STOCKHOLMS UNIVERSITET VT2017
Statistiska institutionen
Michael Carlson (examinator)
Bo Wallentin

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2017-03-20

Skrivtid:	kl. 16.00 - 21.00
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

a) Det är viktigt att man väljer diagramtyp efter variabeltyp och även efter variabelns skalnivå. Anta att du har en numerisk diskret variabel med många olika observerade siffervärden (heltal) där man har värden på kvotskala. Vilket av följande alternativ är minst lämpligast om du vill åskådliggöra den empiriska fördelningen grafiskt? (4p)

A. Stapeldiagram (pareto-ordnat)

B. Stapeldiagram (ordnat efter variabelvärde)

C. Ogive (kumulativ linjegraf)

D. Boxplot

E. Histogram, klassindelat

Man undersökte sambandet mellan antal timmar man studerade i snitt per dag under de två sista veckorna inför sluttentamen och hur många poäng man fick på provet. Ur en population bestående av flera tusen studerande samlade man in följande uppgifter:

Antal timmar/dag	3,5	2,4	4,0	5,0	1,1
Antal poäng på provet	88	76	92	85	60

b) Ange kovariansen och korrelationen mellan dessa två variabler i detta datamaterial. (6p)

A. Kovarians $= 13{,}24$ Korrelation $= 0{,}865$

B. Kovarians $= 16{,}55$ Korrelation $= 0{,}865$

C. Kovarians $= 13{,}24$ Korrelation $= 0{,}692$

D. Kovarians $= 16{,}55$ Korrelation $= 0{,}692$

E. Kovarians $= 13{,}24$ Korrelation $= -0{,}865$

Uppgift 2

Ett försäkringsbolag har beräknat att för 30 % av alla rapporterade bilolyckor var dåligt väder en bidragande orsak. Vidare har man beräknat att 20 % av alla olyckor medför kroppsskador. Slutligen, av alla de olyckor som medförde kroppsskador, så var 40 % dåligt väder en bidragande orsak till olyckan.

a) Vad är sannolikheten att en slumpvist utvald olycka medförde kroppsskada givet att den var orsakad av dåligt väder? (5p)

A. 0,080
B. 0,333
C. 0,150
D. 0,600
E. 0,267

b) Vad är sannolikheten att en slumpvist utvald olycka inte orsakades av vädret och inte medförde kroppsskada? (5p)

A. 0,560
B. 0,333
C. 0
D. 0,580
E. 0,940

Låt nu X vara en slumpvariabel med utfallsrummet $S_X = \{1,2,3,4,5\}$ och sannolikhetsfunktion

$$P(x) = \frac{2 \cdot (3-x)^2 + 1}{25}$$

c) Beräkna sannolikheten att X är ett udda tal. (5p)

A. 0,24
B. 0,36
C. 0,76
D. 0,64
E. 0,50

Uppgift 3

Sannolikheten att ett värdepapper ska öka i värde under en dag antas vara konstant lika med P. Definiera slumpvariabeln X som antalet dagar av n som värdepappret ökar i värde. Om man antar att dagarna är oberoende av varandra så är X binomialfördelad med parametrarna n och P. Och då är slumpvariabeln $Y = n - X =$ antal dagar av n som värdepappret minskar i värde också binomialfördelad men med parametrarna n och $1 - P$.

a) Anta att n är ett jämnt antal dagar. Vilket av nedanstående alternativ är korrekt oavsett värde på P? (5p)

A. $P(X < n) = P(Y > 1)$

B. $P(X = n - 3) = P(Y = 3 - n)$

C. $P(X = n/2) = P(Y = n/2)$

D. $P(X = n/2) = P(Y = 2n)$

E. $P(X \geq 0) = P(Y \leq 0)$

b) Anta nu att $n = 12$ och $P = 0{,}4$. Beräkna och ange sannolikheten $P(3 \leq X \leq 7)$. (5p)

A. 0,505

B. 0,476

C. 0,320

D. 0,717

E. 0,283

Dagliga avkastningar X_i för en placering antas vara slumpvariabler som är normalfördelade med väntevärde $0{,}018\,\%$ och varians $0{,}36\,\%$. Placeringens konstruktion är sådan att avkastningen kan vara negativ, dvs. att man förlorar pengar.

c) Vad är sannolikheten att avkastningen under en slumpmässigt vald dag är positiv? (5p)

A. 0,512

B. 0,051

C. 0,520

D. 0,488

E. 0,030

OBS! Svarsalternativen för uppgifterna b) – c) har avrundats till 3 decimaler.

Uppgift 4

Vid en undersökning studerades variabeln X och på basis av $n = 400$ observationer beräknades ett konfidensintervall för medelvärdet μ_X för X. Konfidensintervallet blev: $(8{,}95\,;\,11{,}05)$ och då utgick man ifrån stickprovsvariansen är 81.

a) Vilken konfidensnivå har använts? (6 poäng)

A. ca 80 %

B. ca 90 %

C. ca 95 %

D. ca 98 %

E. ca 99 %

b) Vid beräkningen av konfidensintervallet i uppgift a) måste fyra av nedanstående fem förutsättningar vara uppfyllda för att konfidensintervallet skall gälla åtminstone approximativt. Ange vilken av förutsättningarna nedan som inte behöver vara uppfylld i detta fall. (4p)

A. X måste vara normalfördelad

B. Observationerna ska vara slumpmässiga observationer på variabeln X

C. Stickprovsmedelvärdet ska vara approximativt normalfördelad

D. Stickprovet måste vara tillräckligt stort för att kunna approximera (skatta) variansen för X med stickprovsvariansen och ändå använda Z-variabeln

E. Observationerna ska vara oberoende

Uppgift 5

Barnafödslar sägs förekomma oftare nattetid men stämmer det? I en engelsk studie från 1953[1] fann man att 29 % av totalt $n = 400$ födslar skedde kl. 0-6; 26 % skedde kl. 6-12; 22 % procent kl. 12-18 samt 23 % kl. 18-24. Sammanfattat i tabellform observerade man följande:

	kl. 0-6	kl. 6-12	kl. 12-18	kl. 18-24
Observerad andel födslar	0,29	0,26	0,22	0,23
Observerat antal födslar	116	104	88	92
Lika andel födslar	0,25	0,25	0,25	0,25

Man vill testa på 5 % signifikansnivå om sannolikheterna för födsel är lika stora i samtliga fyra 6-timmarsperioder.

a) Vilket test nedan är det rätta i detta fall? (4p)

A. Test för differenser av andelar – approximativt normalfördelad (z-test)

B. Test för differenser av andelar – t-fördelad (t-test) med 3 frihetsgrader

C. Homogenitetstest (lika fördelning) – χ^2-fördelad med 3 frihetsgrader

D. Anpassningstest (*goodness-of-fit*) – χ^2-fördelad med 2 frihetsgrader

E. Anpassningstest (*goodness-of-fit*) – χ^2-fördelad med 3 frihetsgrader

b) Vilken av slutsatserna nedan är rätt? (6p)

A. Observerat värde $= 4{,}80 <$ kritiskt värde $= 7{,}815 \Rightarrow H_0$ förkastas

B. Observerat värde $= 4{,}80 <$ kritiskt värde $= 7{,}815 \Rightarrow H_0$ förkastas inte

C. Observerat värde $= 4{,}80 <$ kritiskt värde $= 5{,}991 \Rightarrow H_0$ förkastas inte

D. Observerat värde $= 4{,}69 <$ kritiskt värde $= 5{,}991 \Rightarrow H_0$ förkastas inte

E. Observerat värde $= 4{,}69 >$ kritiskt värde $= 1{,}96 \Rightarrow H_0$ förkastas

[1] Charles, E. (1953). The Hour of Birth: A Study of the Distribution of Times of Onset of Labour and of Delivery throughout the 24-Hour Period. *British Journal of Preventive and Social Medicine*, **7**, No. 2, pp. 43-59.

Fullständig redovisning krävs för följande uppgifter.

Använd separata pappersark för uppgift 6 resp. uppgift 7.

Uppgift 6

Man genomför undersökningar i två olika länder bland annat för att ta reda på hur man ser på den ekonomiska utvecklingen under den närmaste framtiden. Två stickprov av småföretag, ett från respektive land, dras och på frågan om man trodde man det skulle bli en uppgång för det egna företaget får man följande resultat: i land A svarar $y_A = 520$ av $n_A = 1300$ ”Ja” på frågan och i land B svarar $y_B = 385$ av $n_B = 1100$ ”Ja” på samma fråga.

a) Beräkna ett 90 % konfidensintervall för andelen Ja-svar i land A. Ange tydligt förutsättningar och antaganden som krävs, formel, formel med insatta värden. Tolka sedan resultatet. (10p)

b) Beräkna ett 90 % konfidensintervall för skillnaden i andelen Ja-svar mellan länderna. Utgå ifrån samma förutsättningar och antaganden som i a) men ange eventuella ytterliga förutsättningar och antaganden som krävs. Ange formel, formel med insatta värden och tolka sedan resultatet. (10p)

Uppgift 7

Man vill analysera sambandet mellan begärt pris och slutgiltigt pris (mkr, miljontals kr) vid försäljningar av bostadsrättslägenheter med hjälp av en enkel linjär regressionsmodell. Ett datamaterial för tio försäljningar i Uppsala med avslut andra halvan av november 2016 presenteras i bilagan på följande sida (källa: Hemnet).

Utgå ifrån datamaterialet där flera delberäkningar redan är gjorda. Du ska med en enkel linjär regressionsmodell förklara variabeln $Y =$ slutpris med variabeln $X =$ begärt pris. En datorutskrift från Excel för den skattade modellen finns i bilagan på följande sida men flera av uppgifterna har tappats bort och måste räknas om (av dig!).

a) Beräkna de två sista residualerna som har tappats bort och beräkna sedan residualvariansen. (5p)

b) Beräkna förklaringsgraden och tolka resultatet. (5p)

c) Skatta med 95 % konfidens det förväntade slutpriset då begärt pris är 3 000 000 dvs. skatta $\mu_{\text{slutpris}|\text{begärt pris}=3000000}$. Ange vilka förutsättningar och antaganden som krävs, vilken formel du använder, formel med insatta värden, samt beräkningar. Tolka slutligen resultatet. (10p)

BILAGA till Uppgift 7

Obs	Slutpris y (mkr)	Begärt pris x (mkr)	Residualer e
1	2,025	1,895	-0,0150
2	2,800	2,695	-0,1010
3	1,500	1,495	-0,1095
4	2,060	1,595	0,3428
5	2,710	2,495	0,0243
6	2,100	1,895	0,0600
7	1,250	1,300	-0,1497
8	1,295	1,295	-0,0993
9	2,335	2,195	[redacted]
10	1,900	1,695	[redacted]

$\sum y = 19{,}975$

$\sum y^2 = 42{,}4900$

$\sum (y - \bar{y})^2 = 2{,}5900$

$\sum x = 18{,}555$

$\sum x^2 = 36{,}5072$

$\sum (x - \bar{x})^2 = 2{,}0784$

$\sum xy = 39{,}3004$

Modell 1:

UTDATASAMMANFATTNING

Regressionsstatistik	
Multipel-R	[redacted]
R-kvadrat	[redacted]
Justerad R-kvadrat	[redacted]
Standardfel	[redacted]
Observationer	10

(standardavvikelsen för residualerna)

ANOVA

	f.g.	*KvS* (**SS**)	*MKv* (**MS**)	*F*	*p-värde*
Regression (**R**)	1	[redacted]	2,4072	[redacted]	6,99E-06
Residual (**E**)	8	[redacted]	[redacted]		
Totalt (**T**)	9	[redacted]			

	Koefficient	*Standardfel*	*t-kvot*	*p-värde*
Konstant	0,0006 kr	0,2004	[redacted]	0,9975
1000 mil	1,0762 kr	[redacted]	[redacted]	6,99E-06

STOCKHOLMS UNIVERSITET VT2017
Statistiska institutionen
Michael Carlson (examinator)

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2017-03-20

LÖSNINGSFÖRSLAG

Ny version, med reservation för tryck- och slarvfel / 2017-03-24 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)	Rätt svar saknades bland alternativen, räkna och ange svar: **0,859**				
	c)	■	□	□	□	□
Uppgift 4	a)	□	□	□	■	□
	b)	■	□	□	□	□
Uppgift 5	a)	□	□	□	□	■
	b)	□	■	□	□	□

Uppgift 1

a) Rätt svar: **A**

Ett **pareto-ordnat stapeldiagram** börjar från vänster med den kategori/variabelvärde som har störst frekvens, sedan näst störst frekvens, näst-näst störst osv. ner till kategorin/variabelvärdet som har lägst frekvens till höger. Det är inte lämpligt att göra på detta sätt för en numerisk variabel där utfallsrummet består av siffror som kan ordnas, inte om man vill få en uppfattning om fördelningen över utfallsrummet. Pareto-ordningen medför att siffervärdena hamnar huller om buller (även om det går att läsa av frekvensen för enskilda värden).

Stapeldiagram för numeriska variabler ska helst ordnas det från lägst observerat siffervärdet till det högsta; det riskerar att bli ganska många staplar men det fungerar. En ogive linjegraf visar den kumulativa frekvensfördelningen och är ok. En boxplot är också ok eftersom medianer och kvartiler och max-min värden är helt tolkningsbara när det är numeriska variabler. Histogram bygger på klassindelat material och det fungerar eftersom det är väldigt många observerade siffervärden, det blir som en förenklad version av stapeldiagrammet med färre staplar (det diskreta utfallsrummet approximeras med en kontinuerlig).

b) Rätt svar: **B**

Ur en population bestående av flera tusen studerande samlade man in uppgifter motsvarande fem observationer; detta ska tokas som att det handlar om ett stickprov.

	1	2	3	4	5	Summa
Timmar/dag, x_i	3,5	2,4	4	5	1,1	16
$(x_i - \bar{x})$	0,3	-0,8	0,8	1,8	-2,1	
$(x_i - \bar{x})^2$	0,09	0,64	0,64	3,24	4,41	9,02
Poäng på provet, y_i	88	76	92	85	60	401
$(y_i - \bar{y})$	7,8	-4,2	11,8	4,8	-20,2	
$(y_i - \bar{y})^2$	60,84	17,64	139,24	23,04	408,04	648,8
$(x_i - \bar{x})(y_i - \bar{y})$	2,34	3,36	9,44	8,64	42,42	66,2

Stickprovsmedelvärden: $\bar{x} = \frac{16}{5} = 3{,}2 \qquad \bar{y} = \frac{401}{5} = 80{,}2$

Stickprovsvarianser: $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{9{,}02}{4} = 2{,}255$

$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{648{,}8}{4} = 162{,}2$

Stickprovskovarians: $s_{xy} = Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{66{,}2}{4} = \mathbf{16{,}55}$

Stickprovskorrelation: $r_{xy} = Corr(x,y) = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{16{,}55}{\sqrt{2{,}255 \cdot 162{,}2}} = \mathbf{0{,}865}$

Uppgift 2

a) Rätt svar: **E**

Beteckna händelsen ”olycka orsakad av vädret” med V och ”kroppsskada” med K.

Givet: $P(V) = 0{,}30 \quad P(K) = 0{,}20 \quad P(V|K) = 0{,}40$

Sökt: $$P(K|V) = \frac{P(K \cap V)}{P(V)} = \text{[muliplikationssatsen]} = \frac{P(V|K) \cdot P(K)}{P(V)}$$

$$= \frac{0{,}40 \cdot 0{,}20}{0{,}30} = \frac{8}{30} = \mathbf{0,267}$$

b) Rätt svar: **D**

Sannolikheten att en slumpvist vald lycka ”inte orsakas av vädret och inte medförde kroppsskada” $= P(\bar{V} \cap \bar{K})$. Detta är motsatsen dvs. komplementet till ”orsakas av vädret eller medförde kroppsskada eller både och” $= P(V \cup K)$. Man kan rita ett Venndiagam också och ”se” att unionen $V \cup K$ omfattar de tre rutor/händelser som inte är $\bar{V} \cap \bar{K}$:

	K	ej K
V	$V \cup K$	
ej V		$\bar{V} \cap \bar{K}$

$$P(V \cup K) = \text{[additionssatsen]} = P(V) + P(K) - P(V \cap K)$$

$$= (V) + P(K) - P(V|K) \cdot P(K) = 0{,}30 + 0{,}20 - 0{,}40 \cdot 0{,}20 = 0{,}42$$

Sökt: $P(\bar{V} \cap \bar{K}) = \text{[komplementsatsen]} = 1 - P(V \cup K) = 1 - 0{,}42 = \mathbf{0,58}$

c) Rätt svar: **C**

Beräkna sannolikheterna för de tre udda och disjunkta utfallen $\{1,3,5\}$ för X och summera:

$$P(X \text{ är } udda) = P(X = 1 \cup X = 3 \cup X = 5) = P(X = 1) + P(X = 3) + P(X = 5)$$

$$= \frac{2 \cdot (3-1)^2 + 1}{25} + \frac{2 \cdot (3-3)^2 + 1}{25} + \frac{2 \cdot (3-5)^2 + 1}{25}$$

$$= \frac{9}{25} + \frac{1}{25} + \frac{9}{25} = \frac{19}{25} = \mathbf{0,76}$$

Uppgift 3

a) Rätt svar: **C**

Givet: $X \sim Bin(n, P)$ och $Y = n - X \sim Bin(,1 - P)$. Utveckla de olika fallen:

A. $P(X < n) = P(n - X > n - n) = P(Y > 0) \neq P(Y > 1)$ FEL

B. $P(X = n - 3) = P(n - X = n - n + 3) = P(Y = 3) \neq P(Y = 3 - n)$ FEL

C. $P(X = n/2) = P(n - X = n - n/2) = P(Y = n(1 - 1/2)) = P(Y = n/2)$ **RÄTT**

D. $P(X = n/2) = P(n - X = n - n/2) = P(Y = n/2) \neq P(Y = 2n)$ FEL

E. $P(X \geq 0) = P(n - X \leq n - 0)P = P(Y \leq n) = 1 \neq P(Y \leq 0)$ FEL

Kom ihåg att om det är en olikhet ($<, >, \leq eller\ \geq$) så måste man vända på den när man byter från X till $Y = n - X$.

b) Rätt svar: **Saknades bland svarsalternativen, rättelseblad medskickad**

Anta nu att $X \sim Bin(12; 0{,}4)$. Beräkna och ange sannolikheten $P(3 \leq X \leq 7)$

Sökt: $P(3 \leq X \leq 7) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$

$= $ [enl. formel]

$= \binom{12}{3} 0{,}4^3 0{,}6^9 + \binom{12}{4} 0{,}4^3 0{,}6^9 + \cdots + \binom{12}{7} 0{,}4^7 0{,}6^5$

$= 0{,}14189 + 0{,}21284 + 0{,}22703 + 0{,}17658 + 0{,}10090 = 0{,}85925 \approx \mathbf{0{,}859}$

Alternativt använd tabellen:

$P(3 \leq X \leq 7) = P(X \leq 7) - P(X \leq 2) =$ [enl. Tabell 7]

$= 0{,}94269 - 0{,}08344 = 0{,}85925 \approx \mathbf{0{,}859}$

c) Rätt svar: **A**

Givet att avkastningen $= X \sim N(0{,}018; 0{,}36)$.

Sökt: $P(X \text{ är positiv}) = P(X > 0) =$ [standardisera] $= P\left(Z > \dfrac{0 - 0{,}018}{\sqrt{0{,}36}}\right)$

$= P\left(Z > -\dfrac{0{,}018}{0{,}6}\right) = P(Z > -0{,}03) =$ [utnyttja symmetrin, rita!]

$= P(Z \leq 0{,}03) =$ [enl. Tabell 1] $= 0{,}51197 \approx \mathbf{0{,}512}$

Uppgift 4

a) Rätt svar: **D**

Ett 95 % KI för väntevärdet μ ges under dessa förhållanden (stort stickprov, $n = 400$) av

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

Undre (LCL) och övre (UCL) gräns är $8{,}95$ respektive $11{,}05$ och längden på hela intervallet är $UCL - LCL = 11{,}05 - 8{,}95 = 2{,}10$ vilket är detsamma som två gånger felmarginalen:

$$UCL - LCL = (\bar{x} \pm z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}) - (\bar{x} \pm z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}) = 2 \cdot z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} = 2{,}10$$

och därmed att

$$z_{\alpha/2} = \frac{2{,}10}{2} \cdot \frac{\sqrt{n}}{s_x} = 1{,}05 \cdot \frac{20}{9} = 2{,}33333 \approx 2{,}33$$

Enligt definitionen på konfidensgrad gäller att

$$\alpha/2 = P(Z > z_{\alpha/2}) = P(Z > 2{,}33) = 1 - P(Z \leq 2{,}33) = \text{[enl. Tabell 1]}$$

$$= 1 - 0{,}99010 = 0{,}0099 \Rightarrow \alpha \approx 0{,}02$$

Konfidensgraden är alltså $1 - \alpha \approx \mathbf{0{,}98}$

Not. Om man hade antagit att X är normalfördelad och sedan tagit t-värdet med $n - 1 = 399$ frihetsgrader (vilket är ungefär samma som 400 frihetsgrader vilket man hittar i tabellen) istället för Z-värdet, så hade man likafullt landat i ca 98 % konfidensgrad. Men det är lite svårare.

b) Rätt svar: **A**

Att observationerna ska vara ska vara slumpmässiga observationer på variabeln X är detsamma som att säga att de kommer från samma fördelning och detta är en förutsättning.

Man antar dessutom att de är oberoende av varandra annars får man en mängd kovarianstermer att ta hand om och stickprovsvariansen blir felskattad (*biased*).

För att kunna approximera (skatta) den okända variansen för X med stickprovsvariansen och ändå använda Z-variabeln krävs att man har ett ganska stort stickprov men $n = 400$ räcker ofta mer än väl.

Stickprovsmedelvärdet blir approximativt normalfördelad bara man har tillräckligt många observationer tack vare centrala gränsvärdessatsen(CGS), oavsett fördelningen för X.

Av det sista antagandet att $n = 400$ räcker för CGS följer alltså att ***X* inte behöver vara normalfördelad**, alternativ **A** är alltså inte ett krav som behöver vara uppfyllt i detta fall. CGS säger att fördelningen $\bar{X}$ går mot en normalfördelning oavsett hur X'n är fördelade.

Not. Angående CGS så finns det faktiskt ett krav på fördelningen för X och det är att väntevärde μ och varians σ^2 måste existera, de måste gå att härleda (beräkna) dem matematiskt. Detta är överkurs men det finns faktiskt fördelningar som saknar både väntevärde och varians, t.ex. t-fördelningen med endast 1 frihetsgrad.

Uppgift 5

a) Rätt svar: **E**

Anpassningstest, ett $\boldsymbol{\chi^2}$**-test**. Man vill testa om de fyra kategorierna/perioderna skiljer sig åt med avseende på hur sannolikt det är observera en födsel just då. Om de inte skiljer sig åt betyder det att sannolikheterna är lika dvs. $1/K = 0{,}25$.

Man utgår ifrån hypoteserna

H_0: alla sannolikheter lika, $P_1 = P_2 = P_3 = P_4 = 0{,}25$

H_1: minst en sannolikhet P_j skiljer sig från de övriga

Man har $K = 4$ kategorier/perioder alltså är testvariabeln χ^2-fördelad med $(K-1) = \mathbf{3}$ **frihetgrader**.

Övriga svarsalternativ är antingen fel (D, fel antal frihetsgrader) eller totalt fel (A-C): C är inte rätt eftersom man bara har en kategorisk variabel med fyra kategorier/perioder, homogenitetstest (oberoendetest) utgår ifrån två kategoriska variabler. A kan användas om man har två populationer och en ”Ja-Nej” variabel (inget av det stämmer) och B är inte lämpligt eftersom t-test används för normalfördelade variabler (vilket vi inte heller har).

b) Rätt svar: **B**

Testvariabel är $\chi^2 = \sum(O_i - E_i)^2/E_i$ där $E_i = nP_i = 400 \cdot 0{,}25$ under H_0.

Lite räkning krävs men man kommer ganska enkelt fram till att $\chi^2_{obs} = \mathbf{6{,}271}$.

Period	1	2	3	4	
P_i	0,25	0,25	0,25	0,25	
$E_i = nP_i$	100	100	100	100	
O_i	116	104	88	92	
$(O_i - E_i)$	16	4	-12	-8	Summa
$(O_i - E_i)^2$	256	16	144	64	480

$$\chi^2_{\text{obs}} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{256 + 16 + 144 + 64}{100} = \frac{480}{100} = \mathbf{4{,}80}$$

Beslutsregeln är förkasta H_0 om $\chi^2_{obs} = \mathbf{4{,}80} < \chi^2_{3;0{,}05} = \mathbf{7{,}815}$ alltså **förkastas inte nollhypotesen**.

Egenskaper för $\boldsymbol{\chi^2}$**-test**:

- Ett observerat värde på testvariabeln nära noll betyder att det sammantaget är små avvikelser, dvs. bra anpassning och starkt stöd för H_0; man förkastar inte H_0.
- Ett stort observerat värde på testvariabeln uppstår när det är stora avvikelser vilket betyder att det är en dålig anpassning och svagt stöd för H_0; man förkastar H_0.
- Obs! χ^2-test är alltså enkelsidiga test.

Uppgift 6

a) Låt P_A beteckna den okända andelen företag i land A som kan tänkas svara Ja på frågan. Låt Y_A beteckna slumvariabeln som är antalet som svarar Ja i ett stickprov av storlek $n_A = 1300$.

Antaganden: observationerna är **oberoende** av varandra och sannolikheten att slumpmässigt dra ett företag som svarar Ja är **konstant lika med P_A** (iid). Om detta gäller så kan Y_A antas vara fördelad enligt $\boldsymbol{Bin(n_A, P_A)}$. Då n_A är tillräckligt stort ($n > 30$) kan binomialfördelningen enligt **CGS** approximeras med en normalfördelning.

$\hat{p}_A = y_A/n_a$ där y_A är det observerade antalet Ja-svar är en väntevärdesriktig skattning för den sanna andelen P_A. Ett 90 % konfidensintervall för P_A ges av

$$\hat{p}_A \pm z_{0,05} \cdot \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A}}$$

Z-värdet $z_{0,05} = z_{\alpha/2} = $ [enl. Tabell 2] $= 1{,}6449$ används då $\alpha = 0{,}10$. Insättning ger

$$\frac{520}{1300} \pm 1{,}6449 \cdot \sqrt{\frac{\left(\frac{520}{1300}\right)\left(1-\frac{520}{1300}\right)}{1300}} = \mathbf{0{,}400 \pm 0{,}02235}$$

eller avrundat $(0{,}378;\, 0{,}422)$ eller 37,8 – 42,2 % -enheter.

Med 90 % konfidens (ej sannolikhet) ligger den sanna andelen i konfidensintervallet. Tolkningen av konfidensgraden är att vid upprepade stickprovsdragningar kommer det sanna värdet P_A ligga i 90 % av intervallen, allt annat lika.

b) Låt P_B, n_B , Y_B och y_B beteckna motsvarande storheter för land B.

Antaganden: motsvarande antaganden som i a) gäller för B, dvs. $Y_B \sim Bin(n_A, P_A)$. Dessutom antas att de två **stickproven är sinsemellan oberoende**.

Differensen $\hat{p}_A - \hat{p}_B$ är en väntevärdesriktig skattning av den sanna differensen $P_A - P_B$ och ett 90 % konfidensintervall för differensen ges av

$$\hat{p}_A - \hat{p}_B \pm z_{0,05} \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

Insättning ger

$$\frac{520}{1300} - \frac{385}{1100} \pm 1{,}6449 \cdot \sqrt{\frac{\frac{520}{1300}\left(1-\frac{520}{1300}\right)}{1300} + \frac{\frac{385}{1100}\left(1-\frac{385}{1100}\right)}{1100}} = \mathbf{0{,}0500 \pm 0{,}03254}$$

eller $(0{,}0175;\, 0{,}0825)$ eller 1,75 – 8,25 % -enheter.

Med 90 % konfidens (ej sannolikhet) ligger den sanna differensen i konfidensintervallet. Tolkningen av konfidensgraden är att vid upprepade stickprovsdragningar kommer det sanna värdet $P_A - P_B$ ligga i 90 % av intervallen, allt annat lika.

Uppgift 7

a) Från Excel-utskriften har man den skattade modellen

$$\hat{y}_i = b_0 + b_1 x_i = 0{,}0006 + 1{,}0762 x_i$$

Insättning av x- och y-värdena för de två sista observationerna ger residualerna som saknas:

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

$$e_9 = 2{,}335 - 0{,}0006 - 1{,}0762 \cdot 2{,}195 = \mathbf{-0{,}0279}$$

$$e_{10} = 1{,}900 - 0{,}0006 - 1{,}0762 \cdot 1{,}695 = \mathbf{0{,}0752}$$

Kvadrera samtliga residualer och summera dem för att få residualkvadratsumman (SSE) och dela sedan med antalet frihetsgrader ($n - K - 1$) för att få residualvariansen ($s_e^2 = MSE$):

$$s_e^2 = \frac{SSE}{n-K-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-K-1} = [n = 10, K = 1] = \frac{0{,}182847}{8} = 0{,}022856 \approx \mathbf{0{,}0229}$$

b) Vi behöver SST samt SSR eller SSE:

$$SST = \sum (y - \bar{y})^2 = \text{[avläst från bilagan]} = 2{,}59$$

$$SSE = \text{[från a)-uppgiften]} = 0{,}182847$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{0{,}182847}{2{,}59} = 0{,}920403 = \mathbf{0{,}9294}$$

Tolkning: 92,9 % av variationen i Y kan förklaras av X vilket kan anses som riktigt högt med tanke på att det endast är en förklaringsvariabel.

c) Sambandet mellan X_i och Y_i ska vara **linjärt**, dvs. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Feltermerna ε_i ska vara **oberoende** av varandra och oberoende av förklaringsvariabeln X_i. De ska vara **normalfördelade** $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ där **variansen är konstant** (homoskedasticitet).

Ett 95 % konfidensinterval för det betingade medelvärdet $\mu_{Y|X_{n+1}=3 \text{ mkr}}$ ges av

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

Vi behöver $\bar{x} = \sum x / n = 18{,}555/10 = 1{,}8555$ och noterar att $(n-1)s_x^2 = \sum (x - \bar{x})^2 =$ [avläst från bilagan] $= 2{,}0784$. Från Tabell 3 får vi $t_{n-2,\alpha/2} = t_{8;0{,}025} = 2{,}306$.

Insättning ger

$$0{,}0006 + 1{,}0762 \cdot 3 \pm 2{,}306 \cdot \sqrt{0{,}022856 \cdot \left(\frac{1}{10} + \frac{(3 - 1{,}8555)^2}{2{,}0784} \right)} = \mathbf{3{,}2292 \pm 0{,}29791}$$

eller avrundat $3{,}229 \pm 0{,}298$ mkr eller $(2{,}931; 3{,}527)$.

Med 95 % konfidens (ej sannolikhet) ligger det betingade genomsnittliga värdet för Y givet $X = 3$ i konfidensintervallet. Då $X = 3$ mkr ligger utanför det observerade området för X (1,295 – 2,695 mkr) ska man vara försiktig med tolkningen då det är en extrapolering av den skattade modellen.