STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)
Per Gösta Andersson

HT2017

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2017-10-25

Skrivtid:	kl. 15.00 - 20.00
Godkända hjälpmedel:	Miniräknare utan lagrade formler och text
Bifogade hjälpmedel:	Häftet *Formelsamling och Tabeller över statistiska fördelningar* (återlämnas efter skrivningen)

- Tentamen består av 7 uppgifter, i förekommande fall uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

- **Uppgift 1 – 5**: Svar lämnas på särskild **SVARSBILAGA**,
  - Flervalsfrågor där ett av fem alternativ är korrekt svar.
  - Har fler än ett svarsalternativ markerats för en deluppgift ges noll poäng.
  - Uträkningar lämnas ej in för dessa, om uträkningar ändå lämnas in kommer de inte att beaktas vid bedömningen.

- **Uppgift 6 – 7**: Svar med **FULLSTÄNDIGA REDOVISNINGAR** ska lämnas in.
  - Använd endast skrivpapper som tillhandahålls i skrivsalen.
  - För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
  - Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan också ge poängavdrag!

- Tentamen kan maximalt ge 60 + 40 = 100 poäng och för godkänt resultat krävs minst 50.
- Betygsgränser:

  A: 90 – 100 p
  B: 80 – 89 p
  C: 70 – 79 p
  D: 60 – 69 p
  E: 50 – 59 p
  Fx: 40 – 49 p
  F: 0 – 40 p

  OBS! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

- Lösningsförslag läggs ut på Mondo kort efter tentamen.

LYCKA TILL!

Uppgift 1

I samband med de pågående löneförhandlingarna har du hjälpt till med en undersökning på företaget där du arbetar. Ni har bland annat samlat in månadslönen mätt i tusental kronor (tkr) för ett urval anställda. Uppgifterna ni har är:

42, 23, 50, 51, 27, 60, 33, 17, 22, 36, 31, 35

a) Nedan visas fyra olika boxplottar (lådagram) men endast ett avspeglar datamaterialet ovan. Vilket? (5p)

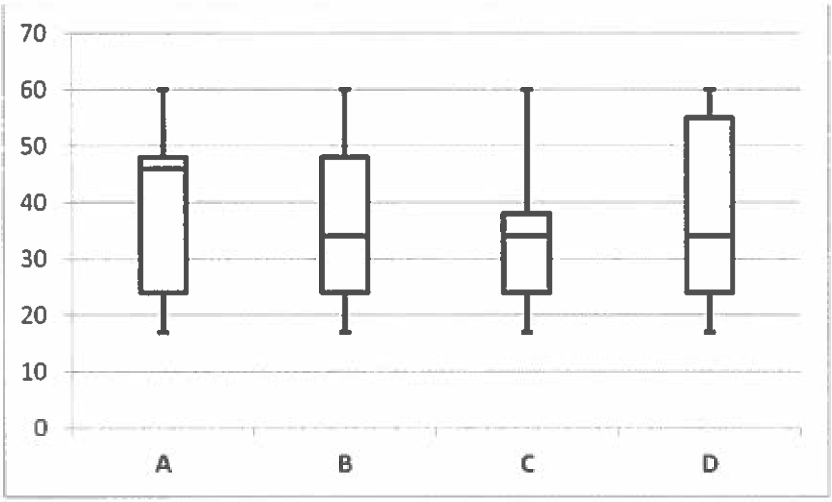
A. Boxplot A

B. Boxplot B

C. Boxplot C

D. Boxplot D

E. Inget av alternativen är rätt



b) Ange variationsvidden (*range*) och standardavvikelse för datamaterialet ovan. (5p)

A. $range = 24$ standardavvikelse = 12,5

B. $range = 24$ standardavvikelse = 13,0

C. $range = 43$ standardavvikelse = 12,5

D. $range = 43$ standardavvikelse = 13,0

E. $range = 43$ standardavvikelse = 10,0

Uppgift 2

Vid upprepade undersökningar visade det sig att 35 % av cyklisterna i Stockholm saknar fungerande cykelbelysning. När polisen i samband med en hastighetskontroll stoppade cyklister som cyklade över hastighetsbegräsningen, kontrollerades även cykelbelysningen. 10 % av alla cyklister cyklade för fort (och stoppades) och av dessa saknade 28 % fungerande belysning.

a) Hur stor andel av alla cyklister bedöms cykla för fort och sakna fungerade belysning. Är händelserna "cyklar för fort" och "saknar belysning" beroende eller oberoende? (4p)

A. 0,035 beroende händelser
B. 0,035 oberoende händelser
C. 0,010 beroende händelser
D. 0,028 beroende händelser
E. 0,028 oberoende händelser

b) Av alla cyklister som saknar fungerande belysning, hur stor andel bedöms hålla korrekt hastighet, dvs. inte cykla för fort? (5p)

A. 0,889
B. 0,206
C. 0,003
D. 0,415
E. 0,920

I tabellen nedan ges de simultana sannolikheterna för två slumpvariabler X och Y, dvs. sannolikheterna $P(x,y) = P(X = x \cap Y = y)$. Tabellen visar alltså den bivariata sannolikhetsfördelningen för (x,y). Utfallsrummet för X är som vi ser $S_X = \{0,1,2,3\}$ och för Y är utfallsrummet $S_Y = \{0,1\}$.

$P(x,y)$	$x = 0$	1	2	3	Summa
$y = 0$	0,06	0,30	0,18	0,06	0,60
1	0	0,24	0,16	0	0,40
Summa	0,06	0,54	0,34	0,06	1,00

c) Vad är kovariansen mellan X och Y och är de oberoende eller beroende slumpvariabler? TIPS: När du beräknar dubbelsumman behöver du bara addera två termer. (6p)

A. $Cov(X,Y) = -0,126$ X och Y är beroende
B. $Cov(X,Y) = 0,126$ X och Y är oberoende
C. $Cov(X,Y) = 0,126$ X och Y är beroende
D. $Cov(X,Y) = 0$ X och Y är oberoende
E. $Cov(X,Y) = 0$ X och Y är beroende

Uppgift 3

Två välkända läskedrycksmärken sägs smaka ungefär likadant och vid olika tester får man fram att 75 % av konsumenterna faktiskt inte kan skilja på dem i ett blint test. I ett underhållningsprogram på tv låter man en testpanel bestående av $n = 6$ slumpvist valda personer testa dryckerna i slutna rum och oberoende av varandra.

a) Vad är sannolikheten att minst 3 av de 6 testpersonerna kan identifiera dryckerna? (4p)

A. 0,169

B. 0,962

C. 0,534

D. 0,831

E. 0,038

Anta att du är anställd på ett stort företag med många anställda och att man precis har förhandlat fram nya löner. Den genomsnittliga lönehöjningen för alla anställda blev 2,5 % med en standardavvikelse på 1 % -enhet. Utöver detta fick samtliga anställda en bonus på 200 kronor i månaden. Din nuvarande lön är 32 000 kr så din nya lön Y kan beräknas som

$$Y = 32000 \cdot X + 200$$

där X är en slumpvariabel motsvarande den procentuella lönejusteringen. För att förenkla det hela antar du att X är normalfördelad, dvs. $X \sim N(1{,}025; 0{,}01^2)$.

b) Vad är sannolikheten att din nya lön är minst 33 500 kr som du blev lovad? (6p).

A. 0,941

B. 0,500

C. 0,059

D. 0,014

E. 0,000

OBS! Svarsalternativen i a) och b) har avrundats till 3 decimaler.

Uppgift 4

Den genomsnittliga tiden i minuter som det tar att tillverka en viss produkt vid två olika tillverkningsställen har undersökts. Vid fabrik X drogs ett *iid* stickprov av storlek $n_x = 400$ observationer och från fabrik Y drogs ett *iid* stickprov (oberoende av fabrik A) av storlek $n_y = 800$. Två 95 % konfidensintervall beräknades för μ_X (förväntad tillverkningstid vid fabrik X) respektive μ_Y (förväntad tillverkningstid vid fabrik Y):

95 % KI för μ_X: $(105, 128)$ 95 % KI för μ_Y: $(98, 108)$

a) Vilket av följande alternativ utgör en väntevärdesriktig skattning av *differensen* mellan μ_X och μ_Y, dvs. en skattning för $\mu_X - \mu_Y$? (4p)

A. 13,5
B. 20,0
C. 7,0
D. 24,0
E. 30,0

b) Anta att du ska beräkna ett 95 % konfidensintervall för differensskattningen i a). Vilket av följande utgör då felmarginalen? TIPS: Lite knepigare uppgift men beräkna först felmarginalerna för $\bar{x}$ och $\bar{y}$ och använd sedan formelsamlingen. (6p)

A. 6,3
B. 12,5
C. 24,5
D. 35,9
E. 80,4

c) Anta att $\bar{X}$ är ett stickprovsmedelvärde som kommer ett stickprov som består av n_x stycken oberoende och lika fördelade (*iid*) observationer X_i, $i = 1, \dots, n_x$. Anta att väntevärde och varians för X_i är μ_X respektive σ_X^2. Ange vilket av följande påståenden som är falskt. (5p)

A. $\bar{X}$ är en väntevärdesriktig (*unbiased*) skattning för μ_X.
B. Centrala gränsvärdessatsen behövs inte om observationerna X_i är normalfördelade.
C. $\bar{X}$ är en effektiv skattning för μ_X om den har en varians som är större än andra tänkbara skattningar för μ_X.
D. Om det gäller att alla X_i endast kan anta värdena 0 och 1 då gäller att fördelningen för $\bar{X}$ går mot en normalfördelning när n_x går mot oändligheten.
E. Felmarginalen går mot noll när n_x går mot oändligheten.

Uppgift 5

En marknadsundersökning genomfördes där man frågade slumpvist utvalda barn på mellanstadiet respektive lågstadiet om de hade mobiltelefon. Det visade sig att 72 av 100 barn på mellanstadiet hade egen mobiltelefon och 60 av 120 barn på lågstadiet hade egen mobiltelefon. Andel barn i populationen som går i mellanstadiet och har mobil betecknas med P_M och motsvarande andel barn i lågstadiet betecknas med P_L.

a) Ange vilken av följande slutsatser du kommer fram till om du testar $H_0: P_M - P_L = 0$ mot $H_1: P_M - P_L > 0$ på 1 % signifikansnivå. (6p)

A. $z_{obs} = 3{,}32$ och H_0 förkastas inte

B. $z_{obs} = 3{,}32$ och H_0 förkastas

C. $z_{obs} = 2{,}28$ och H_0 förkastas inte

D. $z_{obs} = 2{,}28$ och H_0 förkastas

E. $z_{obs} = 6{,}67$ och H_0 förkastas inte

b) Utifrån ditt resultat i a)-uppgiften, vilket av följande påståenden om p-värdet för z_{obs} är sant? (4p)

A. p-värdet ligger mellan 0,025 och 0,05.

B. p-värdet ligger mellan 0,01 och 0,025.

C. p-värdet ligger mellan 0,005 och 0,01.

D. p-värdet ligger mellan 0,0005 och 0,0010.

E. p-värdet ligger mellan 0,00025 och 0,0005.

Fullständig redovisning krävs för följande uppgifter.

Använd separata pappersark för uppgift 6 resp. uppgift 7.

För uppgift 7c används Svarsbilagan

Uppgift 6

Ett företag som importerar och distribuerar vitvaror är väl etablerad på en viss marknad (västkusten) men nu ska företaget expandera sin verksamhet till en ny marknad (östkusten). Man vet sedan tidigare hur kundernas preferenser på västkusten fördelas på fyra olika varumärken betecknade A-D nedan. Om det visar sig att öst skiljer sig från väst kommer man att anta en annan strategi inför lanseringen på östkusten. Man genomför en undersökning bland $n = 400$ potentiella kunder på östkusten och frågar vilket märke de föredrar. De observerade frekvenserna på östkusten redovisas nedan tillsammans med preferenserna (sannolikheterna) på västkusten:

Varumärke	A	B	C	D	
Andel (sannolikhet), västkust	0,20	0,35	0,30	0,15	Summa
Observerade frekvenser, östkust	102	121	120	57	400

a) Testa på 5 % signifikansnivå om östkusten kan antas följa samma fördelning över varumärken som man har på västkusten. Redovisa antaganden, hypoteser, testvariabel samt beslutsregel och kritisk gräns. OBS! Tänk på att stickprovet är taget endast på östkusten. (8p)

b) Redovisa sedan dina beräkningar, dra slutsatser och tolka resultatet i ord. (6p)

c) Förklara kortfattat begreppen Typ I och Typ II fel och hur dessa hänger ihop med testets signifikansnivå. Ingen lång utläggning krävs, cirka 3-5 meningar är allt som behövs. (6p).

Uppgift 7

En mäklare använder data från ett stickprov med $n = 10$ lägenheter (ettor) som nyligen sålts i en stadsdel för att undersöka sambandet mellan Y_i = försäljningspris (miljoner kronor, mkr) och lägenhetsyta X_i (kvadratmeter, m^2). En preliminär analys av stickprovets data gav följande resultat:

$$\sum y_i = 17{,}6 \qquad \sum x_i = 294$$

$$\sum y_i^2 = 26{,}02 \qquad \sum x_i^2 = 9\,060 \qquad \sum x_i y_i = 480{,}7$$

a) Ange den linjära regressionsmodellen med sedvanlig notation. Skatta intercept och lutningskoefficient för den linjära modellen med minsta kvadratmetoden. Skriv ner den skattade regressionslinjen. (6p)

b) Vad är tolkningen av den skattade lutningskoefficienten i den här situationen? Hur tolkar du interceptet och är det en meningsfull tolkning? (4p)

c) Beräkna förklaringsgraden om du vet att $\sum e_i^2 = 0{,}1652$. Tolka siffervärdet du får i ord. (4p)

d) Anta att du är intresserad av en speciell lägenhet i det här området. Storleken är 35 m^2. Beräkna en prognos och ett intervall (95 %, tvåsidigt) för slutpriset av denna lägenhet. Ge en tolkning av intervallet i ord. (6p)

STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2017

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2017-10-25

LÖSNINGSFÖRSLAG

Sista versionen / 2017-11-07 MC

Sammanfattning SVARSBILAGA Uppgifter 1-5
Utförliga beräkningar ges på efterföljande sidor

		A	B	C	D	E
Uppgift 1	a)	□	■	□	□	□
	b)	□	□	□	■	□
Uppgift 2	a)	□	□	□	■	□
	b)	□	□	□	□	■
	c)	□	□	□	□	■
Uppgift 3	a)	■	□	□	□	□
	b)	□	□	■	□	□
Uppgift 4	a)	■	□	□	□	□
	b)	□	■	□	□	□
	c)	□	□	■	□	□
Uppgift 5	a)	□	■	□	□	□
	b)	□	□	□	□	■

Uppgift 1

a) Rätt svar: **B**

Ordna datamaterialet i storleksordning. *Medianen* är mittpunkten som delar materialet i två lika stora delar, och definieras här som medelvärdet av 6:e och 7:e observationen. Här blir medianen 34 och då kan vi stryka boxplot A; återstår boxplot B-D som alla har samma median.

Sedan kan man bedöma ungefär var tredje kvartilen $Q3$ har för värde. Det ska vara den punkt som delar datamaterialet i 75 % nedanför och 25 % ovanför. Den punkten ska ligga mellan 9:e och 10:e observationen dvs. mellan 42 och 50. Den enda boxplot som har $Q3$ mellan dessa två värden är B.

Observera att Min, Max och $Q1$ är identiska i alla fyra boxplottar. Det är alltså med hjälp av medianen och Q3 som vi kan skilja på dem och koppla dem till det givna datamaterialet.

b) Rätt svar: **D**

Variationsvidden (*range*) definieras som $Max - Min$ och i detta fall blir det $60 - 17 = \mathbf{43}$.

Värdet 24 som fanns som alternativ är kvartilavståndet, $IQR = Q3 - Q1 = 24$, men det är inte rätt.

Standardavvikelsen får man t.ex. genom att beräkna

i	1	2	3	...	11	12	Summa
x_i	42	23	50	...	31	35	427
x_1^2	1764	529	2500	...	961	1225	17067

och

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n \cdot \bar{x}^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n \cdot \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2}{n-1}$$

$$= \frac{17067 - \frac{1}{12} \cdot 427^2}{11} = \frac{1872{,}92}{11} = 170{,}26$$

och

$$s_x = \sqrt{s_x^2} = \sqrt{170{,}26} \approx 13{,}0486 \approx \mathbf{13{,}0}$$

Alternativet 12,5 är standardavvikelsen för en *population*, dvs. när man delar med $N = 12$ istället för $n - 1 = 11$ men det är inte korrekt; det stod i inledningen till uppgiften att det handlade om ett urval. Alternativet 10,0 var påhittat och därmed fel.

Uppgift 2

Beteckna händelserna $B =$ "har belysning" och $\bar{B} =$ "saknar belysning" och $H =$ "håller rätt hastighet" och $\bar{H} =$ "cyklar för fort". Sannolikheter som är givna:

$$P(\bar{B}) = 0{,}35 \quad P(B) = 1 - 0{,}35 = 0{,}65 \qquad P(\bar{H}) = 0{,}10 \quad P(H) = 1 - 0{,}10 = 0{,}90$$

samt $P(\bar{B}|\bar{H}) = 0{,}28$ vilket ger

$$P(\bar{B} \cap \bar{H}) = [\text{multiplikationssatsen}] = P(\bar{B}|\bar{H}) \cdot P(\bar{H}) = 0{,}028$$

Sammanställ gärna allt i en tabell (Venndiagram) och räkna ut samtliga sannolikheter:

	B	$\bar{B}$	
H	0,578	0,322	0,900
$\bar{H}$	0,072	0,028	0,100
	0,650	0,350	1,000

a) Rätt svar: **D**

Sökt: andelen "cyklar för fort" <u>och</u> "saknar belysning", dvs. $P(\bar{B} \cap \bar{H}) = \mathbf{0,028}$. Det var redan framräknat ovan.

De är **beroende** vilket ges bland annat av att $P(\bar{B}|\bar{H}) = 0{,}28 \neq 0{,}35 = P(\bar{B})$.

b) Rätt svar: **E**

Sökt: andel "håller hastigheten" (H) bland de som "saknar belysning" ($\bar{B}$), dvs.

$$P(H|\bar{B}) = [\text{läs av från tabellen ovan}] = \frac{P(H \cap \bar{B})}{P(\bar{B})} = \frac{0{,}322}{0{,}350} = \mathbf{0,920}$$

c) Rätt svar: **E**

Beräkna först väntevärdena för X respektive Y och dubbelsumman som behövs i kovarians-formeln:

$$\mu_X = \sum xP(x) = 0 \cdot 0{,}06 + 1 \cdot 0{,}54 + 2 \cdot 0{,}34 + 3 \cdot 0{,}06 = 1{,}4$$

$$\mu_Y = \sum yP(y) = 0 \cdot 0{,}6 + 1 \cdot 0{,}4 = 0{,}4$$

$$\sum\sum xyP(x,y) = 1 \cdot 1 \cdot 0{,}24 + 2 \cdot 1 \cdot 0{,}16 = [\text{alla andra termer är} = 0] = 0{,}56$$

Kovariansen ges av

$$\sum\sum xyP(x,y) - \mu_X \cdot \mu_Y = 0{,}56 - 0{,}4 \cdot 1{,}4 = \mathbf{0}$$

Att X och Y är **beroende** ges t.ex. av att

$$P(X = 0 \cap Y = 1) = 0 \neq 0{,}024 = 0{,}06 \cdot 0{,}4 = P(X = 0) \cdot P(Y = 1)$$

Samtliga kombinationer av x och y ger liknande resultat men det räcker med att hitta ett enda där det skiljer sig för att X och Y ska vara beroende.

Uppgift 3

a) Rätt svar: **A**

Låt $P = 0{,}75$ beteckna sannolikheten att en slumpvist vald person inte kan identifiera dryckerna. Då är sannolikheten att en person kan identifiera dem lika med $1 - P = 0{,}25$.

Låt Y beteckna antalet av $n = 6$ som inte kan identifiera dryckerna. Då kan $X = 6 - Y$ få beteckna antalet som kan identifiera dem.

Sannolikheten P kan antas vara konstant, varje person väljs slumpmässigt och får ta testet oberoende av varandra. Då är $X \sim Bin(6; 0{,}25)$.

Sökt:

$$P(X > 3) = [\text{komplement}] = 1 - P(X \leq 2) = [\text{Tabell 7}]$$
$$= 1 - 0{,}83057 \approx \mathbf{0{,}169}$$

b) Rätt svar: **C**

Givet att $X \sim N(1{,}025; 0{,}01)$ och att $Y = 32\,000 \cdot X + 200$ vilket är en linjär transformation. Beräkna väntevärde och standardavvikelse för Y:

$$\mu_Y = 32\,000 \cdot \mu_X + 200 = 32\,000 \cdot 1{,}025 + 200 = 33\,000$$
$$\sigma_Y^2 = 32\,000^2 \cdot \sigma_X^2 = 32\,000^2 \cdot 0{,}01^2 = 102\,400$$
$$\sigma_Y = \sqrt{102\,400} = 320$$

Sökt:

$$P(Y > 33\,500) = [\text{standardisera}] = P\left(Z > \frac{33\,500 - 33\,000}{320}\right) = P(Z > 1{,}5625)$$
$$\approx P(Z > 1{,}56) = [\text{komplement}] = 1 - P(Z \leq 1{,}56) = [\text{Tabell 1}]$$
$$= 1 - 0{,}94062 \approx \mathbf{0{,}059}$$

Uppgift 4

Ett (två-sidigt) konfidensintevall beräknas som

$$\text{punktksattning} \pm \text{felmarginal}$$

Vi har två sinsemellan oberoende stickprov som var för sig består av *iid* observationer. Stickprovsstorlekarna är tillräckligt stora för att vi ska kunna använda CGS (observera att inget sas om respektive fabriks fördelning över tillverkningstider) och därmed använda $z_{\alpha/2}$. Det står inte heller något om varianserna var kända eller okända men det spelar inte heller någon roll; vi skattar σ_X^2 och σ_Y^2 med stickprovsvarianserna s_x^2 och s_y^2 och använder ändå $z_{\alpha/2}$ som en approximation vilket fungerar hyfsat om stickprovsstorlekarna är tillräckligt stora. I detta fall beräknas alltså konfidensintervallen enligt

$$\bar{x} \pm z_{\alpha/2}\frac{s_x}{\sqrt{n_x}} \qquad \text{respektive} \qquad \bar{y} \pm z_{\alpha/2}\frac{s_y}{\sqrt{n_y}}$$

a) Rätt svar: **A**

Eftersom intervallen är symmetriska ges punktskattningarna för μ_X och μ_Y av mittpunkterna i intervallen, dvs. $\bar{x} = 116{,}5$ respektive $\bar{y} = 103$. En skattning av differensen $\mu_X - \mu_Y$ ges av $\bar{x} - \bar{y} = \mathbf{13{,}5}$.

Denna skattning är väntevärdesriktig eftersom $\bar{X}$ och $\bar{Y}$ är väntevärdesriktiga skattningar av μ_X respektive μ_Y och $E(\bar{X} - \bar{Y}) = [\text{linj. komb.}] = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$.

b) Rätt svar: **B**

På motsvarande sätt kan vi beräkna felmarginalerna för $\bar{x}$ och $\bar{y}$ som *halva* intervallbredden, dvs. $(128 - 105)/2 = 11{,}5$ för μ_X och $(108 - 98)/2 = 5$ för μ_Y.

Om man delar dessa två tal med $z_{\alpha/2} = 1{,}96$ får vi standardavvikelsen (standardfelen) för $\bar{x}$ och $\bar{y}$, dvs.

$$\frac{s_x}{\sqrt{n_x}} = \frac{11{,}5}{1{,}96} = 5{,}8673 \qquad \text{respektive} \qquad \frac{s_y}{\sqrt{n_y}} = \frac{5}{1{,}96} = 2{,}5510$$

och

$$\frac{s_x^2}{n_x} = 5{,}8673^2 = 34{,}4258 \qquad \text{respektive} \qquad \frac{s_y^2}{n_y} = 2{,}5510^2 = 6{,}50771$$

Ett 95 % konfidensintervall för differensen $\mu_X - \mu_Y$ ges av (se formelblad)

$$\bar{x} - \bar{y} \pm 1{,}96 \cdot \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

och felmarginalen blir efter insättning

$$1{,}96 \cdot \sqrt{34{,}4258 + 6{,}50771} = 12{,}53994 \approx \mathbf{12{,}5}$$

c) Rätt svar: **C**

A. SANT: $\bar{X}$ är alltid en väntevärdesriktig (*unbiased*) skattning för μ_X (om μ_X existerar).

B. SANT: Om $X_i \sim N(\mu_X, \sigma_X^2)$ så följer automatiskt att $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$. CGS är det vi använder när X_i inte är $N(\mu_X, \sigma_X^2)$. SANT

C. **FALSKT**: En skattning för μ_X är effektivare om den har en lägre varians jämfört med en annan skattningsmetod.

D. SANT: Enligt CGS ska $\bar{X}$ gå mot en normalfördelning när n_x går mot oändligheten oavsett hur de enskilda X_i är fördelade, även om X_i endast kan anta värdena 0 och 1.

E. SANT: Felmarginalen beräknas $z_{\alpha/2} \cdot s_x/\sqrt{n_x}$ vilket minskar när n_x ökar.

Se även kurslitteraturen och föreläsningsanteckningar.

Uppgift 5

För att testa $H_0: P_M - P_L = 0$ mot $H_1: P_M - P_L > 0$ används testvariabeln (se formelsamlingen)

$$Z = \frac{\hat{p}_M - \hat{p}_L}{\sqrt{\hat{p}_0(1-\hat{p}_0)(\frac{1}{n_M}+\frac{1}{n_L})}} \quad \text{där} \quad \hat{p}_0 = \frac{n_M\hat{p}_M + n_L\hat{p}_L}{n_M + n_L}$$

Under givna förutsättningar och CGS är Z approximativt $N(0,1)$ och H_0 förkastas om $z_{obs} > z_{0,01} = 2{,}3263$.

a) Rätt svar: **B**

Insättning ger $\hat{p}_M = 72/100 = 0{,}72$ och $\hat{p}_L = 60/120 = 0{,}5$ samt $\hat{p}_0 = \frac{72+60}{100+120} = 0{,}6$ och vi får

$$z_{obs} = \frac{0{,}72 - 0{,}5}{\sqrt{0{,}6 \cdot 0{,}4 \cdot (\frac{1}{100} + \frac{1}{120})}} = 3{,}3166 \approx \mathbf{3{,}32} > 2{,}3263$$

och $\boldsymbol{H_0}$ **förkastas.**

b) Rätt svar: **E**

Använd Tabell 2 för att hitta att det observerade värdet på testvariabeln z_{obs} ligger mellan de två kritiska värden z_α som motsvarar $\alpha = 0{,}00025$ och $\alpha = 0{,}0005$ dvs.

$$(z_{0,0005} = 3{,}2905) < (z_{obs} = 3{,}32) < (z_{0,00025} = 3{,}4808)$$

p-värdet måste alltså ligga mellan 0,00025 och 0,0005.

Man kan också (enklare?) slå upp värdet i Tabell 1 i tabellsamlingen och avläsa

$$P(Z \leq 3{,}32) = 0{,}99955 \Rightarrow P(Z > 3{,}32) = 1 - 0{,}99955 = 0{,}00045 = p\text{-värdet}$$

Uppgift 6

a) Anpassningstest (*goodness-of-fit*) dvs. ett **χ^2-test** för att testa om fördelningen på östkusten är densamma som på västkusten (som är given/antagen och inte baserad på ett stickprov; det är alltså inte ett homogenitets-/oberoendetest som ska göras här).

Antagande (2): varje individ i stickprovet (på östkusten) är slumpmässigt dragen, de är oberoende av varandra och de har samma fördelning dvs. samma sannolikheter att välja A, B, C respektive D (dvs. *iid*). Vi kan konstatera att antalet kategorier är $K = 4$ och antalet observationer i stickprovet är $n = 400$.

Hypoteser (2):

H_0: $P_A = 0{,}20$; $P_B = 0{,}35$; $P_C = 0{,}30$; $P_D = 0{,}15$

H_0: inte samma fördelning som i H_0

där P_i betecknar sannolikheten att en slumpmässigt vald individ väljer varumärke i.

Testvariabel (2):

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$ (summeras över de 4 kategorierna A, B, C, D)

som är approximativt χ^2-fördelad med $K - 1 = 4 - 1 = 3$ frihetsgrader under H_0 och om de förväntade frekvenserna $E_i = nP_i > 5$ för varje kategori $i = A, B, C, D$.

Beslutsregel (2): förkasta H_0 om $\chi^2_{obs} > \chi^2_{3;0,05} = 7{,}815$ (enkelsidigt test).

b) Beräkningar ger (3):

i	A	B	C	D	Summa
P_i	0,2	0,35	0,3	0,15	1
$E_i = nP_i$	80	140	120	60	400
O_i	102	121	120	57	400
$O_i - E_i$	22	-19	0	-3	0
$(O_i - E_i)^2$	484	361	0	9	
$\boldsymbol{(O_i - E_i)^2/E_i}$	6,05	2,5786	0	0,15	8,7786

Notera att alla $E_i > 5$

Slutsats (3): Det observerade värdet på testvariabeln är $\boldsymbol{\chi^2_{obs} = 8{,}779} > 7{,}815$ och $\boldsymbol{H_0}$ **förkastas** på 5 % signifikansnivå. Det är statistiskt säkerställt på 5 % nivån att östkusten inte följer samma fördelning som på västkusten.

c) Typ I fel = att förkasta en sann nollhypotes.

Typ II fel = att inte förkasta en falsk nollhypotes.

Sannolikheten att förkasta en sann nollhypotes (Typ I fel) bestäms godtyckligt (ofta 1 eller 5 %) och kallas testets signifikansnivå α.

Efterfrågades inte men: sannolikheten att inte förkasta en falsk nollhypotes (Typ II fel) betecknas ofta med β och beror på vad som faktiskt är sant (och det vet man ju inte). Sannolikheten att förkasta en falsk nollhypotes (vilket man vill göra) kallas testets styrka och betecknas $1 - \beta$.

Uppgift 7

Beräkna först medelvärden, varianser och standaravvikelser för x och y samt kovariansen och eventuellt korrelationskoefficienten mellan x och y.

OBS! Summan $\sum y_i = 17,6$ som gavs var inte korrekt (skulle vara 16) vilket medför följdfel då man får $s_y^2 = -0,55067 < 0$ vilket givetvis är ett felaktigt värde för en varians. Följdfelen har markerats med rött nedan. Hänsyn till felet tas i samband med rättningen och betygsättningen. Observera dock att deluppgifterna a), b) och d) faktiskt kan beräknas utan det blir uppenbara konstigheter; endast i deluppgift c) får man ett identifierbart problem med den negativa variansen för y.

$\bar{y} = 1,76 \qquad s_y^2 = (\sum y_i^2 - n\bar{y}^2)/(n-1) = (26{,}02 - 10 \cdot 1,76^2)/9 = -0,55067$

$\bar{x} = 29{,}4 \qquad s_x^2 = (\sum x_i^2 - n\bar{y}^2)/(n-1) = (9060 - 10 \cdot 29{,}4^2)/9 = 46{,}267$

$$Cov(x,y) = (\sum x_i y_i - n\bar{x}\bar{y})/(n-1) = (480{,}7 - 10 \cdot 1,76 \cdot 29{,}4)/9 = -4,0822$$

a) Modellen skrivs $\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$.

Skattningen av lutningskoefficienten β_1:

$$b_1 = \frac{Cov(x,y)}{s_x^2} = \frac{-4{,}0822}{46{,}267} \approx \mathbf{-0,08823}$$

Skattningen av interceptet β_0:

$$b_0 = \bar{y} - b_1\bar{x} = 1{,}76 + 0{,}08823 \cdot 29{,}4 \approx \mathbf{4,354}$$

Den skattade modellen skrivs:

$$\hat{y}_i = b_0 + b_1 x_i = \mathbf{4,354 - 0,08823}\boldsymbol{x_i}$$

b) Tolkningen av lutningskoefficienten: om $x =$ lägenhetsytan ökar med 1 m^2, minskar priset i *genomsnitt* med 0,0882 mkr eller 88 200 kr.

Tolkningen av interceptet: det *genomsnittliga* priset för en lägenhet med ytan noll är 4,354 mkr. Denna tolkning är inte meningsfull eftersom lägenheter med ytan 0 m^2 är orimliga.

c) Förklaringsgraden kan beräknas på två sätt:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum e_i^2}{(n-1)s_y^2} = 1 - \frac{0{,}1652}{9 \cdot (-0,55067)} \approx \mathbf{1,0333} > 1$$

eller (och då måste man behärska komplexa tal vilket definitivt inte är ett krav på kursen!)

$$R^2 = r_{xy}^2 = \left(\frac{Cov(x,y)}{\sqrt{s_x^2}\sqrt{s_y^2}}\right)^2 = \left(\frac{-4{,}0822}{\sqrt{46{,}267}\sqrt{-0,55067}}\right)^2 = \mathbf{-0,6541}$$

vilket är två olika svar (!) och som båda är **orimliga då $0 \leq R^2 \leq 1$.**

Tolkningen är annars att ca $100 \cdot R^2$ % av variationen i pris (y) beror på att lägenheterna har olika stora ytor (x); det är lägenhetsytornas variation och som påverkar variationen i y.

d) En **prognos** (prediktion) för y_{n+1} när $x_{n+1} = \mathbf{35}$ ges av den skattade regressionslinjen:

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1} = 4{,}354 - 0{,}08823 \cdot 35 = 1{,}26595 \approx \mathbf{1{,}266\ mkr}$$

Ett prognos- eller prediktions**intervall** för y_{n+1} när $x_{n+1} = \mathbf{35}$ ges av (se formelsamlingen):

$$(b_0 + b_1 x_{n+1}) \pm t_{n-2,\alpha/2} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right)}$$

där $t_{n-2,\alpha/2} = t_{8;0{,}025} = [\text{Tabell 3}] = 2{,}306$.

Residualvariansen s_e^2 beräknas enligt (notera att $K = 1$ = antalet förklaringsvariabler):

$$s_e^2 = \frac{\sum e_i^2}{n - K - 1} = \frac{0{,}1652}{8} = 0{,}0206$$

Insättning ger:

$$1{,}266 \pm 2{,}306 \cdot \sqrt{0{,}02065\left(1 + \frac{1}{10} + \frac{(35 - 29{,}4)^2}{9 \cdot 46{,}267}\right)} = 1{,}266 \pm 0{,}3592549$$

eller avrundat $\mathbf{1{,}266 \pm 0{,}359}$ mkr. Intervallet kan också skrivas $\mathbf{(0{,}907\ ;\ 1{,}625)}$

Tolkningen är att man bedömer att 95 % av alla lägenheter av storleken 35 m^2 kommer att säljas till ett pris som ligger i intervallet ovan.

Notera att tolkningen inte är samma som för konfidensintervallet för $\mu_{Y|x_{n+1}}$ som istället är en beskrivning av osäkerheten kring skattningen av det *genomsnittliga* (förväntade) försäljningspriset när $x_{n+1} = 35$.

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 25/10/17

Sal: Laduvikssalen

Tenta: Statistik för ekonomer

Kurs: Grundläggande statistik för ekonomer

ANONYMKOD: SFE-0003

☒ 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	X	X			3
Lär.ant. 10	15	6	15	10	20	20			

POÄNG	BETYG	Lärarens sign.
96	A	Me

STOCKHOLMS UNIVERSITET
Statistiska institutionen

HT2017

SVARSBILAGA till Tentamen i Grundläggande statistik för ekonomer 2017-10-25

Skrivsal: Laduvikssalen

Anonymkod: SFE-0003 *(skriv tydligt!)*

Markera ditt svar med ett tydligt kryss (X) i rutorna nedan.
OBS! Endast ett kryss per uppgift. Har fler än ett svarsalternativ markerats ges noll poäng.

OBS! Om du efter att ha kontrollerat dina beräkningar ordentligt kommer fram till att svaret inte finns bland de angivna svarsalternativen, skriv ditt svar i mariginalen till höger.

		A	B	C	D	E	
Uppgift 1	a)		X				R
	b)				X		R
Uppgift 2	a)				X		R
	b)					X	R
	c)					X	R
Uppgift 3	a)					X	✓
	b)			X			R
Uppgift 4	a)	X					R
	b)		X				R
	c)			X			R
Uppgift 5	a)		X				R
	b)					X	R

56/60

SU, STATISTIK

Skrivsal: Laduvikssalen Anonymkod: SFE-0003 Blad nr: 1

Uppgift 6

H_0: Andelen A,B,C & D på västkusten = Andelen A,B,C,D östkusten

H_1: Andelen A,B,C & D på västkusten ≠ Andelen A,B,C,D östkusten ok

$\alpha = 0{,}05$

Antar att n=400 dragits slumpmässigt, är oberoende och lika fördelade. (gör ett goodness of fit test) ok

χ^2 fördelad R

a) kritisk gräns $= \chi^2_{k-1,\alpha} = \chi^2_{3,\,0{,}05} = 7{,}815$ R

Förkasta Ho om $\chi^2_{obs} > \chi^2_{3,0{,}05}$ R

$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ $E_i = nP_i$ Tumregel $nP_i > 5$

3 R 8

	A	B	C	D	
$nP_i = E_i$	80	140	120	60	
	102	121	120	57	Σ 400
$\frac{(O_i - E_i)^2}{E_i}$	6,05	≈2,578	0	0,15	Σ 8,778521429

$P(A) = 0{,}2$ $P(B) = 0{,}35$
$P(C) = 0{,}3$ $P(D) = 0{,}15$

$$\frac{(102-80)^2}{80} + \frac{(121-140)^2}{140} + \frac{(120-120)^2}{120} + \frac{(57-60)^2}{60}$$

$$= \chi^2_{obs} \approx 8{,}779$$ R

$8{,}779 > 7{,}815 \Rightarrow H_0$ Förkastas

förklara i ord R

8

b) Vi kan med 95% konfidens säga att andelen med preferens för de olika varumärkena på västkusten inte är samma (är skilda från) andelen med preferens för de olika varumärkena på östkusten. R 1/2

6

c)

	H_0 sann	H_0 falsk	
H_0 förkastas	Feltyp I α	rätt beslut $(1-\beta)$	styrka
H_0 förkastas ej	rätt beslut $(1-\alpha)$	Feltyp II β	

testets signifikansnivå α är risken för feltyp I. Desto lägre signifikansnivå α desto större blir risken för feltyp II (β), vilket gör testets styrka $(1-\beta)$ minskar och och omvänt. BRA

6

19

SU, STATISTIK

Skrivsal: Laduvikssalen Anonymkod: SFE-0063 Blad nr: 2

Uppgift 7

$n = 10$

a) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (modellen)

skattad: $\hat{Y} = b_0 + b_1 X_i$

$$\hat{Y} \approx 4{,}3540 - 0{,}0882 X_1$$

$$\bar{X} = \frac{\sum x_i}{n} = \frac{294}{10} = 29{,}4$$

$$\bar{Y} = \frac{\sum y_i}{n} = \frac{17{,}6}{10} = 1{,}76$$

$$b_1 = \frac{\text{cov}(x,y)}{S_x^2} \approx -0{,}0882$$

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{480{,}7 - 10 \cdot 29{,}4 \cdot 1{,}76}{9} \approx -4{,}0822$$

$$S_x^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{9060 - 10 \cdot 29{,}4^2}{9} \approx 46{,}2667$$

interceptet:

$$b_0 = \bar{Y} - b_1 \bar{X} = 1{,}76 + 0{,}0882 \cdot 29{,}4 \approx 4{,}3540$$

/6

b) tolkningen av lutningskoefficienten b_1 är att det förväntade försäljningspriset $\hat{Y}$ i genomsnitt minskar med 88 200kr per kvadratmeter (m^2) större dess yta är.

interceptet b_0 är det genomsnittliga försäljningspriset $\hat{Y}$ vid en yta satt till $0\,m^2$. Detta blir inte meningsfullt i verkligheten då alla lägenheter har en yta större än $0\,m^2$. Ger oss inget meningsfull information.

/4

c)
$$\sum e_i^2 = 0{,}1652 = SSE$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \left(\frac{0{,}1652}{-4{,}956}\right) = 1{,}03333 \approx 103{,}33\%$$

$$SST = (n-1)S_y^2 \approx 9 \cdot -0{,}550667 = -4{,}956$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{26{,}02 - 10 \cdot 1{,}76^2}{9} \approx -0{,}550667$$

variabeln b_1 förklarar över 100% av variansen i $\hat{Y}$. omöjligt svar, ju högre desto mer av variationen förklaras (kan ej vara över 100%) modellen har fel i siffrorna.

/4

d)
$$X_{n+1} = 35\,m^2 \qquad \alpha = 0{,}05 \qquad n = 10$$

prediktionsintervall

$$b_0 + b_1 X_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{S_e^2\left(1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{(n-1)S_x^2}\right)}$$

prognosvärde?

$$4{,}3540 + -0{,}0882 \cdot 35 \pm 2{,}306 \cdot \sqrt{0{,}02065 \cdot \left(1 + \frac{1}{10} + \frac{(35-29{,}4)^2}{9 \cdot 46{,}2667}\right)}$$

$$t_{8;0{,}025} = 2{,}306$$

$$S_e^2 = \frac{SSE}{n-4-1} = \frac{0{,}1652}{8} = 0{,}02065$$

$$\approx 1{,}2659 \pm 0{,}3592 \qquad (1{,}6251;\ 0{,}9067)$$

Slutpriset förväntas ligga mellan ca 1 625 100 kr och 906 700 kr

(~~exakt~~ exakt kr 906 649 kr till 1625147 kr)

I genomsnitt i 95% av stickproven.

/6

/20