STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jörgen Säve-Söderbergh

TENTAMEN I STATISTISK TEORI MED TILLÄMPNINGAR II
2019-10-29

Skrivtid: 15.00-20.00
Godkända hjälpmedel: Miniräknare, språklexikon.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.

OBS! Glöm inte att ange nödvändiga antaganden där det behövs.

Uppgift 1. (20 poäng)
Förklara innebörden av följande begrepp:

a) Bias för en punktestimator

b) p-värde

c) Stora talens lag

d) Medelfel (Standard error)

e) Konfidensgrad

Uppgift 2. (20 poäng)
En tillverkare av solkräm utvecklade på försök en ny «formula» som eventuellt skulle ge ett ännu bättre skydd för solen än den gamla krämen. Tio försökspersoner utvaldes på ett slumpmässigt sätt till att delta i en prövning. De bägge solkrämerna (den nya och den gamla) smordes in på endera ryggen eller bröstet på varje person Vilken solkräm som hamnade var bestämdes av slumpen. Varje försöksperson utsattes därefter för ett intensivt, men kontrollerat solsken. Graden av solbränna mättes på en skala där högre värden innebär kraftigare solbränna.

Försöksperson	1	2	3	4	5	6	7	8	9	10
Gamla solkrämen	43	45	48	38	41	45	21	28	30	17
Nya solkrämen	37	39	31	39	34	47	18	32	25	8

a) Sätt upp lämpliga hypoteser och gör nödvändiga antaganden för att testa om den nya solkrämen är mer effektiv än den gamla.

b) Testa hypoteserna i uppgift a) med hjälp av ett lämpligt icke-parametriskt test.

Uppgift 3. (20 poäng)
Svenska Golfförbundet önskar studera hur lång tid medlemmarnas golfrundor egentligen tar. Sedan lång tid har orsakerna kring detta debatterats i förbundets månatliga publikation. Som en del i projektet önskar man studera hur stor variationen är, vilken skillnad i tid tar olika golfrundor. Tjugo personer väljs ut på ett slumpmässigt sätt och tillfrågas vilken tid deras senaste golfrunda (arton hål) tog. Standardavvikelsen för de tjugo observationerna blev 1.209611 timmar.

a) Ange nödvändiga förutsättningar och beräkna ett 90%-igt konfidensintervall för standardavvikelsen av tiderna för golfrundorna. Ge en tolkning av det intervall du har beräknat.

b) Testa nollhypotesen $H_0 : \sigma^2 = 3$ mot $H_a : \sigma^2 \neq 3$ på signifikansnivån $\alpha = 0.10$.

Uppgift 4. (20 poäng)

Två personer spelar ett spel. De är inte överens om sannolikheten för vinst i spelet. De önskar använda statistisk hypotesprövning för att slita tvisten. De gör n oberoende försök. Låt $Y_1, Y_2, \ldots, Y_n$ beteckna resultatet av n st Bernoulli-försök.

a) Bestäm det förkastelseområde RR som vi erhåller om vi sätter signifikansnivån α till ett visst värde och önskar erhålla maximal styrka i det värde som mothypotesen pekar ut. (Använd Neyman-Pearsons lemma). De hypoteser vi är intresserade av är $H_0 : p = 0.5$ mot alternativet $H_a : p = 0.4$.

b) Antag att den kritiska gränsen för testet i (a) har bestämts till 29. Beräkna sannolikheten β för att begå ett fel av andra slaget då vi gör $n = 70$ försök genom normalapproximation.

c) Är testet i (a) ett likformigt starkaste test om vi istället beaktar mothypotesen $H_a : p < 0.5$? (För full poäng ska du göra ett matematiskt bevis).

Uppgift 5. (20 poäng)
Låt $y_1 = 0.92$, $y_2 = 0.79$, $y_3 = 0.47$, $y_4 = 0.90$ och $y_5 = 0.86$ vara utfall av oberoende stokastiska variabler Y_1, Y_2, Y_3, Y_4, Y_5, respektive, som har den gemensamma täthetsfunktionen $f(y|\theta)$ given av

$$f(y|\theta) = \begin{cases} (\theta + 1)y^\theta & 0 \leq y \leq 1 \\ 0, & \text{annars} \end{cases}$$

där parametern θ är > -1. Härled maximum-likelihood-skattningen av parametern θ på basis av dessa data.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 29/10-2019

Sal: Värtasalen

Tenta: Statistisk teori med tillämpningar II

Kurs: Statistisk teori med tillämpningar

ANONYMKOD: 0013-YWC

[X] 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					9
Lär.ant. 10	20	20	4	20					

MW

POÄNG	BETYG	Lärarens sign.
80	B	JSS

+ 6 bonus

SU, STATISTIK

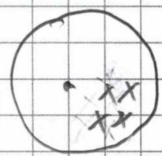
Skrivsal: Vårtasalen Anonymkod: 0013-YWL Blad nr: 1

Uppgift 1

a) Bias, definieras som:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

, alltså är biasen skillnaden mellan den förväntade estimatorn och det sanna värdet på parametern, i detta fall θ. Vid bias för en punktestimator så hamnar alla observationer fel, skattningen $\hat{\theta}$ av θ har alltså hög bias. Nedan visas detta på en bild:



, liten varians, hög bias.

Där mittpunkten är θ medan skattningarna $\hat{\theta}$ är krysen, som alla hamnat nära varandra, men skattningen av parametern har hög bias.

4

b) p-värde är sannolikheten att få samma tal som man observerar (ex. Z_{OBS}), eller ett mer extremt värde på observationen. Alternativhypotesen bestämmer riktningen på det "mer extrema värdet".

4

d) $MSE = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) - (B(\hat{\theta}))^2$

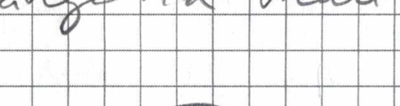
MSE är medelkvadratfelet av en estimator (ex. $\hat{\theta}$). Som visat i formeln så är medelkvadratfelet den förväntade skillnaden mellan estimatorn $\hat{\theta}$ och parametern θ, i kvadrat. Detta är också samma som variansen av estimatorn minus biasen av estimatorn i kvadrat. Om biasen skulle vara noll, alltså om estimatorn är väntevärdesriktig, så blir MSE lika med variansen av estimatorn

0

e) Konfidensgrad: $1 - \alpha$, där α är sannolikheten att förkasta en sann nollhypotes. Konfidensgraden, ofta satt till 90%, 95% eller 99% är en felmarginal, exempelvis: Om konfidensgraden är 95% så kommer konfidensintervall beräknade på identiskt sätt, från samma population, innehålla det sanna värdet på parametern 95% av gångerna man beräknar konfidensintervallet:

2

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 0{,}95$$



c) Detta har vi inte gått igenom på kursen.

0

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWL Blad nr: 2

Uppgift 2

a) Antagande:

I och med att den nya solkrämen sägs ge bättre skydd än den gamla, så antar jag att det jag ska testa är att graden på solbränna minskat, och om det gäller så är den nya solkrämen mer effektiv, och HA är sann.

Detta gör att jag räknar differenserna som: D_i = nya minus gamla = $X_i - Y_i$.

Antagandet för testet är slumpmässigt urval ur en population, där det gäller oberoende mellan observationerna i stickprovet men beroende inom varje observation. Stickprovet är alltså relaterat, den nya och den gamla solkrämen testas på samma person.

Hypoteserna som ska testas är:

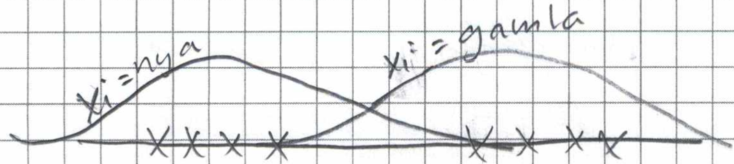
Ho: den nya solkrämen ger samma skydd som den gamla solkrämen, lika effektiv.

HA: Den nya solkrämen är mer effektiv, den ger alltså bättre skydd än den gamla, graden av solbränna är mindre.

Resultatmässigt, "graden av solbränna", ligger fördelningen för den gamla solkrämen till höger om den nya.

där X_i = nya solkrämen och

Y_i = gamla solkrämen.



R

Testet kommer utföras på 5% signifikansnivå.

b) Då jag vill testa om den nya är bättre än den gamla solkrämen kommer jag räkna differenserna som nya minus gamla. Jag använder mig av ett Wilcoxon-matched-pairs, signed rank test.

Teststatistikan $T = T^+$, där T^+ är summan av ranger av positiva differenser

$RR: \{ T \leq T_0 \}$

$D_i = X_i - Y_i$

Obs	Nya kräm (X_i)	Gamla kräm (Y_i)	D_i	$\lvert D_i \rvert$	Rank $\lvert D_i \rvert$
1	37	43	−6	6	6.5
2	39	45	−6	6	6.5
3	31	48	−17	17	10
4	39	38	1	1	1
5	34	41	−7	7	8
6	47	45	2	2	2
7	18	21	−3	3	3
8	32	28	4	4	4
9	25	30	−5	5	5
10	8	17	−9	9	9

Det gör att det observerade värdet på teststatistikan T^+ = summa rangtal positiva differenser $= 1 + 2 + 4 = 7 = T$

R

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWL Blad nr: 3

RR: $\{T_{OBS} \leq T_0\}$ där T_0 fås från tabell nr. 9. Då jag testar ett ensidigt test på 5% signifikansnivå med $n=10$ observationer får jag $T_0 = 11$

RR $\{7 \leq 11\}$ gör att T_{OBS} hamnar i förkastelseområdet om H_0 förkastas på $\alpha = 0,05$.

SVAR: H_0 förkastas på $\alpha = 0,05$ vilket gör att det finns empirisk evidens för att den nya solkrämen ger bättre skydd än den gamla solkrämen. Graden av solbränna var alltså lägre med den nya solkrämen.

20p

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWC Blad nr: 4

Uppgift 3

$n = 20$

$s = 1,209611$

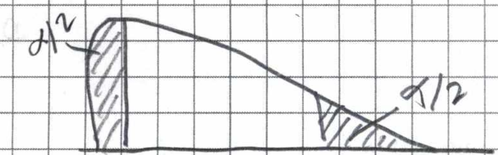
$s^2 = 1,46316$

a) Antaganden: slumpmässigt urval ur en normalfördelad population, oberoende observationer.

$\alpha = 0,1$

Konfidensintervallet för σ^2:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(df)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(df)}\right)$$



$$\left(\frac{(20-1)\,1,46316}{\chi^2_{0,1/2}(20-1)}, \frac{(20-1)\,1,46316}{\chi^2_{1-0,1/2}(20-1)}\right)$$

$$= \left(\frac{19 \cdot 1,46316}{\chi^2_{0,05}(19)}, \frac{19 \cdot 1,46316}{\chi^2_{0,95}(19)}\right)$$

$\chi^2_{0,05}(19) = 30,14$

$\chi^2_{0,95}(19) = 10,12$

$$= \left(\frac{27,80004}{30,14}, \frac{27,80004}{10,12}\right)$$

$$= [0,922;\ 2.747]$$

Detta är konfidensintervallet för variansen, konfidensintervallet för standardavvikelsen är:

$$\left[\sqrt{0,922}\,;\,\sqrt{2,747}\right]$$

$$= [0,960;\ 1,657]$$

SVAR: Ett 90% konfidensintervall för standardavvikelsen ges av:

$$[0,960;\ 1,657]$$

Tolkningen av detta är att 90% av gångerna man räknar konfidensintervall på detta sätt (upprepas identiskt), kommer konfidensintervallet innehålla det sanna värdet på parametern σ (standardavvikelsen). Om det intervallet jag nu beräknade med dessa värden, innehåller det sanna värdet på parametern går inte att svara på.

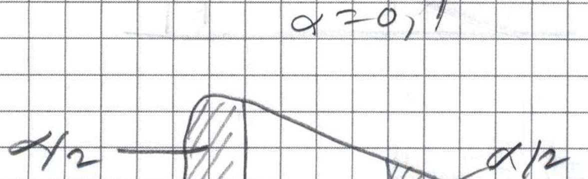
Det intervall jag räknat för standardavvikelsen av 20 personers tid för 18 håls golfrunda, är mellan 0,960 och 1,657 timmar.

b)

$$H_0:\ \sigma^2 = 3$$

$$H_a:\ \sigma^2 \neq 3$$

$$\alpha = 0,1$$

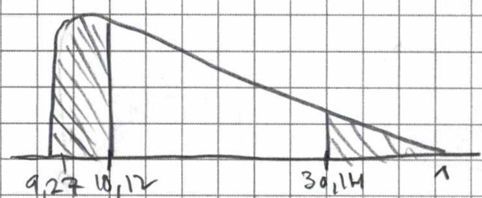


SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0013-YWC Blad nr: 5

testvariabel: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$

$$\chi^2_{OBS} = \frac{(20-1)\cdot 1,46316}{3} = \frac{19 \cdot 1,46316}{3} \approx 9,27$$ R



$\chi^2_{0,1/2}(20-1) = \chi^2_{0,05}(19) = 30,14$ R

$\chi^2_{1-\frac{0,1}{2}}(20-1) = \chi^2_{0,95}(19) = 10,12$ R

på χ^2_{OBS} hamnar inom förkastelseområdet

för $\chi^2_{0,1/2}(19)$ och $\chi^2_{1-\frac{0,1}{2}}(19)$, alltså:

$\chi^2_{OBS} = 9,27 < \chi^2_{0,95}(19) = 10,12$

vilket gör att Ho förkastas på $\alpha = 0,1$

SVAR: Ho förkastas på $\alpha = 0,1$, det finns empirisk evidens för att σ^2 är skiljd från 3. R

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWC Blad nr: 6

~~Uppgift 4~~

$Y \sim Bern(p)$ ~~(p)~~ 1

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

a) $H_0: p = 0,5$

$H_A: p = 0,4$

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

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

$= 0,5^{\sum y_i}(0,5)^{n-\sum y_i}$

$L(p=0,4) = 0,4^{\sum y_i}(1-0,4)^{n-\sum y_i}$

$= 0,4^{\sum y_i}(0,6)^{n-\sum y_i}$

$$\frac{L(p=0,5)}{L(p=0,4)} = \lambda_{LR} = \left(\frac{0,5}{0,4}\right)^{\sum y_i}\left(\frac{0,5}{0,6}\right)^{n-\sum y_i}$$

$\lambda_{LR} < k$ där k är en konstant bestämd av α.

$$\underbrace{-2\ln\lambda_R}_{w} > \underbrace{-2\ln k}_{k^*}$$

df = 1 − 0
= antal parametrar som ska skattas minus fixa parametrar under H_0.

där $w \overset{\text{approx}}{\sim} \chi^2(1)$, ← för stora urval kan fördelningen approximeras till en χ^2-fördelning

Ej LR-test

Ej $-2\ln\lambda$ Neyman-Pearson!

$$W = -2\ln\left(\left(\frac{0,5}{0,4}\right)^{\Sigma y_i}\left(\frac{0,5}{0,6}\right)^{n-\Sigma y_i}\right)$$

$$= -2\left[\Sigma y_i \ln\left(\frac{0,5}{0,4}\right) + (n-\Sigma y_i)\ln\left(\frac{0,5}{0,6}\right)\right]$$

SVAR: RR: $\{ W > u^* \}$

där $W = -2\ln\left[\left(\frac{0,5}{0,4}\right)^{\Sigma y_i}\left(\frac{0,5}{0,6}\right)^{n-\Sigma y_i}\right]$

om jag sätter $\alpha = 0,05$ så får jag:

RR$\{ W > 3,84 \}$. ✓

2p

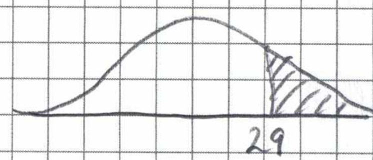
b) $\beta = P(\text{typ II-fel}) = P(\text{acceptera } H_0 \mid H_a \text{ sann})$

$$= P(W \leq 29 \mid p = 0,4)$$

RR: $\{ W > 29 \}$

$$\beta = P(W \leq 29 \mid p = 0,4)$$

$$= \frac{29 - c}{\sqrt{\frac{p(1-p)}{n}}} \leq -z_\beta$$



där c är en okänd konstant.

$$= \frac{29 - c}{\sqrt{\frac{0,4 \cdot 0,6}{70}}} \leq -z_\beta \qquad \{ p = 0,4$$

Jag antar $\alpha = 0,05$ från uppgift a)

$$\Rightarrow \frac{29 - c}{0,05855} = 1,6449$$

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWL Blad nr: 7

4b) forts.

$29 = 1,6449 \cdot 0,05855 + c$

$29 = 0,096 + c$

$c = 28,904$

$$\frac{29 - 28,904}{0,05855} \leq -z_\beta$$

$1,64 \leq -z_\beta$

$1 - \Phi 1,64 = 0,05$

SVAR: $\beta = 0,05$ ✓

0p

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWL Blad nr: 8

c) För att ett test ska vara ett likformigt starkaste test (UMP) så ska RR vara oberoende alternativ hypotesen. Om testet i a) istället skulle testas för följande hypoteser:

$$H_0: p = 0,5$$

$$H_A: p < 0,5$$

där H_A är en sammansatt hypotes av oändligt många enkla hypoteser, där alla är $p < 0,5$. Detta testas med följande testvariabel:

$$\frac{L(p=0,5)}{L(p=\hat{p}_{ML})} = \lambda_{LR}$$

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

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

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

$$\frac{\ell'(p)}{\partial p} = \Sigma y_i \cdot \frac{1}{p} + (n-\Sigma y_i)\frac{1}{1-p}(-1)$$

$$= \frac{\Sigma y_i}{p} - \frac{(n-\Sigma y_i)}{1-p} = 0$$

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

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

$$(1-p)\sum y_i = p(n-\sum y_i)$$

$$\sum y_i - p\sum y_i = pn - p\sum y_i$$

$$\sum y_i = pn$$

$$\bar{y} = \frac{\sum y_i}{n} = \hat{p}_{ML}$$ R

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

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

$$\lambda_{LR} = \frac{L(p=0,5)}{L(p=\hat{p}_{ML})} = \frac{(0,5)^{\sum y_i}(0,5)^{n-\sum y_i}}{(\hat{p}_{ML})^{\sum y_i}(1-\hat{p}_{ML})^{n-\sum y_i}}$$

$$= \left(\frac{0,5}{\hat{p}_{ML}}\right)^{\sum y_i}\left(\frac{0,5}{1-\hat{p}_{ML}}\right)^{n(1-\bar{y})}$$

$\sum y_i = n\bar{y}$

$n - n\bar{y}$

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

där $\hat{p}_{ML} = \bar{y}$

$$\Rightarrow \left(\frac{0,5}{\bar{y}}\right)^{\sum y_i}\left(\frac{0,5}{1-\bar{y}}\right)^{n(1-\bar{y})} = \lambda_{LR}$$

2p

RR: $\{\lambda_{LR} < k\}$ där k är en konstant bestämd av α.

Ovan test har skapat ett RR oberoende av H_A, vilket gör att man kan välja vilket H_A man vill (Bland $H_A: p < 0,5$) och ändå är RR samma. Det gör att testet är det likformigt starkaste testet (UMP). SVAR: Ja, om testet i a) istället beaktar mothypotesen $H_A: p < 0,5$ så är det UMP. Matematiskt bevis finns ovan.

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0013-YWL Blad nr: 9

Uppgift 5

$y_1 = 0{,}92 \quad y_2 = 0{,}79 \quad y_3 = 0{,}47$

$y_4 = 0{,}90 \quad y_5 = 0{,}86$

$n = 5 \qquad Y_1, Y_2, Y_3, Y_4, Y_5$

$$f(y \mid \theta) = (\theta+1)\, y^{\theta} \qquad 0 < y < 1$$

$$L(\theta; y) = \prod_{i=1}^{5} (\theta+1)\, y_i^{\theta}$$

$$= (\theta+1)^n \prod_{i=1}^{5} y_i^{\theta}$$

$$\ell(\theta; y) = n \cdot \ln(\theta+1) + \theta \sum_{i=1}^{n} \ln y_i$$

$$\frac{\ell'(\theta; y)}{\partial\theta} = n \cdot \frac{1}{\theta+1}(1) + \sum_{i=1}^{5} \ln y_i$$

$$= \frac{n}{\theta+1} + \sum_{i=1}^{5} \ln y_i = 0$$

$$\frac{n}{\theta+1} = -\sum_{i=1}^{5} \ln y_i \Rightarrow n = -\sum_{i=1}^{5} \ln y_i\, (\theta+1)$$

$$-\frac{n}{\sum_{i=1}^{5} \ln y_i} = \theta + 1 \Rightarrow -\frac{n}{\sum_{i=1}^{5} \ln y_i} = \hat{\theta}_{ML}$$

$$\sum_{i=1}^{5} \ln y_i = -1{,}33 \qquad n = 5$$

$$\hat{\theta}_{ML} = -\frac{5}{-1{,}33} - 1 \approx 2{,}76$$

SVAR: $\hat{\theta}_{ML} = -\dfrac{n}{\sum_{i=1}^{5} \ln y_i} - 1 \approx 2{,}76$

R

Bra!

20p