STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jolanta Pielaszkiewicz

TENTAMEN I STATISTIKENS GRUNDER 2
2018-11-26

Skrivtid: 12.00-17.00
Godkända hjälpmedel: Miniräknare, språklexikon.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.

Uppgift 1. (20 poäng) Tidsserier för antal hyresrätt och bostadsrätt lägenheter i nybyggda ordinära flerbostadshus (storstadsområdena) från 2012 till 2016 visas i tabellen nedan (källa: SCB).

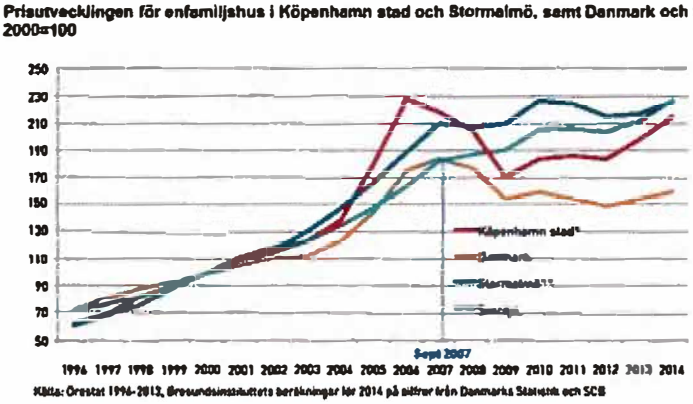
År:	2012	2013	2014	2015	2016
hyresrätt :	2 532	3 413	3 722	4 009	5 457
bostadsrätt:	5 639	7 390	8 105	10 944	13 346

a) Beräkna urvalskorrelation mellan antalet hyres- och bostadsrättslägenheter. Tolka resultatet.

b) Ange en indexserie för utveckling av antal bostadsrätter med basår 2012.

c) Anta att bostadsrättsindex för 2010 är 96.3 (i %, indexerat med år 2012 som basår). Använd tidsserierna som finns i tabellen för att bestämma antalet bostadsrätter i nybyggda ordinära ferbostadshus i 2010.

d) Använd diagrammet nedan för att svara på följande fråga: Kan man säga att priserna för enfamiljshus i Sverige var högre än i Danmark efter september 2007? Motivera svaret.
Bild från: `https://www.oresundsinstituttet.org/fakta/bostadsmarknaden/`



Uppgift 2. (20 poäng)

a) Vi har stickprov: 162, 90, 150, 143, 148, 150, 134, 151, 121. Beräkna median och ge lådagram (boxplot) eller stamblad-diagram för datamaterialet (du får välja en grafisk metod).

b) Hur stor stickprov ska vi välja så att 95% konfidensintervall för populationsmedelvärdet är kortare än 0.5? Anta $\sigma = 2$.

Fortsätt 5 påståenden som finns nedan.

c1) Typ-II-felet är sannolikheten att ....
c2) t-fördelning närmar sig normal-fördelning när ...
c3) Exempel på centralmått som påverkas inte av outliers är ....
c4) p-värdet är ....
c5) Indexserie används för att ...

Uppgift 3. (20 poäng) I en enkätundersökning av 200 slumpmässigt utvalda studenter tillfrågades dessa bland annat om aktuell livskvalitet. Nedan visas svarsfördelningen i olika åldersgrupper.

Svaret (X):	Åldersgrupp			TOTAL
	under 22	22-26	över 27	
Bra	22	58	20	100
Lagom	10	40	20	70
Inte bra	8	12	10	30

a) Hypotespröva om det finns signifikant samband mellan ålder och aktuell livskvalitet på $\alpha = 2.5\%$.

b) Hypotespröva på $\alpha = 5\%$ om andel av studenter som tycker att deras livskvalitet är "Inte bra" ligger under 20%. Storlek av populationen är 800.

OBS. Kom ihåg att skriva H_0, H_1, teststorhet, slutsatser (med motivation).

Uppgift 4. (20 poäng) Ballonger av olika färg produceras i mycket stora mängder. Ballongvikt bör vara 5 (gram). Ett slumpmässigt urval från populationen ger:

$$\bar{x} = 5.5 \qquad s = 1.85 \qquad n = 24$$

Normalfördelningen antas.

a) Hypotespröva $H_0 : \mu = 5$ mot $H_0 : \mu > 5$ på signifikansnivå $\alpha = 5\%$. Tolka resultatet.

b) Uppskatta och tolka p-värdet för testet.

c) Beräkna ett tvåsidigt 95%-konfidensintervall för μ.

Uppgift 5. (20 poäng) Två företag gör marknadsundersökning för en mycket stor population av sina kunder. Resultat:

Det första företaget: $\bar{x} = 87.5 \qquad s_x = 1.5 \qquad n_1 = 100$
Det andra företaget: $\bar{y} = 88.5 \qquad s_y = 3.0 \qquad n_2 = 100$

a) Använd lämpligt test för att kontrollera om det finns signifikant skillnad i företagens resultat på $\alpha = 1\%$. Beräkna p-värdet.

b) Ytterligare information finns tillgänglig:

Det finns en samvariation mellan resultaten och årets säsong.

Kan vi säga att vi har ett orsakssamband (kausalt) mellan resultaten och årets säsong? Motivera svaret.

c) Beräkna urvalsvarians om lika varians för det första och andra företaget antas.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 26/11/18

Sal: Värtasalen

Tenta: Statistikens grunder

Kurs: Statistikens grunder 2

ANONYMKOD: 0012-HPK

☒ 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					11 8C
Lär.ant. 20	14	16.5	19.5	13					

POÄNG	BETYG	Lärarens sign.
83	B	Pedersen

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0012-HPK Blad nr: 1

1. a) Läs baksida!

För att beräkna urvalskorrelationen använder vi oss av

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}$$

där S_{XY} är kovarianssen som vi får av

$$S_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Där $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ och $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

R $\bar{x} = \frac{1}{5}(2532 + 3413 + 3722 + 4009 + 5457) = 3826.6$

R $\bar{y} = \frac{1}{5}(5639 + 7390 + 8105 + 10944 + 13346) = 9084.8$

x representerar hyresrätter och y representerar bostadsrätter

Standardavvikelserna S_X och S_Y får vi med

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$$

R $S_X = \left(\frac{1}{5-1}(2532^2 + 3413^2 + 3722^2 + 4009^2 + 5457^2 - 5 \cdot 3826.6^2)\right)^{1/2} \approx 1066.47$

R $S_Y = \left(\frac{1}{5-1}(5639^2 + 7390^2 + 8105^2 + 10944 + 13346^2 - n \cdot 9084.8^2)\right)^{1/2} \approx 3054.52$

R $S_{XY} = \frac{1}{5-1}(2532 \cdot 5639 + 3413 \cdot 7390 + 3722 \cdot 8105 + 4009 \cdot 10944 + 5457 \cdot 13346 - 5 \cdot 3826.6 \cdot 9084.8) \approx 3137741.9$

R $r_{XY} = \frac{3137741.9}{1066.47 \cdot 3054.52} = 0.9632$ SVAR: $r_{XY} = 0.9632$

Vi ser en väldigt stark korrelation. När antingen byggandet av hyres- eller bostadsrätt ökar så ökar också byggandet av det andra. /8

SU, STATISTIK

Skrivsal: Vårtavägen Anonymkod: 0012-H3K Blad nr: 2

1. b)

Indexserie ges av:

$$I_t^b = \frac{x_t}{x_b} \cdot 100$$

och vi är intresserade av 2012 som bas. Vi kollar enbart på bostadsrätter.

$$I_{2012}^{2012} = \frac{5639}{5639} \cdot 100 = 100$$

$$I_{2013}^{2012} = \frac{7390}{5639} \cdot 100 \approx 131{,}05$$

$$I_{2014}^{2012} = \frac{8105}{5639} \cdot 100 \approx 143.73$$

$$I_{2015}^{2012} = \frac{10944}{5639} \cdot 100 \approx 194.08$$

$$I_{2016}^{2012} = \frac{13346}{5639} \cdot 100 \approx 236.67$$

Svar: Serien blir:

2012	2013	2014	2015	2016
100,	131,05,	143,73,	194,08,	236,67

c)

Vi vet att indexet 2010, med 2012 som basår, är 96.3

Det innebär att vi kan ställa upp:

$$I_{2010}^{2012} = \frac{x_{2010}}{5639} \cdot 100 = 96.3$$

En ekvation med en okänd som vi kan lösa.

$$x_{2010} = \frac{96.3 \cdot 5639}{100} \approx 5430.36$$

Svar: Det byggdes 5431 bostadsrätter 2010.

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: COR-HPK Blad nr: 3

1) d)

Om vi kollar på hur en indexserie skapas

$$I_t^b = \frac{x_t}{x_b} \cdot 100 \qquad (1)$$

så ser vi att vi jämför med ett basvärde i samma serie.
Dvs i förra uppgiften jämförde vi bostadsrätter med bostadsrätter.

Därför när vi kollar på diagrammet som jämför prisutvecklingen i de olika regionerna så jämförs priserna inom regionen och inte mellan dem.

Allt diagrammet säger är att priserna i Sverige har fortsatt öka efter 2007 medan de en period sjönk i Danmark.

För att kunna avgöra det behöver vi veta baspriset, x_b, och lösa ekvation (1). Baspriset framgår inte i diagrammet.

Svar: Nej det går inte att avgöra var priserna är högast med hjälp av diagrammet.

Korrekt svar, bra! /4

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0012-HPK Blad nr: 4

2. a)

Börjar med att ordna listan.

[90, 121, 134, 143, 148, 150, 150, 151, 162]

Antal 4 Antal 4

Vi har ett udda antal värden så det mittersta blir vår median.

Svar: Medianen är 148 K /2

Black-out vad diagrammen innebär.

b) Vi antar att $\sigma = 2$ samt att vi kommer få fler än 30 mätningar så vi kan anta

$Z \sim$ approx $N(0,1)$ enligt CGS.

Det ger oss, där B är halva bredden:

$$2B < 0,5$$

$$2B = 0,5$$

$$B = 0,25$$

Antag att vi använder Z som test-variabel

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0,25$$

K: $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ = B

$$n = \left(\frac{z_{0,025}\cdot 2}{0,25}\right)^2 = \{z_{0,025} = 1.96\} \approx 156.8$$

Svar: Vi behöver ett stickprov på minst 157 för att få 0,5 bredd på vårt konfidensintervall för populationens medelvärdet.

/4

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0012HPK Blad nr: 5

2)

C_1: Typ II fel är att inte förkasta nollhypotesen, H_0, när den är falsk.

R P(Behåll H_0 / H_0 falsk)

C_2: En t-fördelning går mot normalfördelning
R då $n \to \infty$. Vid $n > 30$ kan man börja anta en approximativ normal fördelning.

C_3: Pass... :(

C_4: P-värdet är gränsen mellan när nollhypotesen kan förkastas och inte.
R
Ex: p-värde 0,026 säger att nollhypotesen kan förkastas på signifikansnivåer större än 0,026 som 0.05 och 0.1 men inte under, som 0,025 och 0.01.

C_5: Index-serier används för att kunna jämföra utvecklingen, exempel pris,
R i förhållande till något basvärde på ett enklare sätt (procent).

18

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 00K-HPK Blad nr: 6

3) a) Vi kan inte ta reda på vad för samband som eventuellt skulle existera men för att kunna visa att ett samband använder vi χ^2-testet.

	<22	22-26	27<	Antal
Bra	22 20	58 55	20 25	100
Lagom	10 14	40 38,8	20 17.5	70
Inte bra	8 6	12 16.5	10 7.5	30
	40	110	50	200

/3

Hypotes: H_0: Det existerar ett samband mellan åldersgrupperna och dirskvalitet.

/0

H_A: Det existerar inte ett samband.

För att kunna ta reda på teststorheten behöver vi väntevärdena.

$\frac{100}{200} \cdot 40 = 20$ $\frac{100}{200} \cdot 110 = 55$ $\frac{100}{200} \cdot 50 = 25$

$\frac{70}{200} \cdot 40 = 14$ $\frac{70}{200} \cdot 110 = 38,8$ $\frac{70}{200} \cdot 50 = 17.5$

$\frac{30}{200} \cdot 40 = 6$ $\frac{30}{200} \cdot 110 = 16.5$ $\frac{30}{200} \cdot 50 = 7.5$

/R

Vi observerar att alla väntevärden är över 5 så vi kan fortsätta med att ta fram en testvariabel.

R

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

R

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 00U-HPK Blad nr: 7

3 a) fortsättning

$$\chi^2_{obs} = \frac{(22-20)^2}{20} + \frac{(58-55)^2}{55} + \frac{(20-25)^2}{25} + \frac{(10-14)^2}{14} + \frac{(40-38,8)^2}{38,8} + \frac{(20-17.5)^2}{17,5} + \frac{(8-6)^2}{6} + \frac{(12-16.5)^2}{16.5} + \frac{(10-7.5)^2}{7,5} \approx 5,65$$ /2

om $\chi^2_{obs} > \chi^2_{0,025}$ så kan vi förkasta vår nollhypotes och anta att det saknas samband.

$$\chi^2_{0,025}\left((3-1)(3-1)\right) = \chi^2_{0.025}(4) = 11.143$$ /2

(3-1)(3-1): Rader Kolonner i våra matriser.

/2 Vi ser att $\chi^2_{obs} \not> \chi^2_{0,025}(4)$ vilket innebär att vi inte kan förkasta vår nollhypotes, att det existerar ett signifikant samband.

Svar: Går ej utesluta att ett samband existerar. /9

b) Vi börjar med att ta fram våra två hypoteser:

/1 $H_0: \pi = 0,20$
$H_A: \pi < 0,20$

I enkäten har 30 personer av 200 sagt att deras livssituation inte är bra.

k $p = \frac{30}{200} = 0,15$ /1

$30 \cdot p > 5$
$30 \cdot q > 5$

$n > 30 \longrightarrow Z \sim$ approx $N(0,1)$ enligt CGS. /1

Tar fram vår teststorhet $Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ och om $|Z_{obs}| > Z_{0.05}$ så förkastar H_0.

ett sidigt test

$$Z_{obs} = \frac{0.15-0,2}{\sqrt{\frac{0,2(1-0,2)}{200}}} \approx -1.77$$ /2

$\frac{N-n}{N-1}$

3b) fortsättning.

Vi fick på föregående sida att $z_{obs} = -1.77$.

Fortsätter med att ta fram $z_{0,05}$ som då är ett tabellvärde.

$$z_{0.05} = 1.6449 \quad /1$$

$$|z_{obs}| = 1.77 > 1.6449 = z_{0,05}$$

Testvariabeln är större och vi kan förkasta vårat nollhypotes. på $\alpha = 5\%$.

Svar: Andelen studenter som anser att deras livskvalitet inte är bra är färre än 20%. /1.5

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0012-HPK Blad nr: 9

4) a) Vi vet följande:

$\bar{x} = 5.5$, $s = 1.85$, $n = 24$

Vi antar normalfördelning. Eftersom $n < 30$ använder vi t-test.

R Hypotes: $H_0: \mu = 5$
$H_A: \mu > 5$

med signifikansnivå $\alpha = 0.05$.

R Vi förkastar H_0 om $t_{obs} > t_{0.05}(n-1)$

R $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.5 - 5}{1.85/\sqrt{24}} \approx 1.32$

R $t_{0.05}(24-1) = \{\text{tabell}\} = 1.714$

Svar: Vi ser att $t_{obs} \not> t_{0.05}(23)$ vilket innebär att vi inte kan förkasta att den faktiska medelbatterivikten ligger på 5 gram som den bör ligga på. /9

b) Vi vill ta fram p-värdet för:

$P(X > 5.5)$

$P(T > \frac{5.5-5}{1.85/\sqrt{24}}) = P(T > 1.32) = 1 - P(T < 1.32) = \{\text{Antar normal-fördelning}\}$

$= 1 - \Phi(1.32) = 0.093$

du har t-fördelning $n < 30$

Svar: Vi kan förkasta nollhypotesen, att batterierna vikt ligger runt 5 g, på signifikansnivåer 0,094 och högre. Ex o.l. /3,5

c) Vi får det två-sidiga intervallet med hjälp av:

$\bar{x} \pm t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}$, $t_{0.025}(23) = 2.069$

$5.5 \pm 2.069 \cdot \frac{1.85}{\sqrt{24}} = (4.72, 6.28)$ R

Svar: (4,72, 6,28) /7

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: UD12-HPK Blad nr: 10

5) a) Vi har två företag. Om det skall existera någon skillnad så måste deras värde μ skilja sig.

Hypotes: H_0 $\mu_X = \mu_Y$ /1

H_a $\mu_X \neq \mu_Y$

Eftersom $n_X > 100$ och $n_Y > 100$ så kan vi använda oss av

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \quad N \text{ approx } N(0,1) \qquad /1$$

Resultat från undersökningen: $\bar{x} = 87.5$, $s_X = 1.5$, $n_X = 100$
$\bar{y} = 88.5$, $s_Y = 3.0$, $n_Y = 100$

/2 $$Z_{obs} = \frac{87.5 - 88.5}{\sqrt{\frac{1.5^2 + 3^2}{100}}} \approx -2,98$$

Om $|Z_{obs}| > z_{0.01}$ så kan vi förkasta vår nollhypotes.

$z_{0.01} = \{\text{tabell}\} = 2,3263$ /1.5

Vi ser att $|Z_{obs}| = |-2.98| = 2.98 > 2.3262 = z_{0.01}$

Därmed kan vi förkasta nollhypotesen att de har fått samma resultat. $p < \alpha$ $\alpha = 1.5$

Svar: Det existerar skillnader i resultatet.

Beräknar p-värdet genom att anta ett resultat som μ.

$$P(X > 87.5) = P\left(Z > \frac{87.5 - 88.5}{1.5/\sqrt{100}}\right) = P(Z > -6.67) =$$

$$= 1 - P(Z < -6.67) = 1 - \Phi(-6.67) = \{\Phi(-x) = 1 - \Phi(x)\}$$

$$= 1 - (1 - \Phi(6.67)) = \Phi(6.67) \approx 1$$

6.67 existerar inte på tabellen. Går så gott som garantera att nollhypotesen förkastas. /1

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0012-HPK Blad nr: 4

5) b)

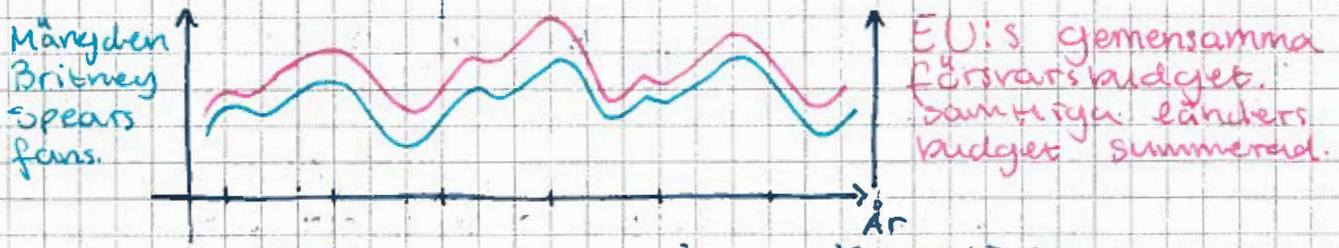
Kausalitet ⟶ Samvariation

R Samvariation ⟶/ Kausalitet

Samvariation innebär att två olika variabler förändras på samma sätt. Ex om en variabel ökar ökar även den andra.

Kausalitet innebär att den ena variabeln påverkar den andra och på så sätt rör sig variablerna på ett gemensamt sätt.

Exempel: (påhittat!!!)



Samvariation men ingen kausalitet.

Svar: Vi kan inte med säkerhet säga att det existerar ett kausalt samband mellan säsongerna då vi enbart vet att samvariation existerar.

c) Vi har redan standardavvikelsena s_x och s_y för de båda företagen.

Variuns är standardavvikelsen i kvadrat.

$S_x = 1.5^2 = 2,25$

$S_y = 3.0^2 = 9.0$

Svar: Urvalsvarianten för företag 1 är $s_x^2 = 2,25$.

Urvalsvariansen för företag 2 är $s_y^2 = 9.0$