theta^2}\right]_0^{\theta} = \frac{2\theta^4}{4\theta^2} = \frac{\theta^2}{2}$ R

$V(Y) = \frac{\theta^2}{2} - \left(\frac{2\theta}{3}\right)^2 = \frac{\theta^2}{2} - \frac{4\theta^2}{9} = \frac{9\theta^2 - 8\theta^2}{18} = \frac{\theta^2}{18}$ R

$V(\hat{\theta}) = V\left(\frac{Y_1 + Y_2 + Y_3}{2}\right) = \frac{1}{4} \cdot V(Y_1 + Y_2 + Y_3) =$ pga oberoende R

$\frac{1}{4}\left(V(Y_1) + V(Y_2) + V(Y_3)\right) = \frac{1}{4} \cdot \left(\frac{\theta^2}{18} + \frac{\theta^2}{18} + \frac{\theta^2}{18}\right) = \frac{3\theta^2}{18} \cdot \frac{1}{4} =$

$\frac{3\theta^2}{72} = \frac{\theta^2}{24}$ R

Svar: Variansen $V(\hat{\theta})$ är $\frac{\theta^2}{24}$

12

(20)