STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jessica Franzén

TENTAMEN I INTRODUKTION TILL STATISTIK FÖR STATSVETARE

2019-05-02

Skrivtid: 09.00-14.00
Godkända hjälpmedel: Miniräknare.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.

Uppgift 1 (20 poäng)
a) Hypoteserna i ett hypotestest är
$H_0 : \mu = 12$
$H_1 : \mu > 12$
p-värdet = 0.12 i detta hypotestest. Vilka slutsatser kan man/kan man inte dra av hypotestestet? Är resultatet statistiskt signifikant?

b) Beskriv skillnaden mellan enkel linjär regression och multipel regression. Ge exempel på en regressionsmodel av vardera sort.

c) Förklara begreppet kausalitet.

d) Beskriv skillnaden mellan en diskret och en kontinuerlig variabel, ge ett konkret exempel på en variabel av vardera sort.

e) Beskriv vad interpolation och extrapolation är.

Uppgift 2 (20 poäng)
a) Sannolikhetsfördelningen för X = antal pojkar i en slumpmässigt vald trebarnsfamilj ges nedan. Den baseras på att det är något högre sannolikhet att få en pojke än en flicka vid en födsel.

x	0	1	2	3
$p(x)$	0.11	0.36		0.14

i) Sannolikheten för $x = 2$ pojkar d.v.s. $p(2)$ saknas. Vad är $p(2)$?
ii) Vad är sannolikheten att en slumpmässigt vald trebarnsfamilj bara har flickor?
iii) Vad är sannolikheten att en slumpmässigt vald trebarnsfamilj har minst en pojke?
iv) Vad är väntevärdet av X d.v.s. väntevärdet av antal pojkar i en trebarnsfamilj? Tolka resultatet.

b) I en kommun har man slumpmässigt valt ut 60 trebarnsfamiljer och noterat antalet familjer med 0, 1, 2 respektive 3 pojkar. Resultatet ges i tabellen.

Pojkar	0	1	2	3
Antal familjer	10	25	19	6

Testa med ett χ^2-test om det bland kommunens trebarnsfamiljer råder samma fördelning mellan antal pojkar som man kan förvänta sig enligt sannolikhetsfördelningen i a). Använd siginifikansnivån 5 procent. Vilka slutsatser kan man/kan man inte dra av hypotestestet?

Uppgift 3 (20 poäng)
a) Vid ett val i USA undersökte man sambandet mellan utgifter för politisk TV-reklam och valdeltagande i det efterföljande valet. X är procent av de totala reklam/annonsutgifterna som läggs på TV-reklam i ett valdistriktet och Y är valdeltagande i procent i valdistriket. Man samlar in data (X och Y) för 20 slumpmässigt valda distrikt och skattar regressionssambandet till $\widehat{y} = 1.18x + 2.17$

i) Tolka den skattade koefficienten 1.18.
ii) Vad är det skattade valdeltagandet i ett distrikt där 45 procent av de totala reklam/annonskostnaderna läggs på TV-reklam?
iii) Korrelationskoefficienten $r = 0.95$. Vad innebär det?
iv) Beräkna determinationskoefficienten och tolka den.

b) Är respektive påstående nedan sant eller falskt. Motivera.

i) Om korrelationskoefficienten r är positiv innebär det att koefficienten b i regressionslinjen $\widehat{y} = bx + a$ också är positiv.
ii) $r = -0.78$ anger ett starkare linjärt samband än $r = 0.68$.
iii) Ju bättre en regressionslinje ansluter till observationerna, desto större är residualvariansen.
iv) Om $r = -0.99$ innebär detta att X och Y varierar så att stora Y-värden ofta förekommer hos de observationer som har stora X-värden och vice versa.

Uppgift 4 (20 poäng)
Vid ett stort företag har man utgått från att tidsåtgången för att utföra ett visst arbetsmoment i genomsnitt är 86 minuter. Från arbetstagarhåll menar man dock att denna tid är för kort och man stöder sig då på en egen undersökning som fackföreningen har gjort. Denna undersökning bygger på tidsstudier av ett slumpmässigt urval av 35 anställda. Dessa 35 anställda behövde i genomsnitt $\overline{x} = 90$ minuter på sig att utföra arbetsmomentet med en standardavvikelse $s = 9$. Man vill nu testa om tidsåtgången för att utföra arbetsmomentet, i genomsnitt för samtliga anställda, är längre än 86 minuter.

a) Vad gäller? Du är konsulterad! Genomför en hypotesprövning på signifikansnivån 1%. Ange noll- och mothypotes. Beskriv vilka slutsatser du kan/inte kan dra av hypotestestet.

b) Varför kan man inte avgöra direkt om tidsåtgången är större än 86 minuter genom att jämföra med värdet som undersökningen gav d.v.s. 90 minuter? Varför måste man utföra hypotestet i a)?

Uppgift 5 (20 poäng)
a) På företaget Alfa AB studerar man årsinkomsten X för de anställda. Medelårsinkomsten per år bland de anställda är $\mu = 270$ tusen kronor med standardavvikelse $\sigma = 20$. Årsinkomsten är Normalfördelad d.v.s. $X \sim N(270, 20^2)$.

i) Ungefär hur stor andel av de anställda tjänar mindre än 250 tusen kronor per år?
ii) Ungefär hur stor andel av de anställda tjänar mer än 310 tusen kronor per år?
Ledning: Hur stor andel av obervationerna i en Normalfördelning ligger ungefär inom $\pm$ 1 respektive 2 standaravvikelser från μ?

b) På företaget Beta AB vet man att årsinkomsten Y för de anställda är Normalfördelad men man vet inte väntevärde och varians. Man tar ett slumpmässigt urval av 7 anställda och noterar lönerna (i tusentals kronor), resultatet blev

350 222 198 256 311 285 333

Beräkna ett 95 procentigt konfidensintervall för årslönerna vid företaget. Hur ska konfidensintervallet tolkas?

Lycka till!

FORMELBLAD

DESKRIPTIV STATISTIK

Ett urval består av n stycken observationer.

Medelvärde:

$$\overline{x} = \frac{\sum x_i}{n}$$

Medelvärde från frekvenstabell:

$$\overline{x} = \frac{\sum f_i x_i}{n}$$

Vägt medelvärde:

$$\overline{x} = \sum_{i=1}^{L} w_i \overline{x}_i$$

Varians:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}$$

Varians från frekvenstabell:

$$s^2 = \frac{\sum f_i (x_i - \overline{x})^2}{n-1} = \frac{\sum f_i x_i^2 - n\overline{x}^2}{n-1}$$

VÄNTEVÄRDE OCH VARIANS

Väntevärde för X:

$$\mu = E(X) = \sum xP(x)$$

Varians för X:

$$\sigma^2 = V(X) = \sum [x-\mu]^2 P(x) = \sum x^2 P(x) - \mu^2$$

SAMPLINGFÖRDELNINGAR OCH CENTRALA GRÄNSVÄRDESSATSEN

Om populationen är normalfördelad med väntevärde μ och varians σ^2 dvs $X \sim N(\mu, \sigma^2)$

så är $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

Om populationen har en annan fördelning (vilken som helst) med väntevärde μ och varians σ^2

så är $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ om $n \geq 30$ enligt centrala gränsvärdessatsen

STATISTISK INFERENS

Konfidensintervall

För medelvärden (Normalfördelade data eller $n \geq 30$)

$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ alternativt $\overline{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$

För proportioner $(np(1-p) > 5)$

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

Antal observationer för en given bredd d på konfidensintervallet:

För medelvärden

$n = \frac{2^2 Z^2_{\alpha/2}\sigma^2}{d^2}$

För proportioner

$n = \frac{2^2 Z^2_{\alpha/2}p(1-p)}{d^2}$

Hypotesprövning

För medelvärden (Normalfördelade data eller $n \geq 30$)

$Z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ alternativt $Z = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$

För jämförelse av medelvärden (Normalfördelade data eller $n_1 \geq 30$ och $n_2 \geq 30$)

$Z = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

För proportioner $(np(1-p) > 5)$

$Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

För jämförelse av proportioner $(n_1p_1(1-p_1) > 5$ och $n_2p_2(1-p_2) > 5)$

$Z = \frac{p_1-p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

χ^2- **test** ($E > 5$ för alla kategorier)

$\chi^2 = \sum \frac{(O-E)^2}{E}$

REGRESSION

Skattning av regressionslinjen $\widehat{y} = bx + a$

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = \overline{y} - b\overline{x}$$

$$\widehat{y} = bx + a$$

Residualvarians:

$$s_e^2 = \frac{\sum e^2}{n-2} = \frac{\sum (y - \widehat{y})^2}{n-2}$$

Residualspridning:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum (y - \widehat{y})^2}{n-2}}$$

Korrelation:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

Determinationskoefficient (enkel linjär regression):

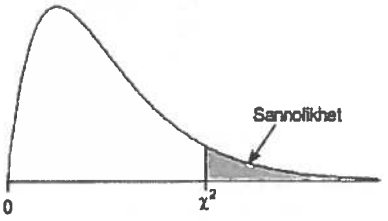
$$R^2 = r^2$$

INDEX

$I_t = \frac{x_t}{x_b} 100$ $\qquad$ $I_t = \frac{I_t}{I_b} 100$

Tabell 4 χ^2-fördelningen

χ^2-värden för vissa sannolikheter.



fg	Sannolikhet 5%	1%	0,1%
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,12
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,72	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,31
21	32,67	38,93	46,80
22	33,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62

fg	Sannolikhet 5%	1%	0,1%
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70
31	44,99	52,19	61,10
32	46,19	53,49	62,49
33	47,40	54,78	63,87
34	48,60	56,06	65,25
35	49,80	57,34	66,62
36	51,00	58,62	67,99
37	52,19	59,89	69,35
38	53,38	61,16	70,71
39	54,57	62,43	72,06
40	55,76	63,69	73,41
41	56,94	64,95	74,75
42	58,12	66,21	76,09
43	59,30	67,46	77,41
44	60,48	68,71	78,75
45	61,66	69,96	80,08
46	62,83	71,20	81,39
47	64,00	72,44	82,72
48	65,17	73,68	84,03
49	66,34	74,92	85,35
50	67,50	76,15	86,66

Tabellvärden för standardiserad Normalfördelning:

α	Z_α
0.005	2.58
0.01	2.33
0.025	1.96
0.05	1.64
0.1	1.28

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 2/5-2019

Sal: Brunnsvikssalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: 6007-DTP

☒ 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. 18	20	20	17	15					

GL

POÄNG	BETYG	Lärarens sign.
90	A	

SU, STATISTIK

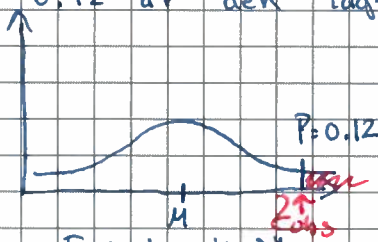
Skrivsal: BR Anonymkod: 0007-DTP Blad nr: 1

1.

a) Hypotestest

$H_0: \mu_0 = 12$
$H_1: \mu_0 > 12$ P-värde = 0.12

Detta säger oss att vi vill testa hypotesen att medelvärdet för något är större än 12. P-värdet säger oss att 0.12 är den lägsta signifikansnivå vi kan förkasta H_0 på.



Om vårt $Z_{obs} > 0.12$ kan vi förkasta H_0. Även på samtliga signifikansnivåer över detta. Då kan vi även säga att resultaten är statistiskt signifikant.

2

b) Enkel linjär regression är en beskrivning av samverkan mellan två variabler. Denna visar hur Y den beroende variabeln påverkas av förändringar i X den oberoende variabeln.

Modellen ser ut såhär: (X = Ålder Y = Lön)

$\hat{Y} = a + bx$ (när man har urval och regressionslinjen är skattad)

$Y = a + bx$ (när den ej är skattad)

Multipel regression är då man beräknar och kan se samverkan mellan fler än 2 variabler. Den innehåller alltså fler förklaringsvariabler. Regressionslinjen behöver här inte vara linjär utan kan påvisa andra icke-linjära samband.

Den kan se ut såhär:

$Y = a + b_1X_1 + b_2X_2$

(Y = Lön, X_1 = ålder X_2 = Arbetslivserfarenhet).

4

c) Kausalitet är när variabeln X (ex. kan va andra variabler) har direkt påverkan på variabeln Y. Till skillnad från korrelation som behandlar samband mellan variabler så handlar kausalitet om orsak-verkan samband.

Det finns tre krav för att kunna bekräfta kausalitet:

1. Att det finns någon form av uppmätt samband ex. Korrelation

2. Att man har uteslutit påverkan från andra förklaringsvariabler

3. Att man säkerställer tidsaspekten - orsak kommer före verkan.

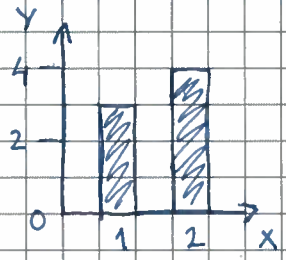
4

d) En diskret variabel är en kvantitativ variabel som enbart kan anta vissa värden. Exempel är antal barn. Där kan bara heltal antas då du ej kan ha 3.7 barn.

En kontinuerlig variabel är också kvantitativ, men kan till skillnad från diskreta variabler anta alla värden. Detta kan exempelvis vara hur långt man har till jobbet i km. Där kan man anta värden som 6.3, 1.4 eller 0,2.

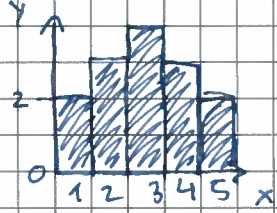
4

De skiljer sig även vad gäller deskription. Diskreta variabler presenteras, grafiskt, oftast i stapeldiagram. Detta för att påvisa avståndet mellan rektanglarna vilket representerar variabelns möjlighet att enbart anta vissa värden.



R

En kontinuerlig variabel redovisas oftast med histogram där de rektangulära staplarna hänger ihop för att påvisa att alla värden kan antas.



R

e) Extrapolation är när man gör estimeringar av värdet på en variabel som ligger utanför intervallet av observerade värden. Här bör man iaktta försiktighet då det kan påvisa falska samband hos variabler. Exempel är om man mäter samband mellan ålder och inkomst. Finns risk att om man tar ålder=0 finns en inkomst - vilket är omöjligt.

Kända värden

Interpolation är tvärtom, det är estimering av värdet på variabler inom intervallet av kända observationer.

4

SU, STATISTIK

Skrivsal: BR Anonymkod: 0007-DTP Blad nr: 2.

2.

a) X = antal pojkar i en slumpmässigt vald 3 barns familj.

i) En sannolikhetsfördelnings olika värden i P(X) ska alltid summeras till 1.

$P(2) = 1 - (P(0)=0.11) - (P(1)=0.36) - (P(3)=0.14) = 0.39$

Sannolikheten för två pojkar är alltså 0.39. (2)

ii) $X = 0$

$P(0) = 0.11$ Sannolikheten att en familj bara har flickor är 0.11 (11%). (2)

iii) Har minst en pojke ges av;

$P(X) = (P(1)=0.36) + (P(2)=0.39) + (P(3)=0.14) = 0.89$ (2)

Sannolikheten att en familj har minsten pojke är 0.89 (89%)

iiii) Väntevärde:

$\mu = E(x) = \sum x\,P(x)$

$\mu = (0 \times 0.11) + (1 \times 0.36) + (2 \times 0.39) + (3 \times 0.14)$
$= 0 + 0.36 + 0.78 + 0.42$
$= 1.56.$ (4)

Vänte värdet är ett teoretiskt lägesmått som i detta fall visar att, baserat på sannolikhets-funktionen, är medelvärdet av antal pojkar i 3 barnsfamiljer 1.56.

b) $n = 60$

X Pojkar:	0	1	2	3	
Y Antal familjer:	10	25	19	6	$\rightarrow \sum Y = 60$

Ö:

X	Y
0	10
1	25
2	19
3	6

H_0: Antal familjer med 0, 1, 2, 3 antal pojkar är lika fördelat som i ovanstående sannolikhetsfunktion.

H_1: Antal familjer med 0, 1, 2, 3 antal pojkar följer inte fördelningen i sannolikhetsfunktionen.

E_i

X	Y
0	$60 \times 0.11 = 6.6$
1	$60 \times 0.36 = 21.6$
2	$60 \times 0.39 = 23.4$
3	$60 \times 0.14 = 8.4$

Signifikansnivå = 5%

Frihetsgrader = $(k-1) = (4-1) = 3$

$\chi^2_{0,05}(3) = 7.81$

Vår kritiska gräns är 7.81. Vi förkastar H_0 då $\chi^2_{obs} > \chi^2_{0,05}(3) = 7.81$

Testvariabel:

$$\chi^2_{obs} = \sum \frac{(O_i - E_i)^2}{E_i}$$

Villkor $E_i > 5$
Samtliga E_i är större än 5 så vi genomför test.

$$\chi^2_{obs} = \frac{(10-6.6)^2}{6.6} + \frac{(25-21.6)^2}{21.6} + \frac{(19-23.4)^2}{23.4} + \frac{(6-8.4)^2}{8.4}$$

$$\chi^2_{obs} = 3.799776505$$

$\chi^2_{obs} = 3.799776505 < \chi^2_{0,05}(3) = 7.81$

Därav kan vi ej förkasta H_0. Vi kan säga att detta urval är fördelat enligt sannolikhetsfunktionen i a).

10

SU, STATISTIK

Skrivsal: BR Anonymkod: 0007 - DTP Blad nr: 3

3. a) x = procent av de totala reklam/annonsutgifterna
y = valdeltagande i procent i valdistrikt.

$n = 20$

$\hat{y} = 1.18x + 2.17.$

i). 1.18 är den skattade riktningskoefficienten som säger hur mycket linjen lutar samt åt vilket håll.

I detta fall säger den att för varje ökad procentenhet av totala reklam/annonsutgifter X, ökar valdeltagande i distriktet Y med 1.18 enheter. (2)

ii). $X = 45\% = 0.45$

$\hat{y} = (1.18 \times 45) + 2.17 = \hat{y} = 55.27$ (2)

Valdeltagande i distrikt där 45% av de totala reklam/annonskostnaderna läggs på TV-reklam är 55.27 %.

iii). Detta ($r = 0.95$) innebär att det finns ett starkt positivt linjärt samband mellan andel av totala reklam/annonskostnader för TV-reklam och andelen valdeltagande i distriktet.

Korrelation mäter samband mellan 2 variabler och kan anta värden mellan -1 och 1. Där sambandet är starkare ju närmare +1 och 1 man befinner sig. 0 = inget samband. -1 = starkt negativt linjärt samband. 1 = starkt positivt linjärt samband. (4)

iv). Determinationskoefficienten

$R^2 = r^2$

$R^2 = 0.95^2 = 0.9025$ $\quad 0.9025 \times 100 = 90.25\%$.

90.25 % av variationen i Y dvs. den beroende variabeln kan förklaras av vårt skattade regressionssamband $\hat{y} = 1.18x + 2.17$. (4)

b).

i). Sant. Då riktningskoefficienten b i regressionslinjen anger lutningen samt håller. Är den positiv betyder Det att för varje ökat värde i X ökar värdet i Y.
Det är en positiv samvariation.

Även matematiskt kan vi se detta då b i regressionslinjen beräknas:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

där för ett positivt värde måste täljaren vara positiv.
Nämnaren kan aldrig vara negativ

Bra!

Därav vid beräkning av korrelation, är täljaren densamma som vid beräkning av b, vilket gör att även den då är positiv.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

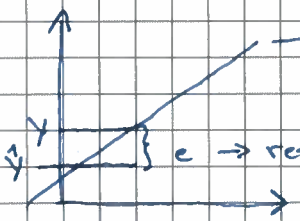
↳ Då nämnaren aldrig kan vara negativ

2

ii). Sant. Som beskrivet om korrelation i del a) av fråga 3 utgörs skalan mellan -1 och 1.
Desto närmare dessa heltal vår riktningskoefficient kommer desto starkare är sambandet. $r = -0,78$ är närmare -1 än $r = 0.68$ är nära 1 vilket gör att $r = -0,78$ har ett starkare samband. Skillnad är att detta är negativt och det andra är positivt, men det spelar ingen roll då $-0,78$ är starkare.

2

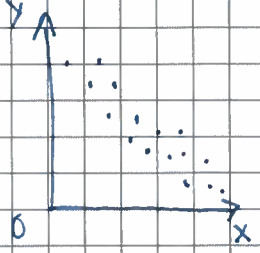
iii). Falskt. Residualvariansen anger residualstandardavvikelsen. Detta innebär ett spridningsmått för spridningen kring regressionslinjen. Dvs. skillnaden mellan y (uppmätt värde) och $\hat{y}$ (det av regressionslinjen skattade värdet.



Därav är residualvariansen mindre desto närmare vår regressionslinje ligger observationerna

2

iv). Falskt. $r = -0,99$ är en stark negativ linjär korrelation som ser ut ungefär såhär:



Här kan vi se att höga värden för y ger mindre värden för x. och stora värden för x ger små värden för y.

2

SU, STATISTIK

Skrivsal: BR Anonymkod: 0007-DTP Blad nr: 4.

4. $\bar{x} = 86$ minuter

$n = 35 \quad \bar{x} = 90 \quad s = 9$

a) Hypotesprövning.

Vi vill testa om tidsåtgången för momentet är längre än 86 minuter.

$H_0: \mu_0 = 86$
$H_1: \mu_0 > 86$ R

Signifikansnivå = 1%. $z_{0,01} = 2.33$ R

Vår kritiska gräns är 2.33 och vi förkastar H_0 då $z_{obs} > z_{0,01} = 2.33$.

Villkor $n \geq 30$ i detta fall är $n = 35$ vilket är ok enligt CGS så vi genomför test. R

Testvariabel:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow z_{obs} = \frac{90 - 86}{9/\sqrt{35}} = \frac{4}{1.521277659}$$

$= 2.629368792$ R

$z_{obs} = 2.629368792 > z_{0,01} = 2.33$ R

Vi kan alltså förkasta H_0, vi kan därigenom dra slutsatsen att det krävs längre tid än 86 minuter i genomsnitt att utföra ett visst arbetsmoment. 12

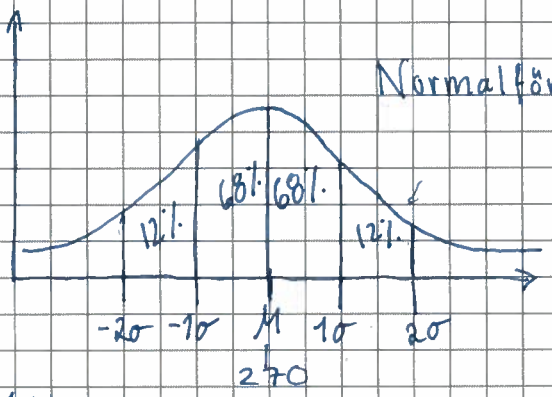
b).

Om man inte genomför ett hypotestest kan man inte säga att det är statistiskt signifikant dvs. statistiskt säkerställt att det krävs mer än, i genomsnitt, 86 minuter att genomföra ett visst arbetsmoment.

5

5. a). X = årsinkomst

$\mu = 270$ tkr $\quad \sigma = 20$ normalfördelad $X \sim N(270, 20^2)$



i). $1\sigma = 20$

$X < 250$ tkr $\quad 100 - 68\% = \underline{32\%}$ tjänar mindre än 250 tkr

ii). $X > 310$ $\quad 100 - 80\% = \underline{20\%}$ tjänar mer än 310 tkr.

✓ ✓

(2)

SU, STATISTIK

Skrivsal: BR Anonymkod: 0007-DTP Blad nr: 5

5. b). Y = Årsinkomst n = 7 Y är Normalfördelat

350, 222, 198, 256, 311, 285, 333

Konfidens intervall för medelvärde:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{350 + 222 + 198 + 256 + 311 + 285 + 333}{7}$$

$$\bar{x} = 279.2857143$$

95 %-igt konfidensintervall, ges av:

$$\bar{x} \pm z_{0,05/2} \frac{s}{\sqrt{n}}$$

$$s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = \frac{(350^2) + (222^2) + (198^2) + (256^2) + (311^2) + (285^2) + (333^2) - (7 \times 279.2857143^2)}{7-1}$$

$$= \sqrt{\frac{565359 - 546003.5714}{6}} = \sqrt{\frac{19355.4286}{6}}$$

$$= \sqrt{3225.904767} = 56.79704893$$

$z_{0,05/2} = z_{0,025} = 1.96$

$$279.2857143 \pm 1.96 \frac{56.79704893}{\sqrt{7}}$$

$$= 279.2857143 \pm 1.96 \times 21.46726667$$

$$= 279.2857143 \pm 42.07584267$$

$$= [237.2098716; 321.3615571] \approx [237.2; 321.4]$$ VR

Utifrån detta 95%-iga konfidensintervall kan vi se att det sanna medelvärdet för årsinkomsterna ligger någonstans mellan vår undre gräns: 237.2 tkr och vår övre gräns: 321.4 tkr.

Nej det vet vi inte säkert!

13