STOCKHOLMS UNIVERSITET
Statistiska institutionen
Ellinor Fackle-Fornius

TENTAMEN I STATISTISK TEORI MED TILLÄMPNINGAR II
2017-10-30

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.
Resultatet meddelas senast den 14 november.

OBS! Glöm inte att ange nödvändiga antaganden överallt

Uppgift 1. (20 poäng)

Förklara innebörden av följande begrepp:
a) Samplingfördelning
b) Centrala gränsvärdessatsen
c) Medelkvadratfel (MSE)
d) Konsistens
e) Momentskattning

Uppgift 2. (20 poäng)

En facklig organisation genomförde en enkätundersökning bland sina medlemmar för att bland annat uppskatta andelen som har haft utvecklingssamtal med sin chef under det senaste året. Av totalt 398 respondenter svarade 287 stycken att de hade haft (minst) ett utvecklingsamtal med sin chef under det senaste året.

a) Beräkna ett 99 %-igt konfidensintervall för andelen medlemmar som har haft utvecklingssamtal med sin chef under det senaste året.

b) Använd konfidensintervallet i a) för att testa hypotesen att 80 % av medlemmarna har haft utvecklingssamtal med sin chef under det senaste året. Vilken signifikansnivå har testet?

c) Beräkna hur stort urval som skulle krävas för att längden av ett 99 %-igt konfidensintervall inte ska överstiga 0.08. Antag vid beräkningen att den sanna andelen är okänd.

d) Hur påverkas erforderlig urvalsstorlek i c)-uppgiften om den statistiska felmarginalen minskar?

e) Hur påverkas erforderlig urvalsstorlek i c)-uppgiften om konfidensgraden minskar?

Uppgift 3. (20 poäng)

Plantor omvandlar koldioxid från luften tillsammans med vatten och energi från solljus till energi som de behöver för att växa. I ett experiment sattes ett antal plantor i normal luft och ett antal plantor i luft som berikats med extra koldioxid för att undersöka effekten på tillväxten. Vattningsmängd och ljusförhållanden var lika för samtliga plantor under experimentet. Nedanstående tabell visar plantornas tillväxt i gram.

Normal luft	4.67	4.21	2.18	3.91	4.09	5.24	2.94	4.71	4.04	5.79
Berikad luft	5.04	4.52	6.18	7.01	4.36	1.81	6.22	5.70		

a) Testa med hjälp av ett lämpligt icke-parameteriskt test om det går att påvisa att koldioxidberikad luft ökar planttillväxten.

b) Testa med hjälp av ett lämpligt parameteriskt test om det går att påvisa att koldioxidberikad luft ökar planttillväxten.

Uppgift 4. (20 poäng)

Antag att tiden det tar i minuter att besvara en viss typ av tentafråga är exponentialfördelad med väntevärde β. I ett slumpmässigt urval av 40 frågor blev den den totala tiden 1220 minuter.

a) Härled och beräkna maximum likelihood-skattningen av den förväntade svarstiden.

b) Vad blir maximum likelihood-skattningen av andelen frågor som besvaras inom en halvtimme?

Uppgift 5. (20 poäng)

För att ta körkort måste man bl.a. först klara ett kunskapsprov med teoretiska frågor. Antag att antalet gånger som en körkortstagare behöver göra kunskapsprovet för att bli godkänd följer en geometriskt fördelning, där parametern p är sannolikheten att klara kunskapsprovet. Man observerar antalet gånger som 40 slumpmässigt utvalda provtagare behöver göra provet och använder maximum likelihood-skattningen $\hat{p}_{ML} = 1/\overline{Y}$ för att uppskatta p. Man vill nu använda likelihoodkvottestet för att testa $H_0 : p = 0.5$ mot $H_a : p \neq 0.5$.

a) Bestäm likelihoodfunktionen för $p = 0.5$.

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

c) Bestäm teststatistikan för likelihoodkvottestet.

d) Ange kritiskt område (RR) för signifikansnivån 0.01.

e) Medelvärdet av antalet gånger som de 40 provtagarna i urvalet behövde göra provet blev $\overline{y} = 1.475$. Genomför testet och tolka resultatet.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 30/10/17

Sal: Ugglevik

Tenta: Statistisk teori med tillämpningar

Kurs: Statistisk teori med tillämpningar II

ANONYMKOD: S7-0014

☒ 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. 16	17	19	14	19					

85 + 5 bonus

POÄNG	BETYG	Lärarens sign.
90	A	

SU, STATISTIK

Skrivsal: U6 Anonymkod: ST-0014 Blad nr: 1

Uppgift 1

a) Samplingfördelning är en sannolikhets-fördelning för en statistika. Statistika är en skattning av en okänd populationsparameter med hjälp av ett stickprov. 4

b) Centrala gränsvärdessatsen säger att ju större stickprov n vi tar ($n \to \infty$) desto mer approximerar medelvärdet till normalfördelningen ($\bar{Y} \sim N(\mu, \sigma^2)$), och z-score $\sim N(0,1)$ (standardiserad n.f.) Förutsättn.? 2

c) Medelkvadratfel (MSE) är mått på kvalité av en parameterskattning, används för beräkningar av konfidensintervaller. ?
Om vi har en väntevärdesriktig estimator θ, då $MSE = V(\hat{\theta})$. För en biased estimator blir $MSE = V(\hat{\theta}) + B(\hat{\theta})^2$. R 4

d) konsistens är en av "properties" av estimatorer.

En estimator är konsistent om gäller om $\hat{\theta}$ vr

$\lim_{n \to \infty} V(\hat{\theta}) = 0$, vilket betyder att variansen minskar och går mot noll när stickprovsstorlek ökar. 2

e) Momentskattning är en av metoder av punktskattningar för parametrar.

Populationsmoment:

μ_1' – första moment

μ_2' – andra moment

μ_k' – k-te moment

jämförs med en motsvarande urvalsmoment.

$$m_k' = \frac{\sum_{i=1}^{n} Y_i^k}{n}$$

Sedan gör man ekvation (eller ekvationssystem om flera parametrar skattas) och löser ut parameter som skattas:

$$\left.\begin{aligned} \mu_1' &= m_1' \\ &\dots \\ \mu_k' &= m_k' \end{aligned}\right\}$$

Momentskattningar är konsistenta, men inte alltid väntevärdesriktiga.

4

16

Uppgift 2

"X" - antal respondenter som hade utvecklingssamtal.
n - totalt antal respondenter.

a) Ett 99% k.i. ges av (då n är stort):

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\sigma^2_{\hat{p}}}$$

$$\hat{p} = \frac{X}{n} = \frac{287}{398} \approx 0,72$$

$$z_{\alpha/2} = z_{0,005} = 2,5758$$

$$\sigma^2_{\hat{p}} = s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1} = \frac{0,2016}{397} = 5,078 \cdot 10^{-4}$$

$$0,72 \pm 2,5758 \cdot \sqrt{5,078 \cdot 10^{-4}} \approx 0,72 \pm 0,058$$

Svar: ett 99% k.i för andelen medlemmar som hade utvecklingssamtal är:
[0,662 ; 0,778] R 5

b) Vi använder 99% k.i., då konfidensgraden är 0,99 och signifikansnivån för testet blir $\alpha = 0,01$. R

Våra hypoteser att testa:

$H_0 : p = 0,8$

$H_a : p \neq 0,8$ R

2-sidig hypotestest, då $\alpha/2 = z_{0,005}$

Beslutsregel: Förkasta H_0 om teststatistika $|z_{obs}| > z_{krit}$.

z_{krit} (från tabell) = 2,5758

$$z_{obs} = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{0,72 - 0,8}{0,0225} = -3,56$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{5,078 \cdot 10^{-4}} = 0,0225$$

$|-3,581| > 2,5758$ då förkastar vi H_0 på 0,01 signifikansnivån.

Svar: Signifikansnivån för testet är 0,01.

H_0 förkastas, (och 0 ingår inte i k.i. i a), då skulle vi inte kunna förkasta H_0). andelen medlemmar som hade utvecklingssamtal är inte 0,8. 3

c) B ska inte överstiga 0,08. Vi kan beräkna urvalsstorlek: B=0.04

$$n = \frac{z_{\alpha/2}^2 \cdot \sigma_p^2}{B^2} = \frac{2,5758^2 \cdot 0,25}{0,08^2} = 259,17$$

Om den sanna andelen är okänd, då tar vi den största möjliga variansen när $p = 0,5$. R

$$\sigma_p^2 = p(1-p) = 0,5 \cdot 0,5 = 0,25$$

Svar: Det skulle krävas ett urval av 260 personer. 3

d) Om vi vill att felmarginalen minskar, då måste n öka. R 3

e) Om konfidensgraden minskar (t ex från 99% till 95%) då ska $z_{\alpha/2}$ också minska, då minskar n också. R 3

17

Uppgift 3.

a) Vi har två oberoende stickprov och observationer är oberoende mellan sig. Vi har olika antal obs. i två stickprov, då ett lämpligt icke-par. test ska vara Mann-Whitneys U-test. R

Vi tar n_1 som har mindre obs. med plantor som hade Berikad luft.

Berikad luft		Normal luft			
n_1	R	n_2	R	y_1^2	y_2^2
5,04	12	4,67	10	25,40	21,81
4,52	9	4,21	7	20,43	17,72
6,18	16	2,18	2	38,19	4,75
7,01	18	3,91	4	49,14	15,29
4,36	8	4,09	6	19,27	16,73
		5,21	13		27,46
1,81	1	2,94	3	3,28	8,64
6,22	17	4,71	11	38,69	22,18
5,70	14	4,04	5	32,49	16,32
Σ_i: 40,84	95	5,79	15	226,89	33,52
		41,78	76		184,42

$\Sigma R = 171$

$\frac{n(n+1)}{2} = 171$

Våra hypoteser: H_0: det finns inget skillnad i läget för fördelningar för tillväxter av plantor som växer i normal eller berikad luft

H_a: fördelningen för tillväxten i berikad luft ligger åt höger åt den i normal luft R

Vi använder teststatistikan U, och beslutsregeln är: förkasta H_0 om $U \leq U_0$.

R

U_0 får vi från tabellen $n_1 = 8$, $n_2 = 10$, signifikansnivån $\alpha = 0{,}05$, och vår test är 1-sidig,

$U_{0\,8,10,0,05} = 21$ R

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W$$

$$U = 8 \cdot 10 + \frac{8 \cdot 9}{2} - 95 = 21$$ R

$H = H_0$. H_0 förkastas på 0,05 signifikansnivå

Svar: koldioxid berikad luft ökar plantillväxten vilket påvisas av icke-parametrisk test. 10

6) Vi kan testa med t-test om det finns skillnad mellan medelvärdena, obs. är inte parvisa. Om vi antar att populationen är normalfördelad, då $\bar{Y}$ n approx. $N(\mu, \frac{\sigma^2}{n})$, sann σ^2 är okänd. Vi antar att $\sigma_1^2 = \sigma_2^2 = \sigma^2$, observationerna är oberoende.

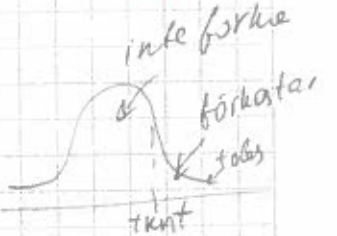
Hypoteser: $H_0: \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

$H_a: \mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 \neq 0$ >

Vi använder $\bar{Y}_1$ och $\bar{Y}_2$, och om H_0 är sann, då teststatistika $T = \frac{\bar{Y}_1 - \bar{Y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2 \text{ f.o.})$ R

Vi förkastar H_0 om $|t_{obs}| > t_{krit}$ på $\alpha = 0{,}05$ sig nivå

$t_{krit} = t_{16;0,05} = 1{,}75$ (tabell) en-sidig test R



$\bar{Y}_1 = 5{,}105 \quad \bar{Y}_2 = 4{,}178$ R

$(n_1 - 1)S_1^2 = \sum y_1^2 - n_1\bar{y}_1^2 = 226{,}89 - 8 \cdot 26{,}061 = 18{,}41$

$(n_2 - 1)S_2^2 = 184{,}42 - 10 \cdot 17{,}456 = 9{,}86$

$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{18{,}41 + 9{,}86}{8 + 10 - 2} = 1{,}7663 \approx R$

$T = \frac{5{,}105 - 4{,}178 - 0}{\sqrt{1{,}7663} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0{,}927}{1{,}329 \cdot 0{,}474} = \frac{0{,}927}{0{,}63} \approx 1{,}47$ R

($D_0 = 0$ under H_0 hypotes)

Slutsats: $T_{obs} < t_{krit} \Leftrightarrow 1{,}47 < 1{,}75$

Svar: H_0 kan inte förkastas. Vi kan inte påvisa med parametrisk test att koldioxiden ökar plantillväxten 9

Uppgift 4

"y" - tiden i minuter det tar att svara en viss typ av tentafråga"

$y \sim exp(\beta)$. I $n = 40$ blev $\sum y_i = 1220$.

a) Härleda $\hat{\beta}_{ML}$ och beräkna

$$f(y) = \frac{1}{\beta} \cdot e^{-y/\beta}, \quad \beta > 0, \quad 0 < y < \infty$$

* Likelihood funktionen blir (den simultana täthetsfunktionen):

$$L(y) = \prod_{i=1}^{n} \frac{1}{\beta} \cdot e^{-y/\beta} = \frac{1}{\beta^n} \cdot e^{-\sum \frac{y_i}{\beta}}$$ R

* Logaritmerar LF:

$$l(y) = \ln \beta^{-n} + \ln e^{-\sum \frac{y_i}{\beta}} = -n \ln \beta + \left(-\sum \frac{y_i}{\beta}\right) \ln e =$$

$$= -n \ln \beta - \sum \frac{y_i}{\beta}$$ R

* deriverar $l(y)$ m.a.p β:

$$\frac{dl(y)}{d(\beta)} = -\frac{n}{\beta} - \sum y_i \cdot (-\beta^{-2}) = -\frac{n}{\beta} + \frac{\sum y_i}{\beta^2}$$ R

* Sätter derivata lika med noll och löser ur β att hitta max.:

$$-\frac{n}{\beta} + \frac{\sum y_i}{\beta^2} = 0$$

$$\sum_i y_i - n\beta = 0$$

$$\hat{\beta}_{ML} = \frac{\sum_i y_i}{n} = \bar{y}$$ R

$t(\hat{\beta})_{ML} = t(\hat{\beta}_{ML})$, beräknar förväntade svarstiden

$$\frac{1220}{40} = 30{,}5$$ R

Svar: ML skattningen för den förväntade svarstiden ges av $\hat{\beta}_{ML} = \frac{\sum y_i}{n}$, och med vår urvalsdata det blir 30,5 minuter. 12

b) andelen frågor som besvaras inom halvtimme:

$\sum y_i = 30$ minuter, $\hat{\beta}_{ML} = 30,5$, lös $\hat{n}_{ML}$ - ?

$t(\beta)_{ML} = t(\hat{\beta}_{ML})$ $\quad n = \frac{\sum y_i}{\hat{\beta}_{ML}} = \frac{30}{30,5} = 0,984$

Svar: andelen frågor blir 0,984 ✓

2

14

Uppgift 5

y - "antalet gånger att göra provet"

$y \sim Geo(p)$. $n = 40$, s.u.

$\hat{p}_{ML} = \frac{1}{\bar{Y}} = \frac{n}{\sum y_i}$.

Använd LR test att testa $H_0: p = 0,5$

$H_a: p \neq 0,5$

a) Likelihoodfunktionen för $p = 0,5$

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

$L(p) = \prod_{i=1}^{n} p(1-p)^{y-1} = 0,5^{n} \cdot 0,5^{\sum y_i - n} = 0,5^{n + \sum y_i - n} = 0,5^{\sum y_i}$ R 2

b) Likelihoodfunktionen för $p = \hat{p}_{ML} \left(= \frac{n}{\sum y_i}\right)$

$L(\hat{p}_{ML}) = \prod_{i=1}^{n} \hat{p}_{ML}(1-\hat{p}_{ML})^{y-1} = \left(\frac{n}{\sum y_i}\right)^{n} \cdot \left(1 - \frac{n}{\sum y_i}\right)^{\sum y_i - n}$ R 2

c) Teststatistika för LR test ges av:

$$\lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} = \frac{0,5^{\sum y_i}}{\left(\frac{n}{\sum y_i}\right)^{n} \cdot \left(1 - \frac{n}{\sum y_i}\right)^{\sum y_i - n}}$$ 4

d) Kritisk område KR för sig. nivå 0,01 ges av.

$KR\{\lambda < k \mid H_0 \text{ är sann}\}$ där k bestäms av sig. nivå $\alpha = 0,01$

$KR\{-2\ln(\lambda) > \underbrace{-2\ln k}_{k^*}\} \sim \chi^2_1$ för sig. nivå 0,01 R

$k^* = \chi^2_{1;0,01} = 3,84$ ✓

$$RK = \{ -2 \ln(\lambda) > 3{,}84 \}$$

7

e) använder data att beräkna λ:

$$\lambda = \frac{0{,}5^{\Sigma y_i}}{\left(\frac{1}{\bar{y}}\right)^{\Sigma y} \cdot \left(1 - \frac{1}{\bar{y}}\right)^{\Sigma y - n}} = \frac{0{,}5^{59}}{0{,}678^{40} \cdot (1-0{,}678)^{59-40}} = \frac{0{,}5^{59}}{0{,}678^{40} \cdot 0{,}322^{19}} =$$

$$\Sigma y = n\bar{y} = 40 \cdot 1{,}475 = 59$$

$$\lambda = \frac{1{,}735 \cdot 10^{-18}}{1{,}775 \cdot 10^{-7} \cdot 4{,}459 \cdot 10^{-10}} = \frac{1{,}735 \cdot 10^{-18}}{7{,}915 \cdot 10^{-17}} = \frac{1{,}735 \cdot 10^{-1}}{7{,}915}$$

$$\approx 0{,}022$$

$$-2 \ln(\lambda) = -2 \ln(0{,}022) = 7{,}633 \quad \in R$$

$$RK \{ 7{,}633 > 3{,}84 \}$$

Svar: Vår beräknad λ ligger i förkastelseområdet. Vi kan förkasta H_0 på 0,01 signifikansnivån baserad på vår data från urvalet. Då accepterar vi den alternativa hypotesen att sannolikheten att klara kunskapsprovet är inte lika med 0,5

4

19