STOCKHOLMS UNIVERSITET
Statistiska institutionen
Ellinor Fackle-Fornius

TENTAMEN I STATISTISK TEORI MED TILLÄMPNINGAR II
2017-03-17

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 2017-04-03 kl. 13 i sal B315.

OBS! Glöm inte att ange nödvändiga antaganden överallt.

Uppgift 1. (20 poäng)

Förklara innebörden av följande begrepp:
a) Samplingfördelning
b) Konfidensgrad
c) p-värde
d) Momentmetoden
e) Neyman-Pearsons lemma

Uppgift 2. (20 poäng)

Ett företag utreder ett nytt tillsatsämne som potentiellt kan minska bensinförbrukningen för bilar. Utan tillsatsämnet är medelbensinförbrukningen för en viss bilmodell 0.8 liter/mil. Företaget planerar att göra en pilotstudie för att testa om tillsatsämnet minskar medelbensinförbrukningen genom att testköra ett antal bilar. Man vill att sannoliheten för typ I-fel ska vara högst 0.05 och att sannolikheten för typ II-fel ska vara högst 0.03 om medelbensinförbrukningen är 0.75 liter/mil. Antag att bensinförbrukningen är approximativt normalfördelad med varians 0.02.

a) Sätt upp lämpliga hypoteser för testet.

b) Hur många bilar ska företaget använda i pilotstudien?

c) Bestäm förkastelseområdet (RR) för testet.

Uppgift 3. (20 poäng)

Två metoder, A och B, för trafik-kontroll testades vid $n = 12$ korsningar under en period av 1 vecka för var och en av metoderna. Antalet olyckstillbud som registrerades under respektive korsning och metod visas i nedanstående tabell. Ordningen av de två metoderna (vilken som testades först) valdes slumpmässigt för varje korsning.

Korsning	Metod A	Metod B
1	5	4
2	6	4
3	8	9
4	3	2
5	6	3
6	1	0
7	2	3
8	4	1
9	7	9
10	5	2
11	6	5
12	1	1

a) Testa med ett teckentest om det går att påvisa någon skillnad mellan metoderna.

b) Testa med Wilcoxons teckenrangtest om det går att påvisa någon skillnad mellan metoderna.

Uppgift 4. (20 poäng)

Den flitige studenten Hasse bor i ett bostadshus med två hissar, en till vänster och en till höger. När man trycker på hissknappen kommer en av hissarna och Hasse undrar om det är samma sannolikhet för båda hissarna att komma. Låt Y vara lika med 1 om den vänstra hissen kommer då Hasse trycker på hissknappen och 0 annars. Då är $Y \sim Bern(p) = Bin(1, p)$ där p är sannolikheten att den vänstra hissen kommer. Hasse har tryckt på hissknappen $n = 100$ gånger och noterat att den vänstra hissen kommit 60 gånger.

a) Härled maximumlikelihood-skattningen $\widehat{p}_{ML}$ och beräkna Hasses skattning.

b) Vad blir maximumlikelihood-skattningen för variansen för Y?

c) Visa att $\widehat{p}_{ML}$ är väntevärdesriktig och konsistent.

Uppgift 5. (20 poäng)

Den flitige studenten Hasse vill nu undersöka om det är samma sannolikhet för båda hissarna att komma genom att testa

$$\begin{aligned} H_0 &: \quad p = 0.5 \\ H_a &: \quad p \neq 0.5 \end{aligned}$$

med hjälp av ett likelihoodkvottest.

a) Bestäm likelihoodfunktionen för $p = 0.5$.

b) Bestäm likelihoodfunktionen för $p = \widehat{p}_{ML}$.

c) Bestäm teststatistikan för likelihoodkvottestet.

d) Ange kritiskt område (RR) för signifikansnivån 0.05 när n är stort.

e) Genomför testet då Hasse observerat att den vänstra hissen har kommit 60 av de 100 gånger som han har tryckt på hissknappen.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 17/3-2017

Sal: Ugglevikssalen

Tenta: Statistisk teori med tillämpningar II

Kurs: Statistisk teori med tillämpningar

ANONYMKOD: STM-0007

☒ 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					5
Lär.ant. 20	20	16	20	18					

94 + 8 bonus

POÄNG	BETYG	Lärarens sign.
102	A	

SU, STATISTIK

Skrivsal: Ugglevik **Anonymkod:** STM-0007 **Blad nr:** 1

1 a) Samplingfördelning är sannolikhetsfördelningen för en statistikan.

4

b) Konfidensgraden är $(1-\alpha)$ och beskriver med vilken säkerhet vi gör en viss skattning. Om vi ex. vill skatta μ med en konfidensgrad på 95%, vill vi att värdet på μ i 95 av 100 stickprov hamnar inom det skattade intervallet. Vid ett hypotestest då H_0 testas på signifikansnivå α, sammanfaller konfidensgraden med A.R, acceptansområde.

4

c) P-värdet är sannolikheten att vi observerar vårt värde eller något mer extremt, tex $P(Z \geq Z_{obs})$. P-värdet anger den minsta sannolikhet med vilken sign.niv H_0 kan förkastas. Om ex RR: $\{Z \geq 1{,}96\}$ $(\alpha = 0{,}025)$ och vi observerar $Z_{obs} = 1{,}97$ dvs p-värdet: 0,0239 så kan vi förkasta hypotesen på 0,025 signifikansnivå. Som lägst skulle vi kunna förkasta hypotesen på signifikansnivå 0,0239. dvs P-värdet.

4

d) Momentmetoden är en metod för att hitta bra estimatorer för parametrar genom att jämföra urvalsmedelvärdet för en stokastisk variabel med väntevärdet för den samma

dvs $\mu_1' = E(Y^1)$ med $m_1' = \frac{\sum Y^1}{n}$

$\mu_2' = E(Y^2)$ med $m_2' = \frac{\sum Y^2}{n}$ m_K' kallas momentet

$\mu_K' = E(Y^K)$ med $m_K' = \frac{\sum Y^K}{n}$

Genom att sätta $\mu_1 = m_1$, $\mu_k = m_k$ kan man få fram estimatorer för parametrar, t.ex $E(Y) = \frac{\sum y_i}{n}$. Det är dock inte säkert att dessa är bra estimatorer som är väntevärdesriktiga, effektiva etc.

4

e) Neyman-Pearsons lemma säger att det starkaste testet för att testa två enkla hypoteser, H_0 & H_A ges när $\frac{L(H_0)}{L(H_A)} < K$, med ord: när likelihooden för hypotesen H_0 genom likelihooden för hypotesen H_A är mindre än ett visst värde k, eller tvärtom,

$$\frac{L(H_A)}{L(H_0)} > k.$$

4

20

SU, STATISTIK

Skrivsal: Ugglevik **Anonymkod:** STM-0007 **Blad nr:** 2

2. Y = "bensinförbrukning"

$Y \sim$ approx $N(\mu, \sigma^2 = 0{,}02)$

a)

$H_0: \mu = 0{,}8$

$H_A: \mu < 0{,}8$ R (för fråga b spec. vi $H_A: \mu = 0{,}75$) 4

b) $\alpha = P(H_0$ förkastas $| H_0$ är sann$) = 0{,}05$ $H_0: \mu = 0{,}8$

$\alpha = P(\bar{Y} < k \mid \mu = 0{,}8) = 0{,}05$ $H_A: \mu = 0{,}75$

$\beta = P(H_0$ accept $| H_A$ är sann$) = 0{,}03$

$\beta = P(\bar{Y} > k \mid \mu = 0{,}75) = 0{,}03$

$p\left(Z_{obs} < \frac{k-0{,}8}{\sqrt{\frac{0{,}02}{n}}}\right) = 0{,}05$ ① $Z_{obs} < \frac{k-0{,}8}{\sqrt{\frac{0{,}02}{n}}} = 1{,}6449$

R

$P\left(Z_{obs} > \frac{k-0{,}75}{\sqrt{\frac{0{,}02}{n}}}\right) = 0{,}03$ ② $Z_{obs} < \frac{k-0{,}75}{\sqrt{\frac{0{,}02}{n}}} = 1{,}88$

$\|$

$P\left(Z_{obs} < \frac{k-0{,}75}{\sqrt{\frac{0{,}02}{n}}}\right) = 0{,}97$ R

~~Sätter ① & ② lika varandra~~ & bryter ut k.

$\frac{k-0{,}8}{\sqrt{\frac{0{,}02}{n}}} = -1{,}6449$ R $\frac{k-0{,}75}{\sqrt{\frac{0{,}02}{n}}} = 1{,}88$ R

$k - 0{,}8 = -1{,}6449\left(\sqrt{\frac{0{,}02}{n}}\right)$ $k - 0{,}75 = 1{,}88\sqrt{\frac{0{,}02}{n}}$

$k = 0{,}8 - 1{,}6449\left(\sqrt{\frac{0{,}02}{n}}\right)$ $k = 0{,}75 + 1{,}88\sqrt{\frac{0{,}02}{n}}$

$0{,}8 - 1{,}6449\left(\sqrt{\frac{0{,}02}{n}}\right) = 0{,}75 + 1{,}88\sqrt{\frac{0{,}02}{n}}$

$0{,}05 = 1{,}88\sqrt{\frac{0{,}02}{n}} + 1{,}6449\left(\sqrt{\frac{0{,}02}{n}}\right)$

$$0{,}05 = \sqrt{\frac{0{,}02}{n}}\left(1{,}88 + 1{,}6449\right)$$

$$\left(\frac{0{,}05}{(1{,}88 + 1{,}6449)}\right)^2 = \frac{0{,}02}{n}$$

$$n = \frac{0{,}02}{\left(\frac{0{,}05}{1{,}88+1{,}6449}\right)^2}$$

$$n = 99{,}39$$ R

Svar: För att slh för ett typ 1 fel ska vara högst 0,05 & slh för ett typ 2 fel ska vara högst 0,03 om medelbensinförbrukningen är 0,74 liter behöver vi 100 bilar i testet. R

10

c). $\alpha = P(\text{Typ 1 fel}) = P(H_0 \text{ förkastas} \mid H_0 \text{ är sann}) =$

$P(\bar{Y} < k \mid \mu = 0{,}8) = P\left(Z < \frac{k - 0{,}8}{\sqrt{\frac{0{,}02}{n}}}\right) = 0{,}05$ från tabell

Vi förkastar H_0 om $Z_{obs} < -1{,}6449$ om $n = 100$

R

$$\frac{k - 0{,}8}{\sqrt{\frac{0{,}02}{n}}} = -1{,}6449$$

$$k = -1{,}6449 \cdot \sqrt{\frac{0{,}02}{100}} + 0{,}8$$

$$k = 0{,}7767$$ R

Svar: Förkastelseområdet är $RR: \{\bar{Y} < 0{,}7767\}$ eller

$RR: \{Z_{obs} < -1{,}6449\}$

6

20 Bra!

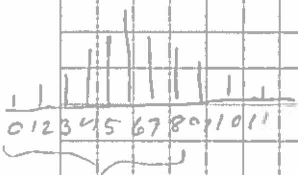
SU, STATISTIK

Skrivsal: Ugglevik, Anonymkod: STM-0007 Blad nr: 3.

3. x Vi har samma kroppningar men två metoder = parvisa observationer
x Slumpmässig ordning mellan metoderna.

a)

Körning	Metod A	Metod B	Tecken	diff	Rang
1	5	4	+	1	3,5
2	6	4	+	2	7,5
3	8	9	−	1	3,5
4	3	2	+	1	3,5
5	6	3	+	3	10
6	1	0	+	1	3,5
7	2	3	−	1	3,5
8	4	1	+	3	11
9	7	9	−	2	7,5
10	5	2	+	3	10
11	6	5	+	1	3,5
12	1	1	=	=	tie



R

a) Vi har $n = 11$ (en tie), vi har $m = 8$

Teststatistika: m = pos diff då $m \sim bin(n = 11, p)$. R

Hypoteser:

H_0: $p = 0,5$ (fördeln är lika)

H_1: $p \neq 0,5$ (fördeln är inte lika) R

Antaganden: oberoende och likafördelade observationer mellan?

Vi beräknar p-värdet för $m \geq 8$ då $n = 11$ & $p = 0,5$

$$2 \times p(m \leq 7 \mid p = 0,5) = 2(1 - 0,88672) = 0,22656$$ R

Svar: Det lägsta värde på vilket vi kan förkasta H_0 skulle vara, då $m_{obs} = 8$, sannolikheten för att få det värdet eller något mer extremt $p(m \geq 8, m \leq 3) = 0,22656$ (dubbelsidigt test). Vi kan inte förkasta H_0 på någon rimlig signifikansnivå. slutsats?

7

d). Wilcoxons teckenrang test. Vi rangordnar obs. (se svar på fråga b)

$\frac{11(12)}{2} = 66$

52,5

H_0 = fördelningarna är lika

H_1 = fördelningarna är inte lika

Teststatistika: $Min(T^-, T^+)$ $T^- = 3,5 + 3,5 + 7,5 = 14,5$ R $n = 11$

Dubbelsidigt test, Vi bestämmer oss för att förkasta H_0 på signifikansnivå 0,05, $\alpha = 0,05$, $\alpha/2 = 0,025$ ger $RR: \{T^-_{obs} \leq 11\}$ R

Svar:

Vårt observerade värde $min(T^-; T^+) = 14,5$. Vi kan inte förkasta H_0 på en signifikansnivå på 0,05.

slutsats 2

9

16

SU, STATISTIK

Skrivsal: Uggleviks. Anonymkod: STM-0007 Blad nr: 4.

4. $Y \overset{iid}{\sim} Ben(p) = Bin(1, p)$

$Y=1$ "hissen kommer"

$Y=0$ "hissen kommer inte"

Vi har oberoende & likafördelade obs,

$Y \overset{iid}{\sim} Bern(p)$

$n=100$

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

a) $L(p/y) = \prod_{i=1}^{n} p^{y}(1-p)^{n-y} = p^{y_1}(1-p)^{n-y_1} \times p^{y_2}(1-p)^{n-y_2} \times p^{y_n}(1-p)^{n-y_n} =$

$= p^{\Sigma y_i}(1-p)^{n-\Sigma y_i}$

$l(p/y) = \Sigma y_i \ln p + (n - \Sigma y_i) \ln(1-p) =$

$= \Sigma y_i \ln p + n \ln(1-p) - \Sigma y_i \ln(1-p)$

R

$$\frac{d(l(p/y))}{dp} = \frac{\Sigma y_i}{p} - \frac{n}{(1-p)} + \frac{\Sigma y_i}{(1-p)} = 0$$

$$\frac{\Sigma y_i}{p} = \frac{n - \Sigma y_i}{(1-p)}$$

$$\frac{(1-p)}{p} = \frac{n - \Sigma y_i}{\Sigma y_i}$$

$$\frac{1}{p} - \cancel{\frac{p}{p}} = \frac{n}{\Sigma y_i} - \cancel{\frac{\Sigma y_i}{\Sigma y_i}}$$

$$\frac{\Sigma y_i}{n} = \hat{p}_{ML}$$

R

Hasses skattning:

$\hat{p}_{ML} = \frac{\Sigma y_i}{n} = \frac{60}{100} = 0,6$

R

Svar: $\hat{p}_{ML} = \frac{\Sigma y_i}{n}$ och

hasses skattning $= \hat{p}_{ML} = 0,6$.

10

b) Vi utnyttjar att ML skattningen av en funktion av $\hat{\theta}_{ML}$ = skattningen av en funktion θ_{ML} $\widehat{f(\theta)}_{ML} = f(\hat{\theta}_{ML})$

$$V(Y) = npq = n\cdot p(1-p) = n\cdot \hat{p}_{ML}(1-\hat{p}_{ML}) = \cancel{n}\cdot\frac{\sum y_i}{\cancel{n}}\left(1-\frac{\sum y_i}{n}\right) =$$

$$= \sum y_i - \frac{(\sum y_i)^2}{n}$$

OK

Svar: Maximumlikelihoodskattningen för variansen av Y = $\sum y_i - \frac{(\sum y_i)^2}{n}$. Använder vi hasses skattning blir

Variansen av $Y = 60 - \frac{60^2}{n} = 24.$

4

c) $\hat{p}_{ML} = \frac{\sum y_i}{n}$ $\quad E(y) = n\cdot p \quad V(y) = n\cdot p(1-p)$ när $Y \overset{iid}{\sim} Ber(p)$

$$E(\hat{p}_{ML}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}\cdot E\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\cdot\sum E(y) = \frac{1}{n}\cancel{n}\cdot np = \underline{\underline{np}}$$

V.V.R

$E(y) = np \quad V(y) = np(1-p)$ när $Y \overset{iid}{\sim} Ber(p)$

Om $\hat{p}_{ML} = \frac{\sum y_i}{n}$ är konsistent ska $V(\hat{p}_{ML}) \to 0$ när $n \to \infty$ $\bar{R}$

vid oberoende observationer

$$V(\hat{p}_{ML}) = V\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2}\cdot V\left(\sum y_i\right) = \frac{1}{n^2}\cdot\sum V(y_i) = \frac{1}{n^2}\cdot n(p(1-p))$$

$= \frac{p(1-p)}{n}$ är konsistent eftersom när $n \to \infty$ kommer

$\frac{p(1-p)}{n} \to 0.$ R

6

20

SU, STATISTIK

Skrivsal: Ugglevik Anonymkod: STM-0007 Blad nr: 5

5. $Y \overset{iid}{\sim} ber(p)$

$H_0: \quad p=0,5$

$H_A: \quad p \neq 0,5$

$p(y) = p^y (1-p)^{\overset{1}{n}-y}$

$L(p|y) = p^{\sum y_i} (1-p)^{n-\sum y_i}$ R

a) $L(p=0,5) = 0,5^{\sum y_i} (1-0,5)^{n-\sum y_i} = 0,5^{\sum y_i} (0,5)^{n-\sum y_i} =$

$= 0,5^{\sum y_i + n - \sum y_i} = \underline{\underline{0,5^n}}$

Svar: $\underline{\underline{0,5^n}}$

R 2

b) $L(p=\hat{p}_{ML})$ då $\hat{p}_{ML} = \frac{\sum y_i}{n}$

$$L(\hat{p}_{ML}) = \underline{\underline{\hat{p}_{ML}^{\sum y_i} (1-\hat{p}_{ML})^{n-\sum y_i} = \left(\frac{\sum y_i}{n}\right)^{\sum y_i} \left(1-\left(\frac{\sum y_i}{n}\right)\right)^{n-\sum y_i}}}$$

Svar: $\underline{\underline{\hat{p}_{ML}^{\sum y_i} (1-\hat{p}_{ML})^{n-\sum y_i} = \left(\frac{\sum y_i}{n}\right)^{\sum y_i} \left(1-\left(\frac{\sum y_i}{n}\right)\right)^{n-\sum y_i}}}$

R 2

c) $$\Lambda_{LR} = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\max_{\theta \in \Omega_0} L(\hat{\theta})}{\max_{\theta \in \Omega} L(\hat{\theta})} = \underline{\underline{\frac{0,5^n}{\hat{p}_{ML}^{\sum y_i} (1-\hat{p}_{ML})^{n-\sum y_i}} < k}}$$

R

Svar: Teststatistikan är $\underline{\underline{\Lambda_{LR} = \frac{0,5^n}{\hat{p}_{ML}^{\sum y_i} (1-\hat{p}_{ML})^{n-\sum y_i}} < k.}}$

4

d) RR när n är stort är att H_0 ska förkastas

när $\Lambda_{LR} < k \underset{\text{samma sak}}{\Leftrightarrow} \frac{0,5^n}{\hat{p}_{ML}^{\Sigma y_i}(1-\hat{p}_{ML})^{n-\Sigma y_i}} < k$

vilket kan bestämmas genom att vi vet att

$-2\ln\Lambda_{LR} \sim \chi^2(1)$ dvs R

$-2\ln\left(\frac{0,5^n}{\hat{p}_{ML}^{\Sigma y_i}(1-\hat{p}_{ML})^{n-\Sigma y_i}}\right) > \underbrace{(-2\ln k)}_{k^*}$

$\alpha = 0,05$ $P(X > \chi^2_\alpha(f)) = 0,05$

$X > 3,84$

Svar:

Vi ska förkasta H_0 när $-2\ln\left(\frac{0,5^n}{\hat{p}_{ML}^{\Sigma y_i}(1-\hat{p}_{ML})^{n-\Sigma y_i}}\right) > 3,84$.

R 7

e) Vi har observerat $\hat{p} = \frac{60}{100} = 0,6$

$-2\ln\left(\frac{0,5^{100}}{0,6^{60}(0,4)^{100-60}}\right) > 3,84$

$4,02710 > 3,84$

H_0 ska förkastas på signifikansnivå 0,05.

Slutsats ?

3

18

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 17/3-2017

Sal: Uggleviksssalen

Tenta: Statistisk teori med tillämpningar II

Kurs: Statistisk teori med tillämpningar

ANONYMKOD: STM-0017

☑ 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					8
Lär ant. 20	20	20	20	20					

HRS

100 + 8 bonus

POÄNG	BETYG	Lärarens sign.
108	A	

SU, STATISTIK

Skrivsal: UB Anonymkod: STM-0017 Blad nr: -1-

Uppgift 1

a) Samplingfördelning är, kort sagt, en sannolikhetsfördelning för en statistika om ex, t. ex, har ett urval bestående av n styck oberoende och likfördelade variabler $Y_1, Y_2, \ldots, Y_n$ från en normalfördelad population, dvs $Y_i \sim N(\mu, \sigma^2)$, dess samplingfördelning kommer att bestämmas av

$$\bar{Y} = \frac{\sum Y_i}{n}$$ (där $Y_1 = y_1$, $Y_2 = y_2$, …, $Y_n = y_n$) och

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$$, där

$\bar{Y}$ $\bar{Y} \sim N\left(\bar{Y}, \frac{\sigma^2}{n}\right)$, som motiverar fördelningen för $\bar{Y}$ då n styck oberoende urval väljs där $\bar{Y}$ kommer att bero på konkreta utfall av Y_i ($Y_1 = y_1, \ldots, Y_n$), $\bar{Y}$ ~~kommer~~ kommer att vara appr. N om n är tillräckligt stort enligt CGS oavsett om Y_i kommer från en normalfördelad population eller ej, dvs fördelningen för $\bar{Y} \overset{appr}{\sim} N\left(\bar{Y}, \frac{\sigma^2}{n}\right)$ om $n \geq 30$

4

b) Konfidensgrad, $100(1-\alpha)\%$ anger antal utfall för det sanna värdet på sökta parametern att befinna sig i den konfidensintervall som vi har beräknat. T.ex. om vi söker ett 95%-igt konfidensintervall för någon parameter θ, kan vi säga att det sanna värdet på θ kommer att befinna sig i det sökta intervallet med 95%-igt konfidens, dvs i 95 fall av 100 (ifall ~~att~~ vi baserar våra beräkningar på data oberoende s.a. av samma storlek och från samma population). Konfidensgrad bör inte förväxlas med sannolikheten att få det sanna värdet i det givet konfidensintervallet. Så är detta [illegible] enligt klassisk inferens. 4

c) p-värde är sannolikheten att få det värdet som vi har observerat eller ännu mer extremt.
Den anger (också) det lägsta sign.nivå på vilket vi kan förkasta vår nollhypotes. T.ex. om vi bestämmer
för $\alpha = 0{,}05$ och vi får ett p-värde som är 0,06 kan vi inte förkasta H_0 på 0,05, däremot kan vi förkasta den på $\alpha \geq 0{,}06$ om det är rimligt

4

SU, STATISTIK STM-0017

Skrivsal: Uk Anonymkod: ______ Blad nr: ______

d) Momentmetoden är en av metoder som används för att skatta (på ett lämpligt sätt) parametrar vi är intresserad i. Det är en enkel metod för att härleda de sökta parametrar men den ger dock inte alltid de bästa skattningarna. (inte alltid vvv). Metoden baseras på "~~moment generating function~~" som anger att unika sannolikhetsfördelningarna för funktioner. Argument för att hitta skattningar av sökta parametrar att följande:

$$\boxed{m_k' = \frac{\sum y_i^k}{n}} \Rightarrow \text{det utgår från}$$

- Sökes första populationsmoment: (antar att $y_i \sim N(\mu, \sigma^2)$).

$$\begin{cases} \mu_1' = \mu \\ m_1' = \frac{\sum y_i}{n} = \bar{y} \end{cases}$$

- $\mu_1' = m_1' = \bar{y}$

- $$\begin{cases} \mu_2' = \sigma^2 + \mu^2 \\ m_2' = \frac{\sum y_i^2}{n} \end{cases} \qquad m_2' = \mu_2'$$

$$\begin{cases} V(y) = \Sigma E(y_i)^2 - (E(y))^2 \\ \Sigma E(y_i)^2 = V(y) + (E(y))^2 = \sigma^2 + \mu^2 \end{cases}$$

$$\frac{\sum y_i^2}{n} = \sigma^2 + \bar{y}^2 \Rightarrow \sigma^2 = \frac{\sum y_i^2 - n\bar{y}^2}{n}$$ (dock v vvv)

d) ~~Neyman-Pearson Lemma~~ anser vara det bästa testet för ett givet nivå av α (sign nivå, typ I-fel) och ett nivå β som kan anpassas så att det kommer att vara det riktigt test. NPL baseras på att hitta skattningsfunktionerna för de parametrar som passar, dvs det testar vilken modell kommer på ett bästa sätt beskriva vår data givet parametrarna som vi har.

Om: $H_0 = \theta_0$

$H_a = \theta_a$, t [illegible]

Testvariabel

$\lambda = \frac{L(\theta_0)}{L(\theta_a)} < k$, Om $\lambda < k$, då kan vi förkasta H_0. För att hitta RR behöver vi hitta den bästa testsstatistik som kan beskriva data och bestämma dess fördelning. Detta kan göras genom att lösa ekvationen

$\frac{L(\theta_0)}{L(\theta_a)} < k$ och genom att lösa ut "y" och bestämma dess fördelning

NPL kan också användas för sammansatta hypoteser: $H_0: \theta = \theta_0$

$H_a: \theta < \theta_0,\ \theta > \theta_0$

(uniformly most powerful test)

LR

4

20

SU, STATISTIK

Skrivsal: U6 Anonymkod: STM-0017 Blad nr: -3-

Uppgift 2

- y – Bensinförbrukningen
- $y_i \overset{app.}{\underset{iid}{\sim}} N(0{,}8;\ \sigma^2 = 0{,}02)$
- $\alpha = 0{,}05$
- $\beta = 0{,}03$ om $\bar{y} = 0{,}75$ (liter/mil).

a) Sätt upp lämpliga hypoteser för testet.

- Vi vet att medelbensinförbrukningen utan tillsatsämnet för en viss bilmodell är 0,8 liter/mil.
- Företaget vill testa om tillsatsämnet minskar medelbensinförbrukningen.
- Vi vill testa därför följande hypoteser:

Svar: H_0: $\mu = 0{,}8$ (att tillsatsämnet inte påverkar medelbensinförbrukningen)

$\alpha = 0{,}05$ H_a: $\mu < 0{,}8$ (att tack vare tillsatsämnet medelbensinförbrukningen faktiskt minskar)

Detta är en enkelsidig (sammansatt (composite)) hypotes.

R

R

4

b) Hur många bilar ska företaget använda i pilotstudien om $\alpha \leq 0,5$, $\beta \leq 0,3$ om $\bar{y} = 0,75$ liter/mil? $n = ?$

- Eftersom $\bar{y} \approx$ appr. NF med känd varians $\sigma^2 = 0,02$ vi kan använda oss av teststatistika Z i våra beräkningar av n.

- Vi har en enkelsidig test, kan vi använda oss av följande formel för att beräkna n:

$$n = \frac{(Z_\alpha + Z_\beta)^2 \cdot \sigma^2}{(\mu_a - \mu_0)^2}, \text{ där}$$

$\Rightarrow Z_\alpha = Z_{0,05} = 1,6449$

$\Rightarrow P(Z > Z_\beta) = 0,03 \Rightarrow 1 - P(Z < Z_\beta) = 0,03 \Rightarrow$

$\Rightarrow P(Z < Z_\beta) = 0,97$ (ur tabell 1) $\Rightarrow$

$Z_\beta \approx 1,88$

$$\Rightarrow n = \frac{(1,6449 + 1,88)^2 \cdot 0,02}{(0,75 - 0,8)^2} = \frac{12,42 \cdot 0,02}{0,0025} =$$

$$\approx \frac{0,2485}{0,0025} = 99,4 \Rightarrow n = 100$$

<u>Svar</u>: företaget ska använda minst 100 bilar i pilotstudien ($\alpha \leq 0,05$, $\beta \leq 0,03$).

SU, STATISTIK

Skrivsal: UG Anonymkod: STM-0017 Blad nr: -4-

[Fort. Upp. 2]

c) RR = ?

$\alpha = P(\text{förkasta } H_0 \mid H_0 \text{ är sann}) =$
$= P(\bar{y} < k \mid \mu = 0{,}8) = P\left(Z < \underbrace{\frac{k-0{,}8}{\sqrt{\frac{0{,}02}{100}}}}_{Z_{obs}}\right) =$
$= 0{,}05 \Rightarrow (\text{ur Tabell 1})\ Z_{obs} = -1{,}6449$

$\frac{k-0{,}8}{\sqrt{\frac{0{,}02}{100}}} = -1{,}6449 \Rightarrow$

$= k - 0{,}8 = -1{,}6449 \cdot 0{,}01414$
$k = -1{,}6449 \cdot 0{,}01414 + 0{,}8 \simeq 0{,}7767 \simeq 0{,}78$

Svar: $RR = \{\bar{y} \le 0{,}78 \mid H_0 \text{ är sann}\}$ R 6

20

[Additionell info för upp 2, b)

$n = \frac{(Z_\alpha + Z_\beta)^2 \cdot \sigma^2}{(\mu_a - \mu_0)^2}$ härleds ur följande:

$\alpha = P(\bar{y} \le k \mid \mu = 0{,}8) \Rightarrow \frac{k-0{,}8}{\sqrt{\frac{0{,}02}{n}}} = -1{,}6449$

$\Rightarrow k = -1{,}6449 \cdot \sqrt{\frac{0{,}02}{n}} + 0{,}8$

$\beta = P(\bar{y} > k \mid \mu = 0{,}75) \Rightarrow \frac{k-0{,}75}{\sqrt{\frac{0{,}02}{n}}} = 1{,}88$

$\Rightarrow k = 1{,}88 \sqrt{\frac{0{,}02}{n}} + 0{,}75$

$-1{,}6449 \cdot \sqrt{\frac{0{,}02}{n}} + 0{,}8 = 1{,}88 \cdot \sqrt{\frac{0{,}02}{n}} + 0{,}75 \rightarrow n = \frac{(Z_\alpha + Z_\beta)^2 \cdot \sigma^2}{(\mu_a - \mu_0)^2}$

Bra!

Uppgift 3

- A, B - två metoder för trafik-kontroll
- $n = 12$ korsningar under 1 vecka

Korsning	A	B	Tecken	Diff	Rang
1	5	4	+	1	3,5
2	6	4	+	2	7,5
3	8	9	−	1	3,5
4	3	2	+	1	3,5
5	6	3	+	3	10
6	1	0	+	1	3,5
7	2	3	−	1	3,5
8	4	1	+	3	10
9	7	9	−	2	7,5
10	5	2	+	3	10
11	6	5	+	1	3,5
12	1	1	0	0	→ tie

R

(vi utesluter från testerna 1 obs., dvs $n = 11$.)

a) Skillnad mellan A och B med teckentest?

- X = antal olycks-tillbud, metod A
- Y = antal — " — metod B.

Antagande:
- parvisa observationer (beroende mellan X_i och Y_i)
- oberoende observationer inom X_i och Y_i

Bra!

⇒

SU, STATISTIK STM-0017

Skrivsal: U6 Anonymkod: Blad nr: -5-

Fort. Upp 3

H_0: det finns ingen skillnad mellan metoderna A och B (ingen skillnad i läge för fördelningarna x_i och y_i).

H_a: metoderna A och B skiljer sig åt (dvs det finns en skillnad i läge mellan fördelningarna x_i och y_i). R

- M = antal positiva differenser mellan x_i och y_i
- Om H_0 är sann, dvs att det finns ingen skillnad i fördelningarnas läge blir chk $p = \frac{1}{2} = 0,5$ (chk att få både positiva och negativa differenser $= 0,5$).

$\Rightarrow$

- $H_0: p = \frac{1}{2}$

  $H_a: p \neq \frac{1}{2}$ (dubbelsidig hypotes). R

$\Rightarrow$

- $M \sim Bin(n = 11, p = 0,5)$ (om H_0 är sann).
- RR = {förkastar för extremt små eller extremt stora M} R
- Finns ingen α specificerad i uppgiften, därför väljer jag att räkna p-värdet.
- $m = 8$, $P(M \geq 8 | H_0 \text{ är sann}) =$

$= 1 - P(M \leq 7 | p = 0,5) = \{$ ur tabell för Bin $\} =$

$= 1 - 0,88672 = 0,11328$ R $\Rightarrow$

p-value $= 2 \cdot P(M \geq 8) = 2 \cdot 0{,}11328 = 0{,}22656$
p.g.a dubbelsidigt test
R

Svar: eftersom p-value $= 0{,}22656$ det går inte att förkasta H_0 på något rimligt signifikansnivå, dvs vi fick inte tillräckligt med stöd för att bevisa att det (finns) skillnader mellan metoderna

10

b) Skillnad mellan metoderna med Wilcoxons teckenrangtest.

- Antagende: samma antagande som i a)
  - symmetrisk fördelning (ung. $\frac{n}{2}$ kommer att ligga på båda sidor av rangmedianen om det inte finns någon skillnad i fördelningarnes läge).

- Hypoteser: H_0: det finns ingen skillnad i läge mellan fördelningarne X_i och Y_i

  (dubbel-sidig test) H_a: fördelningarne X_i och Y_i skiljer sig åt i läge (åt höger eller åt vänster). R

- $T^- = 3{,}5 + 3{,}5 + 4{,}5 = 14{,}5$ R

  $T = \frac{n(n+1)}{2} = \frac{11 \cdot 12}{2} = 66$, $T^+ = 66 - 14{,}5 = 51{,}5$ R

- $T_{obs} = \min\{T^+; T^-\} = \min\{51{,}5; 14{,}5\} = T^- = 14{,}5$

R

ŞU, STATISTIK

Şkrivsal: Ur   STM-0017 Anonymkod: ______ Blad nr: -6-

Fort. Upp. 3

- RR = { $T_{obs} < T_0$ } om H_0 är sann
  Eftersom sign. nivå inte är specificerat i uppgiften jag bestämmer för $\alpha = 0,05$ för dubbelsidigt test med $n = 11$. ⇒ Ur tabell 9 hittar vi att $T_0 = 11$.
  ⇒ RR = { $T_{obs} < 11$ } H_0 förkastas.
  R
- $T_{obs} = T^- = 14,5 > T_0 = 11$ ⇒ vi kan inte förkasta H_0.

Svar: Med Wilcoxons teckenrangtest fick vi samma svar som med teckentest, nämligen att vi inte kan förkasta H_0 på ($\alpha = 0,05$, även för $\alpha = 0,10$ då $T_0 = 14$). Vi fick inte ett tillräckligt med stöd för att påstå att det finns någon skillnad mellan metoderna.

10

20 Bra!

Uppgift 4

- $Y \sim Bern(p) = Bin(1, p)$.
- $n = 100$ (tryckt på knappen).
  Success: 60 gånger (den vänstra hissen kommit). Y = antal gånger hissen kommit = 60

$\hat{p}_M = ?$

y

$\bar{y} = $

$\hat{p}_M$ ... 6 , $\hat{p}_{M_L} = $

SU, STATISTIK STM-0017

Skrivsal: UG Anonymkod: Blad nr: -7-

Fort. Upp. 4

b) Max-likelihood skattningen för variansen för y.

- $y \sim Bin(1, p)$
- $V(y) = np(1-p) = \{n = 1\} = p(1-p)$.
- $V(y)$ är funktionen av $p \Rightarrow t(p)$. R

Enligt "invariance property" för funktioner:

$$\widehat{V(y)} = \widehat{t(p)} = t(\hat{p}) = t(\hat{p}_{ML}) =$$ R

$$= \hat{p}_{ML}(1-\hat{p}_{ML}) = \bar{y}(1-\bar{y})$$

Svar: $\hat{V}_{ML}(y) = \bar{y}(1-\bar{y})$. R 4

c) Visa att $\hat{p}_{ML}$ är VVR och konsistent.

- Estimatorn kan anses vara VVR om $E(\hat{\theta}) = \theta$
- I vårt fall

$$E(\hat{p}_{ML}) = E(\bar{y}) = E\left(\frac{\sum y_i}{n}\right) =$$

$$= \frac{1}{n} E(\sum y_i) = \frac{1}{n} \underbrace{E(n y_i)}_{n E(Y_i)} = \{y_i \sim Bin(1,p)\} =$$

$$= \frac{n}{n} \cdot p = p \Rightarrow VVR.$$

R

• Om estimatorn är VVK då kan vi kolla dess konsistens enligt följande:

$$\lim_{n\to\infty} V(\hat{p}_{ML}) \to 0 \Rightarrow \text{konsistent}$$ R

$$\Rightarrow V(\hat{p}_{ML}) = V(\bar{y}) = V\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2} V\left(\sum y_i\right) =$$

$$\underset{\substack{p.g.a\\ i.i.d}}{=} \frac{1}{n^2} \cdot n \cdot V(y_i) = \{ y_i \sim Bin(1,p) \} =$$

OK!

$$= \frac{1}{n} \cdot p(1-p) = \frac{p(1-p)}{n}$$ R

$$\Rightarrow \lim_{n\to\infty} \frac{p(1-p)}{n} \to 0, \text{ dvs. konsistent}$$

R

Svar: $\hat{p}_{ML}$ är både VVR och konsistent då

$E(\hat{p}_{ML}) = p$ och $\lim_{n\to\infty} V(\hat{p}_{ML}) \to 0$. 6

20 Bra!

Uppgift 5 (bygger på upp. 4).

H_0: $p = 0,5$

H_a: $p \neq 0,5$

a) Likelihoodfunktionen för $p = 0,5$.

$$L(p) = p^{\sum y_i}(1-p)^{n-\sum y_i} = p^{n\bar{y}}(1-p)^{n-n\bar{y}}$$

från upp. 4.

$$L(p=0,5) = 0,5^{n\bar{y}} \cdot (1-0,5)^{n-n\bar{y}} = 0,5^{n\bar{y}} \cdot 0,5^{n-n\cdot\bar{y}} =$$

$$= 0,5^{n\bar{y}+n-n\bar{y}} = 0,5^{n}$$ Svar: $L(p=0,5) = 0,5^n$

R 2

SU, STATISTIK STM-0017

Skrivsal: Ub Anonymkod: Blad nr: -8-

b) Bestäm $L(p = \hat{p}_{ML})$

$$L(p = \hat{p}_{ML} = \bar{y}) = \hat{p}_{ML}^{\,n\bar{y}} \cdot (1 - \hat{p}_{ML})^{n - n\bar{y}} =$$

[från upp. 4 vet vi att $\hat{p}_{ML} = \bar{y}$]

$$= \bar{y}^{\,n\bar{y}} \cdot (1 - \bar{y})^{n - n\bar{y}} =$$

$$= \bar{y}^{\,n\bar{y}} \cdot (1 - \bar{y})^{n(1-\bar{y})}$$

Svar: $L(\hat{p}_{ML}) = y^{n\bar{y}} \cdot (1-y)^{n(1-\bar{y})}$ R 2

c) $$\lambda_{LR} = \frac{\cancel{max}\, L(0{,}5)}{\cancel{max}\, L(\hat{p}_{ML})} = \frac{0{,}5^n}{\bar{y}^{\,n\bar{y}} \cdot (1-\bar{y})^{n(1-\bar{y})}} =$$

$$= \frac{(1-\bar{y})^{n\bar{y}}}{2^n \cdot \bar{y}^{\,n\bar{y}} \cdot (1-y)^n} = \left(\frac{1-\bar{y}}{\bar{y}}\right)^{n\bar{y}} \cdot \left[2(1-\bar{y})\right]^{-n} =$$

$$= \left(\frac{1}{\bar{y}} - 1\right)^{n\bar{y}} \left[2(1-\bar{y})\right]^{-n} < \underline{k} \quad R$$

Svar: teststatistikan är $\lambda_{LR} = \left(\frac{1}{y} - 1\right)^{n\bar{y}} \cdot \left[2(1-\bar{y})\right]^{-n}$ som borde vara mindre än en viss konstant k för att kunna förkasta H_0 (λ_{LR} måste vara tillräcklig litet). 4

d) Om n är stort vet vi att $-2 \ln \lambda_{LR}$ är χ^2 fördelad med 1 f.g. Med $\alpha = 0{,}05$ bestäms kritiskt område som: R

$\lambda_{LR} < k$
$-2\ln\lambda_{LR} > k - 2\ln k$

RR $\{-2\ln\lambda_{LR} > 3{,}84\}$ R

$-2\ln\lambda_{LR} = -2\left(\ln\left(\frac{1-\bar{y}}{\bar{y}}\right)^{n\bar{y}} + \ln(2(1-\bar{y}))^{-n}\right) =$

$= -2\left(n\bar{y}\left(\ln(1-\bar{y}) - \ln\bar{y}\right) - n\left(\ln 2 + \ln(1-\bar{y})\right)\right) =$

$= -2n\left(\bar{y}\ln(1-\bar{y}) - \bar{y}\ln\bar{y} - \ln 2 - \ln(1-\bar{y})\right) =$

$= -2n\left((\bar{y}-1)\cdot\ln(1-\bar{y}) - \bar{y}\cdot\ln\bar{y} - \ln 2\right) =$

$= 2n\left[(1-\bar{y})\cdot\ln(1-\bar{y}) + \bar{y}\cdot\ln\bar{y} + \ln 2\right)$ R

Svar: RR $= \{-2\ln\lambda_{LR} = 2n[(1-\bar{y})\cdot\ln(1-\bar{y}) +$
$+ \bar{y}\cdot\ln\bar{y} + \ln 2] > 3{,}84\}$ om H_0 är sann.

8

c) $p = \bar{y} = \frac{\sum y_i}{n} = \frac{60}{100} = 0{,}6.$

$-2\ln\lambda_{LR} = 2\cdot 100\left((1-0{,}6)\cdot\ln(1-0{,}6) +\right.$
$\left.+ 0{,}6\cdot\ln 0{,}6 + \ln 2\right) \simeq 200\left(0{,}4\cdot(-0{,}9163) +\right.$
$+ 0{,}6\cdot(-0{,}5108) + 0{,}6931) \approx 200(-0{,}3665 -$
$\left.- 0{,}306 + 0{,}693\right) \approx 200\cdot 0{,}02013 \approx \underline{4{,}027}$ R

Svar: $-2\ln\lambda_{LR} = 4{,}027 > 3{,}84$, dvs
vi kan förkasta H_0 på $\alpha = 0{,}05$.
Vi fick stöd att för mothypotesen
att $p \neq 0{,}5$.

4

20

Genomgående tydliga och bra lösn.!