STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jessica Franzén

TENTAMEN I STATISTIKENS GRUNDER 2

2016-10-26

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

Förklara kortfattat följande begrepp:
a) Samplingfördelning
b) P-värde
c) Förklara tolkningen av ett 95%-igt konfidensintervall
Förklara kortfattat följande begrepp och skillnaden mellan dem.
d) Korrelation och kausalitet
e) Enkelt index och sammansatt index

Uppgift 2

En fastighetsinvesterare överväger 3 olika investeringsalternativ - bygga lägenheter, kontor eller ett köpcentrum. Avkastningen (i tusentals kronor) för varje investering beror på den framtida populationsökningen i området vilken delas in i tre scenarion - långsamt, normalt eller snabbt växande population. Avkastningen redovisas i tabellen.

	Långsamt	Normalt	Snabbt
Lägenheter	-750	100	1500
Kontor	-2000	500	1000
Köpcentrum	150	300	750

a) Vilket investeringalternativ bör väljas enligt maximinkriteriet?

b) Vilket investeringalternativ bör väljas enligt Laplacekriteriet?

c) Vilket investeringalternativ bör väljas enligt maximaxkriteriet?

d) Vilket investeringalternativ bör väljas enligt minimax regretkriteriet?

Uppgift 3
Det påstås att Stockholmare är världens mest tatuerade storstadsfolk. I genomsnitt är 14% av världens storstadsinvånare (18-49 år) tatuerade. För att skatta andelen tatuerade stockholmare tog man ett slumpmässigt urval av 550 stockholmare i åldern 18-49 år, 182 av dessa var tatuerade.

a) Uppskatta proportionen tatuerade stockholmare i åldern 18-49 år.

b) Uppskatta variansen hos urvalsproportionen. Bör ändlighetskorrektion användas? Motivera.

c) Gör ett 90 procentigt konfidensintervall för proportionen tatuerade stockholmare i åldern 18-49 år.

d) Hur stort urval bör man dra om man önskar att längden på hela konfidensintervallet inte ska överstiga fem procentenheter?

Uppgift 4
Folkhälsomyndigheten rapporterar kontinuerligt andelen vuxna svenskar i fyra olika viktkategorier - undervikt (BMI <18.5), normalvikt (BMI 18.5-24.9), övervikt (BMI 25-29.9) och fetma (BMI>30). År 2000 var den procentuella viktfördelningen av Sveriges vuxna befolkning följande.

Undervikt	*Normalvikt*	*Övervikt*	*Fetma*
7%	49.5%	34%	9.5%

a) 2014 gjordes en undersökning bland 600 slumpmässigt utvalda vuxna svenskar. Urvalets viktfördelning blev då följande.

Undervikt	*Normalvikt*	*Övervikt*	*Fetma*
2%	48.5%	35.5%	14%

Testa med lämpligt hypotestest om fördelningen över viktkategorierna 2014 är densamma som den var 2000. Använd signifikansnivån 1%.

b) I 2014 års undersökning var 290 kvinnor och 310 män. Testa med lämpligt test om fördelningen mellan viktklasserna är samma för kvinnor och män. Använd signifikansnivån 1%.

	Undervikt	*Normalvikt*	*Övervikt*	*Fetma*	Antal
Kvinnor	9	157	84	40	290
Män	3	133	130	44	310

Uppgift 5
Det har gjorts 40 mätningar av mängden av ett visst tillsatsämne i en livsmedelsprodukt. Mätningarna betraktas som ett utfall av den stokastiska variabeln X med väntevärde μ och varians σ^2. Vi antar att alla X är oberoende. Urvalsmedelvärdet blev 1.033 gram och urvalsvariansen 0.01188. Det rekommenderade nivån av tillsatsämnet i det aktuella livsmedlet är 1.0 gram. Vid tillverkningen av livsmedlet är det en maskin som proportionerar tillsatsämnet i varje enhet. Maskinen kan ställas in på önskad nivå (gram per enhet) men det blir alltid en mindre variation av mängden som tillsätts de olika enheterna.

a) Tyder data på att maskinen är inställd på en högre nivå än den rekommenderade? Utför ett lämpligt test och beräkna p-värdet, kommentera resultatet.

Vi tänker oss att vi har 40 andra mätningar $Y_1, Y_2, ..., Y_{40}$ från samma produktion. Dessa Y har samma väntevärde som X men osäkerheten i dessa mätningar är större vilket medför en varians för Y på $2.25\sigma^2$. Vi antar att alla Y är oberoende och att X och Y är oberoende. Två estimatorer för μ har föreslagits

$$\widehat{\mu}_1 = \frac{1}{2}\left(\overline{X} + \overline{Y}\right) \qquad \widehat{\mu}_2 = \frac{2}{3}\overline{X} + \frac{1}{3}\overline{Y}$$

b) Visa att både $\widehat{\mu}_1$ och $\widehat{\mu}_2$ är väntevärdesriktiga. Räkna ut variansen för båda estimatorerna. Vilken av estimatorerna är den bästa? Motivera.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 26/10-2016

Sal: Värtasalen

Tenta: Statistikens grunder 2

Kurs: Statistikens grunder

ANONYMKOD: SGD-0059

☑ 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					
Lär.ant.									11
20	20	18,5	20	20					

POÄNG	BETYG	Lärarens sign.
98,5	A	JF

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: SGD-0059 Blad nr: 1

Uppgift 1

Förklara kortfattat begrepp:

a) Samplingfördelning beskriver hur urvalsmedelvärdet kommer att avvika i populationen, dvs om vi tar ett urval från en population och beräknar dess medelvärde ($\bar{x}$) och stand. avvikels (S_x) och sedan beräknar urvalsmedelvärdets avvikelse $\sigma_{\bar{x}}$ som $\sigma_{\bar{x}} = \frac{S_x}{n}$, där n är urvalsstorlek, kommer dessa parametrar ($\bar{x}$, $\frac{S_x}{n}$) beskriva samplingfördelning för $\bar{x}$ i populationen (om vi tar n styck lika stora urval från samma population visar de parametrarnas spridningen för populationensmedelvärdet).

4

b) P-värde: kan tolkas som:

- sannolikheten att få en lika stor extremvärde eller större som vi fick i våra observationer,
- den anger den lägsta sign. nivå på vilket vår nollhypotes kan

förkastas. Om p-värdet är mindre än sign. nivå, då säger vi att våra observationerna är stat. signifikanta på sign. nivå större än p-värdet. Om p-värdet är större än sign. nivå, då säger vi att våra observationerna inte är stat. signifikanta och att H_0 inte kan förkastas på sign. nivå < p-värdet. 4

c) 95%-igt k.i anger intervall som kan innehålla det sanna värdet på populationsmedelvärdet med 95% säkerhet. Dvs om vi drar n stycken lika stora urval från samma population kommer i 95 fall av 100 det sanna μ ligga i konfidensintervallen för dessa urval. 4

d) Korrelation visar hur starkt eller svagt samband (linjärt) finns mellan variablerna. Korrelationen kan vara positiv (om en variabel ökar i värde, så ökar den andra och vice versa), eller negativ (om en ökar, minskar den andra och vice versa). Korrelationen kan också vara 0, dvs vi kan inte påvisa en linjärt samband mellan variablerna.

SU, DEPARTMENT OF STATISTICS

Room: Vårbussalen Anonymous code: SGD-0059 Sheet number: -2-

Kausalitet: visar ett orsakssamband mellan variablerna, dvs hur en variabel påverkar den andra (alltid i en riktning).

Om variablerna är korrelerade behöver det inte vara att det föreligger ett kausalt samband mellan dem. Det kan vara så att en tredje variabel som vi inte "ser" påverkar de andra och därför ger intryck att det kan finnas orsaksamband mellan två andra variabler, som är korrelerade.

Som regel måste man komma ihåg att en stark korrelation mellan variablerna är inte bevis för orsaks-samband mellan dem. Också använder man sunt förnuft när man tolkar korrelationen och samband mellan variablerna.

4

c) Enkel index: $I_t^b = \frac{x_t}{x_b} \cdot 100$ – (basperiod)

visar hur priset (etc) för en enhet (vara 4) i tidpunkt t har ändrats i jämförelse med bastidpunkten.

Sammansatt index t.ex. KPI visar hur priser (t.ex. för en korg av varor) har ändrats i jämförelse med någon given tidpunkt (för KPI tidpunkten är ofta år 1980).

4

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: SGD-0059 Blad nr: 3

Uppgift 2

Handlingsalt (HA)	Tillstånd: Långsamt	Normalt	Snabbt
HA_L - Lägenheter	-750	100	1500
HA_K - Kontor	-2000	500	1000
HA_{KC} - Köpcentrum	150	300	450

a) Maximinkriteriet ("pessimistiskt kriteriet"): vi försöker gardera oss mot dåliga utfall. Vi väljer den minsta nyttan för varje handlingsalternativ och sedan väljer det HA bland dem som ger den största nyttan.

$$\bar{a}_{ij} = \{ \bar{a}_i = -750,\ \bar{a}_K = -2000,\ \bar{a}_{KC} = 150 \}$$

Svar: en fastighetsinvesterare bör bygga ett köpcentrum enligt maximinkriteriet.

5

b) Laplacekriteriet: vi antar att varje tillstånd, alltså varje scenario i vårt fall, har samma sannolikhet att hända (p(långsamt) = p(normalt) = p(snabbt) = $\frac{1}{3}$) och sedan räknar det förväntade värde för varje

→

handlingsalternativ. Vi väljer sedan det förväntade värde som har den största nyttan (som för beslut under risk)

	Långs ($\frac{1}{3}$)	Norm. ($\frac{1}{3}$)	Snabbt ($\frac{1}{3}$)	E(U)
HA_L	−750	100	1500	283,33
HA_K	−2000	500	1000	−166,67
HA_{KC}	150	300	750	400

$E(U_L) = -750 \cdot \frac{1}{3} + 100 \cdot \frac{1}{3} + 1500 \cdot \frac{1}{3} =$
$\approx 283,33$

$E(U_K) = -2000 \cdot \frac{1}{3} + 500 \cdot \frac{1}{3} + 1000 \cdot \frac{1}{3} =$
$\approx -166,67$

$E(U_{KC}) = 150 \cdot \frac{1}{3} + 300 \cdot \frac{1}{3} + 750 \cdot \frac{1}{3} \approx 400$

Svar: en fastighetsinvesterare bör bygga ett köpcentrum enligt Laplacekriteriet.

5

c) Maximaxkriteriet ("optimistisk kriteriet").
Vi väljer den största nyttan för varje handlingsalternativ och sedan väljer den största bland dem.

$a^+_{ij} = \max\{a^+_L = 1500, a^+_K = 1000, a^+_{KC} = 750\}$

Svar: en fastighetsinvesterare bör bygga lägenheter enligt maximaxkriteriet.

5

SU, STATISTIK

Skrivsal: VA Anonymkod: Blad nr: 4

d) Minimaxkriteriet: vi försöker minimera "ånger" med minimaxregret-kriteriet. Vår max alternativförlusten bör vara så liten som möjligt.
Alt. förlusten räknas enligt följande

$r_{ij} = S_j - u_{ij}$, där

S_j - den maximala nyttan för varje tillstånd. I vårt fall:

$S_L(max) = 150$, $S_N(max) = 500$, $S_S(max) = 1500$

Regretmatrix

	S_L	S_N	S_S
MA_L	150 + 750 = 900	500 - 100 = 400	1500 - 1500 = 0
MA_K	150 + 2000 = 2150	500 - 500 = 0	1500 - 1000 = 500
MA_{KC}	150 - 150 = 0	500 - 300 = 200	1500 - 750 = 750

$r_{ij} = \{ r_L = 900, r_K = 2150, r_{KL} = 750 \}$

Svar: fastighetsinvesterare bör bygga köpcentrum enligt minimaxregret-kriteriet.

5

Uppgift 3

Urval: $n = 550$ (pers, stockholmare), 18-49 år,
182 av dessa är tatuerade.

a) Proportionen tatuerade stockholmare i åldern 18-49 år:

$$p = \frac{182}{550} \cong 0{,}331 \ (\approx 33{,}1\ \%)$$

2

Svar: proportionen tatuerade stockholmare i åldern 18-49 år är cirka 33,1 % ($p \cong 0{,}331$).

b) Variansen hos urvalsproportionen.

Ändlighetskorrektionen:

Vi gör urval ($n = 550$) utan återläggning för att skatta proportionen tatuerade stockholmare från populationen "stockholmare" som omfattar personer i åldern 18-49 år. Åldersspridningen i populationen är tillräckligt stor och omfattar såväl unga som medelåriga personer i arbetsför åldern, dvs populationen kan sägas vara oändlig, dvs ändlighetskorrektionen behöver inte användas.

Variansen kan räknas enligt följande formel: $V(p) = \frac{p(1-p)}{n}$ såsom för urval med återläggning (oberoende föreligger).

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: SGD-0059 Blad nr: 5

Vidare: $n = 550$ (stort, >30), $np = 550 \cdot 0{,}331 =$
$= 182{,}05$ samt $n(1-p) = 550 \cdot 0{,}669 =$
$= 364{,}95$ (≥ 5), dvs fördelning för proportionen kan approximeras med normalfördelning enligt CGS.

$$\Rightarrow V(p) = \frac{p(1-p)}{n} = \frac{0{,}331 \cdot 0{,}669}{550} =$$

$$= \frac{0{,}221439}{550} \approx 0{,}000403$$

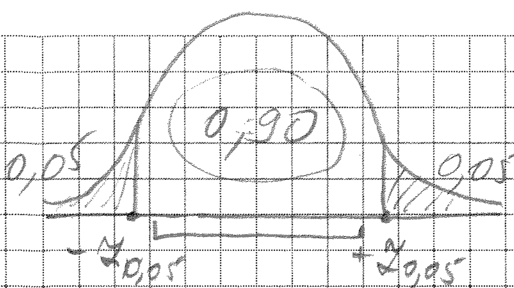
Svar: variansen hos urvalsproportionen är lik med c. 0,000403. Ändlighets-korrektionen bör inte användas eftersom vi gör urval från en oändlig population.

(3)

Nej, antal Stockholmare är ändligt

c) 90%-igt K.i: fördelningen för proportionen tuberade stockholmare kan approximeras med normal-fördelning enligt CGS (mer utförlig motivering i punkt (b)). 90%-igt K.i kan därför beräknas enligt följande:

$$p \pm z_{0,05} \cdot \sqrt{\frac{p(1-p)}{n}} \quad (Z \overset{app}{\sim} N(0,1))$$



Ur tabellen för n.f.:

$z_{0,05} = 1,6449$

$\sqrt{\frac{p(1-p)}{n}} = \sqrt{0,000403} \approx 0,02007$

Felmarginalen $B = 1,6449 \cdot 0,02007 \approx 0,033$

$p \pm B$

$0,331 \pm 0,033$

$[0,298 \,,\, 0,364]$

8

Svar: 90%igt konfidensinterval för vårt urval är [0,298; 0,364], vilket betyder att i 90 fall av 100 kommer det sanna värdet för proportionen ligga i intervallet [0,298; 0,364] (eller [29,8%; 36,4%]).

d) $n = ?, \quad 2B = 5\%$.

$2B = 0,05 \Rightarrow B = 0,025$

Formeln:

$$1,6449 \cdot \sqrt{\frac{p(1-p)}{n}} \leq B$$

$$1,6449^2 \cdot \frac{p(1-p)}{n} \leq B^2$$

$$n \cdot B^2 \geq 1,6449^2 \cdot p \cdot (1-p)$$

$$n \geq \frac{1,6449^2 \cdot p(1-p)}{B^2}$$

Alternativ 1: $p = 0,331 \Rightarrow$

SU, STATISTIK

Skrivsal: ____________ **Anonymkod:** ____________ **Blad nr:** 6

$$n \geq \frac{1{,}6449^2 \cdot 0{,}331 \cdot 0{,}669}{(0{,}025)^2} \geq$$

$$\geq \frac{0{,}5991}{0{,}0004625} \geq \underline{1295{,}351}, \text{ dvs}$$

$n \geq \underline{1296}$

<u>Eller, alternativ 2</u>: där $p = 0{,}5$ (ger den största variansen).

$$n \geq \frac{1{,}6449^2 \cdot 0{,}5 \cdot 0{,}5}{(0{,}025)^2} \geq \frac{0{,}676424}{0{,}0004625} \geq$$

$$\geq \underline{1462{,}54}, \text{ dvs}$$

$n \geq 1463$.

<u>Svar</u>: Man bör dra minst a) $n = 1296$ (urvalsvariansen skattas med $p = 0{,}331$), eller b) $n = 1463$ (urvalsvariansen anges som maximal (0,25) med $p = 0{,}5$).

5,5

Uppgift 4

Vår info

Undervikt: BMI < 18,5

Normalvikt: 18,5 < BMI < 24,9

Övervikt: 25 < BMI < 29,9

Fetma: BMI > 30

År 2000, fördelning

Undervikt	Norm.	Övervikt	Fetma
7%	49,5%	34%	9,5%

a) år 2014, undersökning bland 600 unga svenskar ($n = 600$).

År 2014, fördelning:

Undervikt	Norm.	Övervikt	Fetma
2%	48,5%	35,5%	14%

Testa med lämpligt hypotestest om fördelningen över viktkategorierna 2014 är densamma som den var 2000 med $\alpha = 0,01$

Lösning: Eftersom data som vi disponerar är på ordinalskala så använder vi i det här fallet χ^2-måttet för att lösa uppgiften och antar att observationerna på vikt är st oberoende.

Ett lämpligt test i detta fall är "goodness of fit" test som testar om en viss stokastisk variabel

SU, STATISTIK

Skrivsal: Väntesalen Anonymkod: SGD-0059 Blad nr: 7

följer en viss fördelning, i vårt fall om fördelningen över viktkategorierna år 2014 är densamma som den var år 2000.

Hypoteser: H_0: fördelningen viktkategorier 2014 är densamma som den var år 2000;

H_a: fördelningen – " – skiljer sig från fördelningen år 2000.

Om n = 600 kommer vår fördelning år 2014 och den förväntade fördelningen över viktkategorierna given att H_0 stämmer se ut som i tabell nedan:

	n_i(2014)	$E(n_i)$ om H_0 är sann (7,5)
Undervikt	$600 \cdot 0{,}02 = 12$	$600 \cdot 0{,}07 = 42$
Normalvikt	$600 \cdot 0{,}485 = 291$	$600 \cdot 0{,}495 = 297$
Övervikt	$600 \cdot 0{,}355 = 213$	$600 \cdot 0{,}34 = 204$
Fetma	$600 \cdot 0{,}14 = 84$	$600 \cdot 0{,}095 = 57$
Σ	600	600

Kritiskt värde: SN = 1%, $\alpha = 0,01$

$$\chi^2_{0,01}(n-1) = \chi^2_{0,01}(4-1) =$$
$$= \chi^2_{0,01}(3) = 11,345$$

Beslutsregel: vi förkastar H_0 om $\chi^2_{obs} > \chi^2_{0,01}(3) = 11,345$ på SN 1%.

Testvariabel:

$$\chi^2 = \sum \frac{(n_i - E(n_i))^2}{E(n_i)} \sim \chi^2(n-1)$$

$$\chi^2_{obs} = \frac{(12-42)^2}{42} + \frac{(291-297)^2}{297} + \frac{(213-204)^2}{204} +$$
$$+ \frac{(84-57)^2}{57} = 21,43 + 0,12 + 0,397 +$$
$$+ 12,79 = 34,737$$

Svar: $\chi^2_{obs} = 34,737 > \chi^2_{0,01} = 11,345$, dvs vi kan förkasta nollhypotesen på sign.nivå 1%. Det finns tillräckligt med stöd att fördelningen över viktkategorierna 2014 inte är densamma som den var år 2000. Testet som används är "goodness of fit test"

10

SU, STATISTIK

Skrivsal: Vårbergsalen Anonymkod: SBD-0059 Blad nr: 8

Fort. uppgift 4.

b) År 2014, $n = 600$, av dem 290 kvinnor och 310 män.

Är fördelningen mellan viktklasser är samma för kvinnor och män?

Lösning: vi har data på nominal/ordinal-skala, därför använder vi χ^2-måttet för att genomföra testet. I detta fall använder vi ett av "homogenitets-testet som tester om två eller flera variabler kommer från samma fördelningen.

$E(n_i) \geq 5$

zon	Kvinnor	Män	Σ	Prop.	Förväntade frekv.	Kvinnor	Män	Σ
Undervikt	9	3	12	$\frac{12}{600}$		$\frac{12}{600} \cdot 290 = 5,8$	$\frac{12}{600} \cdot 310 = 6,2$	1
Normalv.	157	133	290	$\frac{290}{600}$		$\frac{290}{600} \cdot 290 = 140,17$	$\frac{290}{600} \cdot 310 = 149,83$	29
Övervikt	84	130	214	$\frac{214}{600}$		$\frac{214}{600} \cdot 290 = 103,43$	$\frac{214}{600} \cdot 310 = 110,57$	21
Fetma	40	44	84	$\frac{84}{600}$		$\frac{84}{600} \cdot 290 = 40,6$	$\frac{84}{600} \cdot 310 = 43,4$	84
Σ	290	310	600	1	Σ	290	310	60

Hypoteser: H_0: Fördelningen mellan viktklasserna är samma för kvinnor och män.

H_A: — " — " — " — är inte samma (skiljer sig).

Om det är så fördelningen "viktklasser" är samma för kvinnor och män, då borde proportionen "undersöktgå" personer bli samma för kvinnor och män, i vårt fall $\frac{12}{600} = 0{,}02$ (eller 2%).

Det betyder 5,8 kvinnor ($0{,}02 \cdot 290 = 5{,}8$) och 6,2 män ($0{,}02 \cdot 310 = 6{,}2$). På samma sätt beräknar vi de andra förväntade frekvenser (se tabell i början av uppgiften).

Testvariabel: $\chi^2 = \sum \frac{(n_i - E(n_i))^2}{E(n_i)} \sim \chi^2(\nu)$,

där ν = (antal kategorier − 1) × (antal populationer − 1).

Kritiskt värde: $\alpha = 0{,}01$

$$\chi^2_{0{,}01}(4-1)\cdot(2-1) = \chi^2_{0{,}01}(3) = 11{,}345$$

Beslutsregeln: förkasta H_0 om $\chi^2_{obs} > \chi^2_{0{,}01}(3) = 11{,}345$.

$$\chi^2_{obs} = \frac{(9-5{,}8)^2}{5{,}8} + \frac{(3-6{,}2)^2}{6{,}2} + \frac{(154-140{,}17)^2}{140{,}17} +$$

$$+ \frac{(133-149{,}83)^2}{149{,}83} + \frac{(84-103{,}43)^2}{103{,}43} + \frac{(130-110{,}57)^2}{110{,}57} +$$

$$+ \frac{(40-40{,}6)^2}{40{,}6} + \frac{(44-43{,}4)^2}{43{,}4} = 1{,}77 + 1{,}65 + 2{,}021 +$$

$$+ 1{,}89 + 3{,}65 + 3{,}41 + 0{,}009 + 0{,}008 \approx 14{,}408$$

→

SU, STATISTIK

SGD-0059

Skrivsal: Värtasalen Anonymkod: Blad nr: 9

Fort Upp 4, b

Svar: $\chi^{2}_{obs} = 14{,}408 > \chi^{2}_{0,01}(3) = 11{,}345$; dvs
Ho förkastas på sign nivå 1%.
Det betyder att vi har bevis för
att fördelningen mellan viktklasser
inte är samma för kvinnor och
män

10

Uppgift 5

a) Data: $n = 40$ mätningar. X - st. variabel (μ, σ^2).
X är oberoende.

$\bar{x} = 1{,}033\,g$

$S_x^2 = 0{,}01188$

Rekommenderade nivå: 1,0 g.

Vi vet att X är en stok variabel och
att alla X är oberoende. Vi vet ej vilken
fördelning X följer, och σ^2 är okänd.
Men eftersom vårt urval $(n = 40)$ är
tillräckligt stort (>30), kan vi
approximera fördelningen till normal-
fördelning och använda Z-variabeln
som testfstatistika

Hypoteser:

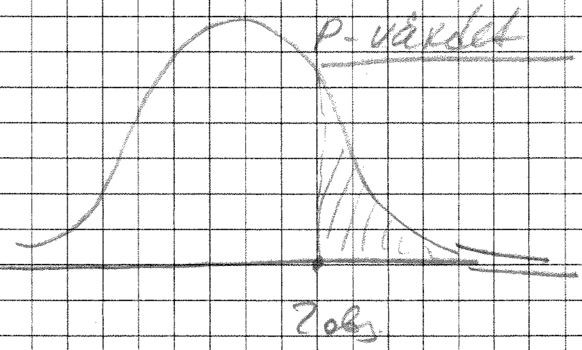
H_0: $\mu = 1,0$

H_A: $\mu > 1,0$ (ensidig test)

(Vi sätter $\mu > 1,0$ för H_A för att kolla om maskinen är inställd på högre nivå än rekommenderade 1,08).

Testvariabel: $Z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \overset{appr}{\sim} N(0,1)$ om H_0 är sann.

$$Z_{obs} = \frac{1,033 - 1,0}{\sqrt{0,01188/40}} = \frac{0,033}{0,0172337} \approx 1,91$$ R



$$p\text{-värdet} = P(Z > 1,91) = 1 - P(Z < 1,91) = 1 - 0,97193 = 0,028$$ R

Svar: p-värdet = 0,028 eller 2,8 %, vilket betyder att vi kan förkasta H_0 på alla sign. nivå som är större än p-värdet. T.ex. H_0 kan förkastas på sign. nivå 3 %, 5 %.. (dvs vår resultat i ett sådant fall kan sägas vara stat. signifikant). Men på alla nivå som är mindre än 2,8 % kan vi inte förkasta H_0. T.ex. om vi väljer

SU, STATISTIK

Skrivsal: Värtasalen   Anonymkod: SGD-0059   Blad nr: 10

utföra testet på sign. nivå 1%, då kan vi inte förkasta H_0 (vår resultat kan inte sägas vara stat. signifikant) det är för läsaren att utgöra vilken sign. nivå hen behöver för att tolka testet. I övrigt kan man säga att c. 28 av 1000 mätningar kommer att visa en högre nivå än den rekommenderade.

10

b) $n = 40$ mätningar på Y $(y_1, y_2, \ldots, y_{40})$

$E(Y) = E(X)$

$V(Y) = 2{,}25 \cdot \sigma^2$

$\hat{\mu}_1 = \frac{1}{2}(\bar{X} + \bar{Y})$ $\hat{\mu}_2 = \frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}$

(i) Visa att $\hat{\mu}_1$ och $\hat{\mu}_2$ är väntevärdesriktiga

(1) $\hat{\mu}_1$

$$E(\hat{\mu}_1) = E\left(\frac{1}{2}(\bar{X} + \bar{Y})\right) = \frac{1}{2}E(\bar{X} + \bar{Y}) =$$

$$= \frac{1}{2}\left(E(\bar{X}) + E(\bar{Y})\right) = \frac{1}{2}\left(E\left(\frac{\sum^n x_i}{n}\right) + E\left(\frac{\sum^n y_i}{n}\right)\right) =$$

$$= \frac{1}{2}\left(\frac{1}{n} \cdot E(n \cdot x_i) + \frac{1}{n} \cdot E(n \cdot y_i)\right) =$$

$$= \frac{1}{2}\left(\frac{n}{n} \cdot E(X_i) + \frac{n}{n} E(y_i)\right) = \frac{1}{2}\left(E(X) + E(X)\right) =$$

$$= \frac{1}{2} \cdot 2\mu = \mu.$$

I vårt fall $E(\bar{X}) = 1{,}033$, $E(\bar{X}) = E(X) =$
$= \mu = 1{,}033$, på samma sätt
$E(\bar{Y}) = E(\bar{X}) = 1{,}033$.
$\Rightarrow \quad E(\hat{\mu}_1) = \frac{1}{2}(1{,}033 + 1{,}033) = 1{,}033$

(2) $\hat{\mu}_2 = \frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}$

$E(\hat{\mu}_2) = E\left(\frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}\right) = \frac{2}{3}E(\bar{X}) + \frac{1}{3}E(\bar{Y}) =$

$= \frac{2}{3} \cdot \frac{E(nX_i)}{n} + \frac{1}{3}\frac{E(nY_i)}{n} = \frac{2}{3} \cdot \frac{n}{n} \cdot E(X) +$

$+ \frac{1}{3} \cdot \frac{n}{n} \cdot E(Y) = \frac{2}{3}E(X) + \frac{1}{3}\underbrace{E(Y)}_{E(X)} =$

$= \frac{2}{3}E(X) + \frac{1}{3}E(\cancel{Y}X) = E(X) = \mu$

I vårt fall $E(\bar{X}) = 1{,}033$, $E(\bar{X}) = E(X) =$
$= 1{,}033$, $E(\bar{Y}) = 1{,}033$.
$\Rightarrow E(\hat{\mu}_2) = \frac{2}{3} \cdot E(\bar{X}) + \frac{1}{3} \cdot E(\bar{Y}) =$

$= \frac{1}{3} \cdot 1{,}033 + \frac{2}{3} \cdot 1{,}033 = \underline{1{,}033}$

Svar: Vi har visat att både $\hat{\mu}_1$ och $\hat{\mu}_2$ är VVR genom att visa att $E(\hat{\mu}_1) = E(\hat{\mu}_2) = \mu$ och i vårt fall $E(\hat{\mu}_1) = E(\hat{\mu}_2) = 1{,}033$.

$\rightarrow$

SU, STATISTIK

Skrivsal: Västrasalen SGD-0059 Anonymkod: ______ Blad nr: 11

(ii) Räkna ut varianserna för $\hat{\mu}_1$ och $\hat{\mu}_2$.

oberoende

$$\underline{V(\hat{\mu}_1)} = V\left(\frac{1}{2}(\bar{x}+\bar{y})\right) = \frac{1}{4}V(\bar{x}+\bar{y}) =$$

$$= \frac{1}{4}\left(V(\bar{x})+V(\bar{y})\right) = \frac{1}{4}\left(\frac{V(x)}{n}+\frac{V(y)}{n}\right) = \quad (n=40).$$

$$= \frac{1}{4}\left(\frac{\sigma^2}{40}+\frac{2{,}25\cdot\sigma^2}{40}\right) = \frac{1}{4}\cdot\frac{3{,}25\cdot\sigma^2}{40} =$$

$$= \frac{3{,}25\,\sigma^2}{160} \approx 0{,}0203\cdot\sigma^2$$

oberoende

$$V(\hat{\mu}_2) = V\left(\frac{2}{3}\cdot\bar{x}+\frac{1}{3}\bar{y}\right) = \frac{4}{9}V(\bar{x})+\frac{1}{9}V(\bar{y}) =$$

$$= \frac{4}{9}\cdot\frac{\sigma^2}{n}+\frac{1}{9}\cdot\frac{2{,}25\cdot\sigma^2}{n} = \frac{4\sigma^2}{9\cdot 40}+\frac{2{,}25\cdot\sigma^2}{9\cdot 40} =$$

$$= \frac{6{,}25\cdot\sigma^2}{360} = 0{,}0173\cdot\sigma^2$$

Om vi skattar σ^2 med S_x^2 (både $\bar{x}$ och $\bar{x}$ kan approx. med n.f då n är stor), får vi: $V(\hat{\mu}_1) = 0{,}0203\cdot\sigma^2 = 0{,}0203\cdot S_x^2 =$

$= 0{,}0203\cdot 0{,}01188 \approx \underline{0{,}000241}$

$V(\hat{\mu}_2) = 0{,}0173\cdot\sigma^2 = 0{,}0173\cdot S_x^2 =$

$= 0{,}0173\cdot 0{,}01188 = \underline{0{,}000206}$

Svar: Varianserna för både estimatorerna är: $V(\hat{\mu}_1) \approx 0{,}0203\cdot\sigma^2$ (eller i vårt fall $V(\hat{\mu}_1) = 0{,}000241$) och $V(\hat{\mu}_2) = 0{,}0173\cdot\sigma^2$;

eller i vårt fall $V(\hat{\mu}_2) = 0{,}000206$):

(iii) Vilken av estimatorerna är bästa?

<u>Svar</u>: den bästa estimatorn är den som har minst varians, dvs spridningen kring kring väntevärdet är mindre. För att välja den bästa estimatorn måste vi i så fall välja den minsta variansen bland:

$V(\hat{\mu}_1) = 0{,}0203 \cdot \sigma^2$

$V(\hat{\mu}_2) = 0{,}0173 \cdot \sigma^2$

$V(\hat{\mu}_2) < V(\hat{\mu}_1)$ eftersom $0{,}0173 < 0{,}0203$, vad också uträknas om vi ersätter σ^2 med $S_x^2 = 0{,}0188$:

$V(\hat{\mu}_1) = 0{,}0203 \cdot 0{,}0188 = 0{,}000381$

$V(\hat{\mu}_2) = 0{,}0173 \cdot 0{,}0188 = 0{,}000325$.

Vi kan se att $V(\hat{\mu}_2)$ är något mindre än $V(\hat{\mu}_1)$.

<u>Svar</u>: den bästa estimatorn är $\hat{\mu}_2 = \frac{2}{3}\bar{x} + \frac{1}{3}\bar{y}$, då den har minsta varians.

10

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 26/10-2016

Sal: Värtasalen

Tenta: Statistikens grunder 2

Kurs: Statistikens grunder

ANONYMKOD: SGD-0055

☒ 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. 19	20	20	20	17					

HS

POÄNG	BETYG	Lärarens sign.
96	A	JF

SU, STATISTIK

Skrivsal: Värtasalen **Anonymkod:** SGD-0055 **Blad nr:** 1

urvalskarakteristikan

1 a) Samplingsfördelning är fördelningen av observationerna, en urvalsundersökning kring medelvärdet $\bar{X}$, dvs värdet på de olika observationerna och hur många observationer som gjorts på de olika värdena.

b) P-värdet är sannolikheten att få ett värde i en urvalsundersökning som är lika sannolikt eller mindre sannolikt än det erhållna värdet förutsatt att populationsmedelvärdet är sant. Om man gör en hypotesprövning och beräknar exempelvis Z_{obs} är p-värdet sannolikheten att få ett värde på Z_{obs} som är lika eller mindre sannolikt än det erhållna värdet givet att H_0 är sann. Storleken på p-värdet kan alltså användas för att testa hur troligt det är att en hypotes kan förkastas. Ett mycket litet p-värde gör att H_0 kan förkastas. Hur litet det ska vara beror på vilken signifikansnivå man väljer. Ju mindre p-värde desto större signifikansnivå. Om p-värdet är stort har vi fått ett värde som är sannolikt och vi kan inte förkasta H_0.

4

forts a)

Samplingsfördelningen är motsvarigheten till den teoretiska sannolikhetsfördelningen för hur utfall kommer att fördelas i teorin.

3

c)

Ett 95%-igt konfidensintervall ger värden som om man gör en stor mängd urvalsundersökningar så kommer urvalsmedelvärdet att hamna inom gränserna för konfidensintervallet i 95% av fallen. I 5% av fallen kommer värdet att hamna utanför intervallet för ett visst populationsmedelvärde, (eller proportion).

4

d)

linjärt bra!

Korrelation betyder att det finns ett samband mellan två variabler. T.ex om den ena ökar så ökar den andra också, eller om den ena ökar så minskar den andra. Det behöver dock inte betyda att orsaken att den ena ökar är den andra variabeln. Det kan lika gärna vara en tredje variabel som gör att de båda variablerna ökar samtidigt tex.

Ex. i samhällen där det finns många storkar föds det många barn, men storkarna är inte orsaken till att det föds fler barn, det är att det finns fler storkar och det föds fler barn i små orter.

Kausalitet är att det finns ett orsakssamband mellan två variabler. Den ena variabelns förändring orsakas av den andra. Det är inte alltid lätt att avgöra om det finns kausalitet mellan två varia

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: S6D-0055 Blad nr: 2

1d forts.

Man måste visa att förändringen inte skett om det inte hade varit för variabelns påverkan, och att det finns en tidsmässig ordning, påverkan skedde före verkan.

4

e) Ett enkelt index visar förändringen i tid för en variabel, t ex pris över tid, i förhållande till ett basvärde. Det visar alltså procentuell förändring över tid från en bastidpunkt.

Ett sammansatt index väger samman flera olika variabler, t ex priser på flera olika varor. Ex. KPI som väger samman en mycket stor mängd konsumentprodukter, varor och tjänster, för att visa hur den allmänna prissättningen har förändrats över tid. I sammansatta index brukar man använda viktade värden där vikterna kan vara produktionen eller försäljning, så att varor som är vanligare får större vikt i indexet. Exempel på sammansatta index är Laspeyres index och Paasches index.

4

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: SGD-0055 Blad nr: 3

(2) En fastighetsinvesterare överväger 3 olika investerings-alternativ: lägenheter, kontor eller köpcentrum.

Handlingsalt.	Naturtillstånd S_1 Långsamt	S_2 Normalt	S_3 Snabbt	Maximin $\max(a_1^-, a_2^-, a_3^-)$	Maximax $\max(a_1^+, a_2^+, a_3^+)$
A_1 Lägenheter	−750	100	1500	−750	1500
A_2 Kontor	−2000	500	1000	−2000	1000
A_3 Köpcentrum	150	300	750	150	750

a) Maximinkriteriet: Investeraren väljer det största värdet av de minsta avkastningsalternativen, dvs alternativ 3. – Köpcentret som ger den lägsta avkastningen 150 tkr.

Svar Alternativet Köpcentrum välj.

5

b) Laplacekriteriet: Man antar att alla alternativ är lika sannolika, dvs $\frac{1}{3}$ sannolikhet var beträffande naturtillstånden långsamt, normalt och snabbt växande population. Sedan välj det största väntevärdet. Laplace: $\max(E(u_1), E(u_2), E(u_3))$

A_1 Lägenhet	$\frac{1}{3}(-750) + \frac{1}{3}\cdot 100 + \frac{1}{3}\,1500 = 283{,}3$
A_2 Kontor	$\frac{1}{3}(-2000) + \frac{1}{3}(500) + \frac{1}{3}\,1000 = -166{,}67$
A_3 Köpcentrum	$\frac{1}{3}\cdot 150 + \frac{1}{3}\,300 + \frac{1}{3}\,750 = 400$

Vänd →

b)

Svar: Enligt Laplacekriteriet välis alternativ 3, köpcentrum, där avkastningen förväntas bli 400 tkr.

5

c)

Enligt maximaxkriteriet jämförs de största avkastningarna för varje naturtillstånd och så välis det alternativ som har den största avkastningen.

Svar: Alterativ 1, Lägenheter välis enligt maximaxkriteriet med en avkastning på 1500 tkr.

5

d) minimax regret: alternativet som ger den minsta "ånger" välis. Skillnaden mellan det största värdet och värdet för varje naturtillstånd beräknas och det högsta värdet för varje handlingsalternativ jämförs. Det minsta värdet av dessa välis.

$S_1^+ = 150$ $S_2^+ = 500$ $S_3^+ = 1500$

	Lågsanbt $S_1^+ = 150$	Normalt $S_2^+ = 500$	Snabbt $S_3^+ = 1500$	
A_1 Lägenheter	150 − (−750) = 900	500 − 100 = 400	1500 − 1500 = 0	900
A_2 Kontor	150 − (−2000) = 2150	500 − 500 = 0	1500 − 1000 = 500	215
A_3 Köpcentrum	150 − 150 = 0	500 − 300 = 200	1500 − 750 = 750	75

Svar d: Enligt minimax regret välis alternativ 3, köpcentrum, som har det minsta värdet för den största "ångern"

5

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: SGD-0055 Blad nr: 4

③ Uppskatta andelen tatuerade Stockholmare, 18-49 år

Urvalsundersökning gjordes med $n = 550$. Av dessa var 182 tatuerade. Anta att observationerna är oberoende och likafördelade

a) proportionen som är tatuerad uppskattas med

$$p = \frac{\sum_{i=1}^{n} X_i}{n}$$ där $X_i = 1$ om personen är tatuerad och $X_i = 0$ om personen inte är tatuerad.

X är stokastiskt oberoende variabler.

$$p = \frac{182}{550} = 0{,}3309$$

2

Svar: Proportionen tatuerade Stockholmare är 33,09%

b) Variansen hos urvalsproportionen är:

$$S_p^2 = \frac{p(1-p)}{n} = \frac{0{,}3309(1-0{,}3309)}{550} = 0{,}000402$$

Ändlighetskorrektion behöver inte användas eftersom N är så stort. N är populationen, dvs alla Stockholmare mellan 18 och 49 år, flera hundratusen.

Ändlighetskorrektionen $\frac{N-n}{N-1}$ går mot 1 när N är stort. Som regel brukar man inte behöva använda ändlighetskorrektion när den är $> 0{,}95$.

Här:

$$\frac{N-550}{N-1} = 0{,}95 \Leftrightarrow N - 550 = 0{,}95(N-1) \Leftrightarrow$$

$$N - 550 = 0{,}95N - 0{,}95 \Rightarrow N = 10981$$

$$0{,}05N = 549$$

Så för N över ca 11000 behövs inte ändlighetskorrektion och i det här fallet är N betydligt större.

Vänd →

Svar: Variansen för urvalsproportionen är 0,0004 och ändlighetskorrektion behövs inte eftersom N är så stort.

4

c)

Konfidensintervall för proportioner

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Eftersom n är så stort och $p \cdot n = 0{,}33 \cdot 550 = 181{,}5$ och $(1-p)n = 368{,}5$ båda är större än 5 kan vi approximera fördelningen med normalfördelningen enligt centrala gränsvärdessatsen.

90%-igt konfidensintervall $\Rightarrow \alpha = 0{,}1$

$\alpha/2 = 0{,}05$

$z_{\alpha/2} = z_{0{,}05} = 1{,}6449$ ur tabell 2

$p = 0{,}3309$

Konfidensintervallet blir:

$$0{,}3309 \pm \sqrt{\frac{0{,}3309(1-0{,}3309)}{550}} \cdot 1{,}6449$$

$$0{,}3309 \pm 1{,}6449 \cdot 0{,}02006 \Rightarrow$$

$$0{,}3309 \pm 0{,}033$$

Svar: Ett 90%-igt konfidensintervall ges av $0{,}3309 \pm 0{,}033$ eller $[0{,}2979; 0{,}3639]$

8

SU, STATISTIK

Skrivsal: Värtasalen **Anonymkod:** SGD-0055 **Blad nr:** 5

3 forts

d) Hur stort urval? Hitta n

Längden på hela konfidensintervallet ska vara max 5 procentenheter dvs 0,05

Halva konfidensintervallet = 0,025

Halva konfidensintervallet = felmarginalen = B

ges av

$$B = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow$$

$$0{,}025 = 1{,}6449\sqrt{\frac{0{,}3309(1-0{,}3309)}{n}}$$

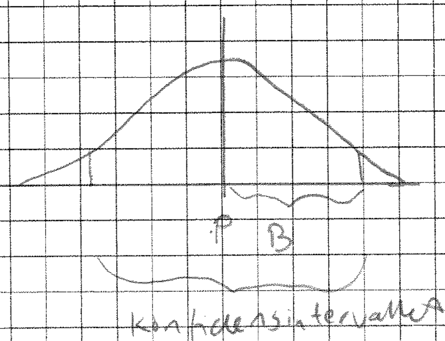
$$\left(\frac{0{,}025}{1{,}6449}\right)^2 = \frac{0{,}3309(1-0{,}3309)}{n}$$

$$2{,}3099 \cdot 10^{-4} = \frac{0{,}2214}{n}$$

$n = 958{,}49 \Rightarrow n = 959$ man avrundar alltid uppåt för att säkerställa att konfidensintervallet ska bli max 5%-enheter

Svar $n = 959$ st

6



SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: SGD-0055 Blad nr: 6

(4) Andelen vuxna svenskar i BMI-kategorier:

	2000 (hela populationen)	2014 (urval)
Undervikt	0,07	0,02
Normalvikt	0,495	0,485
Övervikt	0,34	0,355
Fetma	0,095	0,14

$\alpha = 0{,}01$

$n = 600$

Testa om fördelningen över viktkategorierna är samma.
Vi använder ett goodness-of-fittest med χ^2 för att avgöra om fördelningen i urvalsundersökningen överensstämmer med populationsfördelningen.

Antal	2000	2014
Undervikt	42	12
Normalvikt	297	291
Övervikt	204	213
Fetma	57	84

I den första kolumnen är de förväntade värdena om den teoretiska fördelningen stämmer, dvs $0{,}07 \cdot 600 = 42$ personer bör vara underviktiga i urvalet osv. Alla värden är > 5 så vi kan gå vidare utan att slå ihop klasser.

Vi sätter upp en hypotes:

H_0: fördelningen i urvalsundersökningen 2014 är samma som populationsfördelningen 2000. VC

H_A: Fördelningen i urvalsundersökningen 2014 är <u>inte</u> samma som 2000.

För båda H_0 och H_A gäller en signifikansnivå på 0,01.

Vi antar att urvalsundersökningen gjorts på lika fördelade oberoende stickprov.

Vi förkastar H_0 om $X_{obs} > X_{0,01}(v)$

där v = frihetsgrader = (antal klasser − 1) = 4−1 = 3

$$X_{obs} = \sum_{i=1}^{k} \frac{(E(n_i) - n_i)^2}{E(n_i)} \qquad X_{0,01}(3) = 11,345$$ VC

k = antal klasser

$$X_{obs} = \frac{(42-12)^2}{42} + \frac{(297-291)^2}{297} + \frac{(204-213)^2}{204}$$

$$+ \frac{(57-84)^2}{57} = 34,736$$ VC

$$X_{obs} = 34,736 > 11,345 = X_{0,01}(3)$$

H_0 kan förkastas.

Svar H_0 kan förkastas på signifikansnivån 1%.

dvs fördelningarna är inte lika 2014 som 2000.

10

SU, STATISTIK

Skrivsal: Värtasalen **Anonymkod:** SGD-0055 **Blad nr:** 7

4 5)

	Kvinnor	Män	
Undervikt	9	3	12
Normalvikt	157	133	290
Övervikt	84	130	214
Fetma	40	44	84
	290	310	600

Vi vill testa om fördelningen mellan män och kvinnor är samma mellan de olika BMI-kategorierna. Det gör vi med ett homogenitetstest med χ^2. Vi beräknar fördelningen för det totala urvalet och jämför det med fördelningen för kvinnor och män. Vi gör en tabell över de förväntade värdena

	Kvinnor $E(n_i)$	Män $E(n_i)$
Underviktig	$\frac{12}{600} \cdot 290 = 5{,}8$	$\frac{12}{600} \cdot 310 = 6{,}2$
Normalviktig	$\frac{290}{600} \cdot 290 = 140{,}17$	$\frac{290}{600} \cdot 310 = 149{,}83$
Överviktig	$\frac{214}{600} \cdot 290 = 103{,}43$	$\frac{214}{600} \cdot 310 = 110{,}57$
Fetma	$\frac{84}{600} \cdot 290 = 40{,}6$	$\frac{84}{600} \cdot 310 = 43{,}4$

Alla värden på de förväntade antalet är större än 5.

Vi sätter upp hypotesen:

H_0: Fördelningen mellan de olika BMI-kategorierna är lika för män och kvinnor

H_A: Fördelningen är inte lika.

Testvariabel $\chi^2_{obs} = \sum_{alla\ klasser} \frac{(E(n_i) - n_i)^2}{E(n_i)}$

$\chi^2_{krit} = \chi^2_{0,01}$ (antal rader -1)(antal kolumner -1) $=$

$\chi^2_{0,01}(2-1)(4-1) = \chi^2_{0,01}(3) = 11,345$

Vi förkastar H_0 om $\chi^2_{obs} > \chi^2_{0,01}(3) = 11,345$

$$\chi^2_{obs} = \frac{(5,8-9)^2}{5,8} + \frac{(140,17-157)^2}{140,17} + \frac{(103,43-84)^2}{103,43}$$

$$+ \frac{(40,6-40)^2}{40,6} + \frac{(6,2-3)^2}{6,2} + \frac{(149,83-133)^2}{149,83} +$$

$$\frac{(110,57-130)^2}{110,57} + \frac{(43,4-44)^2}{43,4} =$$

$1,7655 + 2,02075 + 3,65005 + 0,008866945$

$+ 1,65161 + 1,890469 + 3,414352 + 0,0082949 = 14,41$

$(14,40989963)$

$\chi^2_{obs} = 14,41 > 11,345 = \chi_{0,01}(3) \Rightarrow H_0$ förkastas

Svar. H_0 förkastas på signifikansnivån 1%

dvs fördelningarna är inte lika.

10

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: SGD-0055 Blad nr: 8

(5) $n = 40$

X - stokastisk variabel med väntevärde μ och varians σ^2. X är oberoende.

$\bar{X} = 1,033$ g

$S_x^2 = 0,01188$

Rekommenderad nivå $\mu_0 = 1,0$ gram

a) Tyder data på att maskinen är inställd på en högre nivå än den rekommenderad.

Hypotes:

$H_0: \mu = 1,0$ gram

$H_A: \mu > 1,0$ gram ensidigt test

Testvariabel: $Z = \dfrac{\bar{X} - \mu_0}{S_x / \sqrt{n}}$ där $Z \sim N(0,1)$ om H_0 är sann.

Vi kan använda oss av normalfördelningen eftersom (> 30) är så stort, så det spelar ingen roll vilken fördelning X kommer i från, enligt centrala gränsvärdessatsen.

$X \sim N(\mu, \sigma^2)$

Vi kan förkasta H_0 om $Z_{obs} > Z_{0,05} = 1,6449$

Vi sätter $\alpha = 0,05$ tills vidare.

$Z_{obs} = \dfrac{1,033 - 1}{\sqrt{0,01188} / \sqrt{40}} = 1,9149$

$1,9149 > 1,6449$ Vi kan förkasta H_0

på signifikansnivå 0,05, dvs det finns belägg för att tro att $\mu > 1$.

Vänd

Vi vill veta p-värdet

P-värdet är signifikansnivån vid det observerade värdet på $Z_{obs} = 1,9149$

Vi tittar i tabellen och får $z = 1,9149 \Rightarrow P(Z \leq z)$ $0,97139$

P-värdet blir då $1 - 0,97139 = 0,0286$

Det är alltså sannolikheten att få ett värde på 1,9149 eller större om H_0 är sann.

Så på signifikansnivå 5% kan vi förkasta H_0 men inte på signifikansnivå 1%.

Så om vi vill förkasta hypotesen att $\mu = 1$ och slå fast att vi tror att $\mu > 1$ beror på hur säkra vi vill vara, dvs vilken signifikansnivå vi väljer.

Svar a) Data tyder på att maskinen är inställd på en högre nivå, i alla fall om man nöjer sig med en signifikansnivå på $\alpha = 0,05$.

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: SGD-0055 Blad nr: 9

5b) $n = 40$ Y = vikt av ämnes, oberoende

$\mu = 15$

$\sigma_Y^2 = 3{,}25\sigma^2$

X och Y är oberoende.

$\hat{\mu}_1 = \frac{1}{2}(\bar{X} + \bar{Y})$ $\hat{\mu}_2 = \frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}$

$E(\bar{X}) = \mu$ $E(\bar{Y}) = E(\bar{X}) = \mu$

$E(\hat{\mu}_1) = E(\frac{1}{2}(\bar{X} + \bar{Y})) = \frac{1}{2} \cdot E(\bar{X} + \bar{Y}) =$

$\frac{1}{2}(E(\bar{X}) + E(\bar{Y})) = \frac{1}{2}(\mu + \mu) = \frac{1}{2} 2\mu = \mu$

$E(\hat{\mu}_1) = E(\frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}) = E(\frac{2}{3}\bar{X}) + E(\frac{1}{3}\bar{Y}) =$

$\frac{2}{3} \cdot E(\bar{X}) + \frac{1}{3} E(\bar{Y}) = \frac{2}{3}\mu + \frac{1}{3} \cdot \mu = \frac{2\mu + \mu}{3} =$

$\frac{3\mu}{3} = \mu$

$\Rightarrow$ $\hat{\mu}_1$ och $\hat{\mu}_2$ är båda väntevärdesriktiga, v.s.v.

Variansen:

$V(\hat{\mu}_1) = V(\frac{1}{2}(\bar{X} + \bar{Y})) = (\frac{1}{2})^2 \cdot V(\bar{X} + \bar{Y}) =$

$\frac{1}{4}(V(\bar{X}) + V(\bar{Y})) = \frac{1}{4}(\sigma^2 + 3{,}25\sigma^2)$

$= \frac{1}{4} \cdot (3{,}25\sigma^2) = \frac{3{,}25\sigma^2}{4} = 0{,}8125\sigma^2$

$V(\hat{\mu}_2) = V(\frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}) = V(\frac{2}{3}\bar{X}) + V(\frac{1}{3}\bar{Y}) =$

$\left(\frac{2}{3}\right)^2 V(\bar{X}) + \left(\frac{1}{3}\right)^2 V(\bar{Y}) = \frac{4}{9} V(\bar{X}) + \frac{1}{9} V(\bar{Y}) =$

$\frac{4}{9} \cdot \sigma^2 + \frac{1}{9} \cdot 2{,}25\sigma^2 = \frac{4\sigma^2 + 2{,}25\sigma^2}{9} = \frac{6{,}25\sigma^2}{9}$

$= 0{,}694\sigma^2$

Variansen av $\hat{\mu}_1$ är $0{,}8125\sigma^2$ och variansen för $\hat{\mu}_2$ är $0{,}694\sigma^2$

Svar: Den bästa estimatorn är $\hat{\mu}_2$. Båda estimatorerna är väntevärdesriktiga men $\hat{\mu}_2$ har mindre varians vilket är bra.

7