STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jessica Franzén

TENTAMEN I STATISTIKENS GRUNDER 2

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.

Uppgift 1

Efter flera års erfaranhet vet man att för ett speciellt inteligenstest på barn är testresultatet normalfördelat med populationsmedelvärde 3.4. Man studerade nyligen 13 slumpmässigt valda barn som genomfört testet och fick följande resultat:

4.8 3.1 6.0 1.8 2.6 3.9 4.0 3.7 2.0 5.5 4.3 2.9 3.2

a) Ange medelvärde, median och standardavvikelse för detta urval.

b) Räkna ut ett 90 procentigt konfidensintervallet för populationsmedelvärdet baserat på de 13 testresultaten. Förklara begreppet konfidensintervall.

c) Man misstänker att testresultaten på senare tid har ökat. Undersök detta genom att utföra ett lämpligt test på signifikansnivån 5%.

Uppgift 2

Grön starr (Glaukom) har egentligen ingenting med starr att göra, utan beror på ett ökat vätsketryck i ögat. Två trycksänkande mediciner A och B mot grön starr testades på ett slumpmässigt urval av 10 hundar som har denna ögonsjukdom på båda ögonen. På varje hund användes Medicin A på ett slumpmässigt valt öga och medicin B på det andra ögat. Trycket i ögonen mättes en timme senare och resultaten visas i tabellen.

X_1 = Trycket i ögat 1 timme efter medicin A gavs
X_2 = Trycket i ögat 1 timme efter medicin B gavs

Hund	1	2	3	4	5	6	7	8	9	10
X_1	0.17	0.20	0.14	0.18	0.23	0.19	0.12	0.10	0.16	0.13
X_2	0.15	0.18	0.13	0.18	0.19	0.12	0.07	0.09	0.14	0.08

a) Testa med ett lämpligt test på signifikansnivån 5% om det är någon skillnad mellan medicin A och B. Ange vilka antaganden du gör.

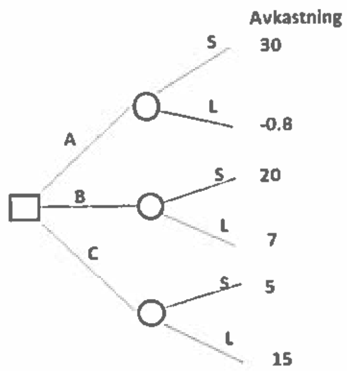
b) Korrelationen mellan X_1 och X_2 är 0.8586. Kommentera storleken på korrelationen. Resonera kring vad som är den främsta anledningen till korrelationen i just detta exempel. Råder kausalitet? Förklara varför eller varför inte.

Uppgift 3
En pokermaskin ger ut spelkort. Det är meningen att maskinen ska ge ut spelkorten slumpmässigt och utan stopp d.v.s. som om spelkorten kom från en aldrig sinande kortlek. I ett test av maskinen lät man den dela ut 1600 spelkort och man fick då 408 spader, 432 hjärter, 384 ruter och 376 klöver.

a) Är andelen klöver som maskinen ger ut 1/4? Testa med lämpligt test på signifikansnivån 5%.

b) Testa med lämpligt test om maskinen delar ut korten slumpmässigt med avseende på att proportionerna av spader/hjärter/ruter/klöver överenstämmer med en vanlig kortlek. Använd signifikansnivån 5%.

Uppgift 4
Ett företag har utvecklat en ny produkt. Ledningen för företaget ska besluta om en produktions- och marknadsstrategi för den nya produkten. Man överväger tre strategier - A (aggressiv), B (normal) och C (försiktig). Efterfrågan för produkten kan antingen vara stor (S) eller liten (L). Följande beslutsträd med de resulterande nyttorna i form av avkastning (miljoner kronor) sattes upp av företagsledningen.



a) Sätt upp en korrekt uppställd beslutsmatris med hjälp av informationen i beslutsträdet.

b) VD:ns främsta fokus är att välja en strategi där skillnaden mellan den faktiska och den potentiellt bästa avkastningen ska bli så liten som möjligt. Vilken strategi bör man välja enligt VD:ns fokus?

c) Företagets statistiker har gjort en marknadsundersökning och han förutspår att sannolikheten är 0.62 att efterfrågan av den nya produkten blir stor. Vilken strategi bör företaget välja baserat på den informationen?

d) Skattningen i c) grundar sig på att företagets statistiker gjort ett urval av n stycken potentiella butiker och frågat om de är intresserade att köpa in en stor mängd (stor efterfrågan) eller ingen/liten mängd (liten efterfrågan) av varan. Urvalsproportionen för stor efterfrågan blev då 0.62. VD:n vill veta mer om osäkerheten kring skattningen och statistikern presenterar då det 95%-iga konfidensintervallet som blev [0.522;0.718]. För att räkna ut konfidensintervallet använde statistikern största möjliga varians. Hur många butiker n hade statistikern i sitt urval?

Uppgift 5
Temperaturen vid tillverkning av en plastmodul är en normalfördelad stokastisk variabel med väntevärde μ och variansen $\sigma^2 = 16$. Om processen är under kontroll är $\mu = 75$ grader medan förväntad temperatur sjunker till $\mu = 70$ grader om ett värmeelement går sönder. För att undersöka om elementet fungerar tar man $n = 9$ observationer på temperaturen och prövar nollhypotetsen

$$H_0 : \mu = 75$$

mot alternativhypotesen

$$H_A : \mu = 70$$

Nollhypotesen förkastas om medelvärdet av observationerna understiger 72.27 grader.

a) Bestäm testets signifikansnivå.

b) Bestäm testets styrka.

c) Medelvärde av de 9 observationerna blev 72.09. Bestäm testets p-värdet. Förklara begreppet p-värde.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 17/3-2017

Sal: Ugglevikssalen

Tenta: Statistikens grunder 2

Kurs: Statistikens grunder

ANONYMKOD: SG-0015

☒ 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					6
Lär.ant. 20	19	20	20	20					

POÄNG	BETYG	Lärarens sign.
99	A	[illegible]

SU, STATISTIK

Skrivsal: UG **Anonymkod:** SG-0015 **Blad nr:** 1

1 $n = 13$ ~~X = testresultat~~

a) $X \sim N(\mu = 3{,}4,\ \sigma^2)$

$$\bar{X} = \frac{\sum x_i}{n} = \frac{4{,}8+3{,}1+6{,}0+1{,}8+2{,}6+3{,}9+4{,}0+3{,}7+2{,}0+5{,}5+4{,}3+2{,}9+3{,}2}{13}$$

$$\approx \underline{\underline{3{,}67692}}$$

SVAR: Medelvärdet är ca $\underline{\underline{3{,}68}}$ R

Median = det tal som delar gruppens värden mitt itu.

1,8 2,0 2,6 2,9 3,1 3,2 3,7 3,9 4,0 4,3 4,8 5,5 6,0

SVAR: Medianen är $\underline{\underline{3{,}7}}$. R

Standardavvikelse

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{(4{,}8^2+3{,}1^2+6{,}0^2+1{,}8^2+2{,}6^2+3{,}9^2+4{,}0^2+3{,}7^2+2{,}0^2+5{,}5^2+4{,}3^2+2{,}9^2+3{,}2^2) - (13 \cdot 3{,}67692^2)}{13-1} = \frac{194{,}94 - 13(3{,}67692)^2}{12} =$$

$$= \frac{19{,}1831}{12} \approx 1{,}59859$$

$$s = \sqrt{1{,}59859} \approx 1{,}26435$$

SVAR: Standardavvikelsen är ca $\underline{\underline{1{,}26}}$ R 6

SG-0015

b) Ett 90%-igt ki ges av

$n < 30$
CGS kan ej tillämpas.

$\bar{x} \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$ $\quad \frac{\alpha}{2} = \frac{0,10}{2} = 0,05$

$3,67692 \pm 1,782 \cdot \frac{1,26435}{\sqrt{13}}$ $\quad t_{0.05}^{(12)} = 1,782$

$\Rightarrow 3,67692 \pm 0,624891$

K

SVAR: Konfidensintervallet är $[3.05 ; 4.30]$

Ett 90%-igt konfidensintervall kan tolkas enligt följande: om vi utför försöket många gånger på samma sätt med samma villkor. (dvs oberoende observationer och ett lika stort urval) och varje gång räknas ut ki för medelvärdet, så kommer 90 av 100 konfidensintervall att täcka det samma populationsmedelvärdet μ. ?

c) Hypotestest - vi testar om testresultaten har ökat $\alpha = 0.05$

$H_0: \mu_0 = 3.4$ K

$H_A: \mu_A > 3.4$ $\Rightarrow$ enkelsidigt test $\quad t_{0.05}^{(n-1)} = t_{0.05}^{(13-1)}$

$t_{0.05}^{(12)} = 1,782$ Testvariabel $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$t_{obs} = \frac{3.67692 - 3.4}{1.26435/\sqrt{13}} = \frac{0.27692}{0.350669} \approx \underline{\underline{0.789692}}$

nästa blad

SU, STATISTIK

Skrivsal: VG Anonymkod: SG-0015 Blad nr: 2

forts. uppgift 1c) $H_0: \mu_0 = 3.4$ $H_A: \mu_A > 3.4$

Beslutsregel: Förkasta H_0 om $t_{obs} > 1,782$

$t_{obs} = 0.789692 < 1,782 \Rightarrow H_0$ kan ej förkastas.

SVAR: $t_{obs} \approx 0.79$. Vi kan inte förkasta hypotesen om att testresultatet fortfarande har medelvärde på 3.4. Vi har alltså inte stöd för att testresultaten har ökat

7

2. Parvisa observationer ✓ X_1 och X_2 är beroende. (samma hund)

a) Antaganden: normalfördelade variabler X_1 och X_2. oberoende mellan hundarna. ✓

$H_0: \mu_1 - \mu_2 = 0$ Dubbelsidigt test. $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

$H_A: \mu_1 - \mu_2 \neq 0$ ✓

	1	2	3	4	5	6	7	8	9	10
X_1	0.17	0.20	0.14	0.18	0.23	0.19	0.12	0.10	0.16	0.13
X_2	0.15	0.18	0.13	0.18	0.19	0.12	0.07	0.09	0.14	0.08
D	0.02	0.02	0.01	0.0	0.04	0.07	0.05	0.01	0.02	0.05

$\bar{D} = \frac{\sum D_i}{n} = \frac{0.29}{10} = 0.029$

SG-0015

$$S^2_D = \frac{\sum D_i^2 - n\bar{D}^2}{n-1} = \frac{0.0129 - 10 \cdot 0.029^2}{10-1} = \frac{0.00449}{9} =$$

$$= 0.000499$$

$$T = \frac{\bar{D} - \overset{=0}{D_0}}{\sqrt{S^2 D / n}} \qquad t_{obs} = \frac{0.029}{\sqrt{\frac{0.000499}{10}}} = \underline{\underline{4.10578}} \; K$$

$t_{0.025}^{(9)} = \underline{\underline{2.262}}$ K

H_0: ingen skillnad $\mu_A - \mu_B = 0$

H_A: skillnad $\mu_A - \mu_B \neq 0$

Beslutsregel: Förkasta H_0 om $t_{obs} > 2.262$

Slutsats: $4.11 > 2.26 \Rightarrow H_0$ förkastas på 5%-niv

Det finns en skillnad mellan medicin A och B.

14

b) Korrelation beskriver det linjära sambandet mellan två variabler. Korrelationen är ett värde mellan -1 och 1 där 0 betyder inget linjärt samband, -1 betyder starkt negativt samband och 1 betyder starkt positivt samband. En korrelation på 0,8586 är starkt positivt. Det innebär att om X_1 ökar så kommer också X_2 att öka.

Detta är logiskt i exemplet då X_1 och X_2 är samma hund, och om hunden har

nästa blad

SU, STATISTIK

Skrivsal: UG Anonymkod: 00-15 Blad nr: 3

forts uppgif 2b

→ högt tryck vid behandling av den ena medicinen så kommer den troligtvis ha högt tryck även vid behandling av den andra medicinen (även om trycket kan vara olika)

Kausalitet betyder orsakssamband, att en variabel har en direkt påverkan på en annan variabel. Det finns 3 kriterier för kausalitet. orsaksvariabel → responsvariabel

1. Samvariation (tex korrelation)

2. Intern validitet - eliminera påverkan från andra variabler.

3. Tidsordning, orsak måste komma före verkan.

Kausalitet gäller i en riktning tex inkomst påverkar konsumtion, men konsumtion påverkas inte inkomst.

I exemplet råder det inte kausalitet eftersom beroendet gäller i båda riktningarna.

Ett högt värde på X_1 → högt värde på X_2

Ett högt värde på X_2 → högt värde på X_1

5

SG-0015

3. $n = 1600$ 408 spader
432 hjärter
384 ruter
376 klöver

Vi testar om andelen klöver som maskinen ger ut är 0.25

a) $H_0: \pi_0 = 0.25$ Dubbelsidigt test.

$H_A: \pi_A \neq 0.25$ ✓ $\alpha/2 = 0.025$

$$p = \frac{\text{antal kort med klöver}}{\text{totalt antal kort i urvalet}} = \frac{376}{1600} = 0.235$$

$np = 1600 \cdot 0.235 = 376 \geq 5$ ✓ $\Rightarrow$ normalfördelnings approximation tillåten enligt CGS.

$n(1-p) = 1600 \cdot (1-0.235) = 1224 \geq 5$

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \text{approx } N(0,1) \text{ om } H_0 \text{ är sann.}$$

Kritiskt värde $= z_{0.025} = 1.96$

Beslutsregel: Förkasta H_0 om $z_{obs} < -1.96$ ✓

$$z_{obs} = \frac{0.235 - 0.25}{\sqrt{\frac{0.25(1-0.25)}{1600}}} = \frac{-0.015}{\sqrt{0.000117}} = \underline{\underline{-1.38564}}$$ ✓

SVAR: $z_{obs} \approx -1.39$. H_0 kan inte förkastas på 5% nivån. Vi kan inte förkasta hypotesen om att andelen klöver som maskinen ger ut är 0.25.

10

nästa blad

SU, STATISTIK

Skrivsal: UG **Anonymkod:** SG-0015 **Blad nr:** 4

3b) Goodness of fit test. $\alpha = 0.05$

Vi vill testa om kortutdelningen följer en viss fördelning.

	SPADER	HJÄRTER	RUTER	KLÖVER	
obs. n	408	432	384	376	$\Sigma 1600$
förv. $E(n)$	400	400	400	400	$\Sigma 1600$

I en vanlig kortlek är fördelningen:

$\frac{1}{4}$ spader, $\frac{1}{4}$ hjärter, $\frac{1}{4}$ ruter, $\frac{1}{4}$ klöver. $\left(\frac{1}{4} = 0.25\right)$

Våra förväntade värden blir:

$0.25 \cdot 1600 = 400$ för varje kategori

Vi använder oss av ett χ^2 test $v = (\text{antal kategorier} - 1)$

$$\chi^2 = \Sigma \frac{(n - E(n))^2}{E(n)} \qquad \chi^2_{0.05}(4-1) = \chi^2_{0.05}(3) = \underline{\underline{7.815}}$$

$$\chi^2_{obs} = \frac{(408-400)^2}{400} + \frac{(432-400)^2}{400} + \frac{(384-400)^2}{400} + \frac{(376-400)^2}{400}$$

$$= \frac{24}{5} = \underline{\underline{4.8}}$$

K

SG-0015

H_0: fördelningen är samma som i en vanlig kortlek, dvs $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$

H_A: fördelningen är inte samma

Beslutsregel: Förkasta H_0 om $\chi^2_{obs} > 7,815$

SVAR: $\chi^2_{obs} = 4.8 < 7,815 \Rightarrow H_0$ kan inte förkastas på signifikansnivån 5%. Vi har inte stöd för att fördelningen är olik en vanlig kortlek.

10

4 a) Beslutsmatris: Efterfrågan

strategi	Stor	Liten
A	30	-0.8
B	20	7
C	5	15

regretmatris

	S	L
A	0	15.8
B	10	8
C	25	0

Välj B

2

b) VD:n skall välja minimax regret. Man räknar då ut $S_1^* = 30$ $S_2^* = 15$ och skapar en regretmatris (se ovan till höger). Man väljer sedan det handlingsalternativ som gör att den maximala alternativförlusten blir så liten som möjligt.

6

→ nästa blad

SU, STATISTIK

Skrivsal: UG **Anonymkod:** SG-0015 **Blad nr:** 5

4c)

	Efterfrågan	
	(0.62) S	(0.38) L
A	30	−0.8
B	20	7
C	5	15

P(Liten efterfrågan) $= 1 - 0.62 = 0.38$

$E(U_1) = 30 \cdot 0.62 + 0.38 \cdot (-0.8) = 18.296$

$E(U_2) = 20 \cdot 0.62 + 7 \cdot 0.38 = 15.06$

$E(U_3) = 5 \cdot 0.62 + 15 \cdot 0.38 = 8.8$

Nu när vi vet slh för de olika tillstånden (efterfrågan) kan vi för varje handlingsalternativ räkna ut den förväntade nyttan.
Vi väljer det handlingsalternativ med störst förväntad nytta (enligt Laplace kriteriet)
I exemplet blir det handlingsalternativ A.

6

4d) $p = 0.62$ 95%-igt KI $[0.522 ; 0.718]$

$B = 0.718 - 0.62 = 0.098$ $2B = 0.718 - 0.522 = 0.196$

Största möjliga varians används $(0.5 \cdot 0.5)/n$

$np > 5$

$n(1-p) > 5$

$z_{0.025} = 1.96$

$B = 1.96 \cdot \sqrt{\frac{0.5 \cdot 0.5}{n}} \Longrightarrow 1.96 \cdot \sqrt{\frac{0.25}{n}} = 0.098$

$\frac{0.25}{n} = \left(\frac{0.098}{1.96}\right)^2 \Longrightarrow n = 100$

6

SVAR: Statistikern hade 100 butiker i sitt urval.

S -0015

5 X = temperaturen vid tillverkning av en plastmodul

$X \sim N(\mu, \sigma^2 = 16)$ $n = 9$ $\sigma = \sqrt{16} = 4$

σ är känt, vi kan använda z som testvariabel, trots att urvalet är litet.

$H_0: \mu_0 = 75$

$H_A: \mu_A = 70$

Nollhypotesen förkastas om $\bar{x} < 72.27$

(a) $\alpha = P(H_0 \text{ förkastas} \mid H_0 \text{ är sann}) =$

$= P(\bar{X} < 72.72 \mid H_0 \text{ är sann}) = P\left(z < \frac{72.27 - 75}{4/\sqrt{9}}\right) =$

$= P\left(z < \frac{-2.73}{4/3}\right) = P(z < -2.0475) \approx$

$\approx 1 - \Phi(2.05) = \{\text{från tabell}\} = 1 - 0.97982 = \underline{\underline{0.02018}}$

SVAR: Signifikansnivån är ca 2 %

(b) Testets styrka $= 1 - \beta$

vilket här är samma sak som att H_A är sann

$\beta = P(H_0 \text{ inte förkastas} \mid H_0 \text{ är falsk}) =$

$= P(\bar{X} > 72.27 \mid \mu = 70) = P\left(z > \frac{72.27 - 70}{4/\sqrt{9}}\right) =$

$= P(z > 1.7025) \approx 1 - \Phi(1.70) = \{\text{från tabell}\} =$

$= 1 - 0.95543 = 0.04457$

→ nästa blad

SU, STATISTIK

Skrivsal: VG **Anonymkod:** SG-0015 **Blad nr:** 6

forts. 5b)

Testets styrka $= 1-\beta = 1-0.04457 = 0.95543$

SVAR: Testets styrka är ca 0.96

7

c) $\bar{x} = \frac{\sum x_i}{n} = 72.09$

givet att H_0 är sann

P-värdet är slh att få det observerade värdet eller något ännu mer extremt på teststatistikan. P-värdet är det lägsta värdet som H_0 kan förkastas på.

$$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{72.09-75}{4/\sqrt{9}} = -2.1825$$

$$P(z < -2.18) = 1-\Phi(2.18) = \{\text{från tabell}\} =$$

$$= 1-0.98537 = 0.01463.$$

6

Detta betyder alltså att vi kan förkasta H_0 på alla signifikansnivåer som är högre 1,463 %. T.ex förkastas H_0 på signifikansnivå 2,5 % men inte på 1 %.