STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)
Bo Wallentin

HT2016

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2016-12-01

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

En kartläggning av studentpopulationen vid Stockholms universitet (SU) syftade till att studera rekryteringen till universitetet för olika grupper (SU-rapport, nov 2010, www.su.se/foralla). Man samlade bl.a. in uppgifter om nyregistrerade studenters **kön**, **ålder**, **födelseregion** (svenskfödd, eller utlandsfödd fördelad på 9 olika regioner) samt **föräldrarnas utbildningsnivå** (3 nivåer: förgymnasial, gymnasial och eftergymnasial). Vidare samlade man in uppgifter om **gymnasiebakgrund**, där man för olika kommuner Stockholms län och län för övriga Sverige, redovisar andelen gymnasieelever som påbörjar studier vid SU; redovisningen sker per kommun i. Man redovisade även delar av materialet efter vilken **fakultet** vid SU där studenterna började.

a) Vilket av följande alternativ är ett korrekt påstående? (5p)

A. Utbildningsnivå är en kategorisk variabel på nominalskala

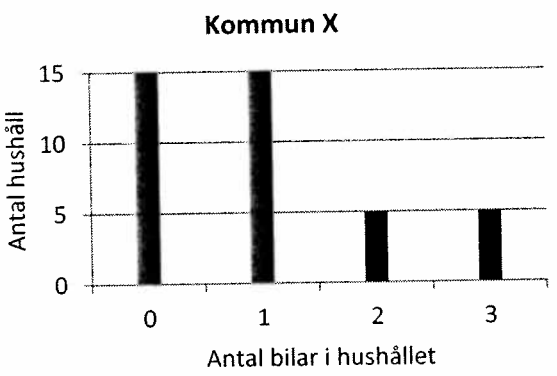
B. Utbildningsnivå är en diskret kvantitativ variabel på ordinalskala

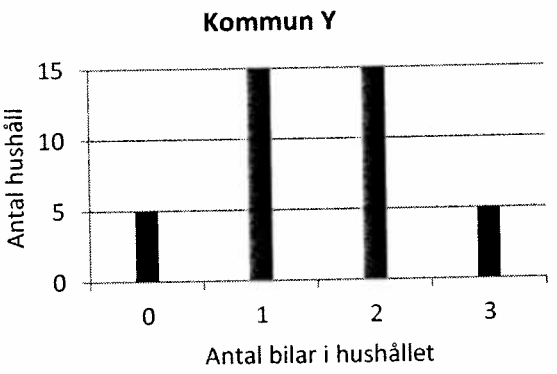
C. Andelen gymnasieelever som börjar vid SU är en numerisk variabel på kvotskala

D. Man kan beräkna medelvärdet för variabeln utbildningsnivå på ett meningsfullt sätt

E. Man kan beräkna medianen för fakultet på ett meningsfullt sätt

I en annan undersökning vill man jämföra antalet bilar per hushåll mellan olika kommuner. För två kommuner, X och Y, samlade man in två oberoende stickprov båda av storlek $n = 40$. Man fick ett resultat som visas i stapeldiagrammen nedan.





Låt $\bar{x}$ och $\bar{y}$ beteckna det genomsnittliga antalet bilar per hushåll och s_x^2 och s_y^2 variansen i antal bilar för respektive kommun.

b) Vilken av följande alternativ är ett korrekt påstående? (5p)

A. $\bar{x} = \bar{y}$ $\quad s_x^2 < s_y^2$

B. $\bar{x} < \bar{y}$ $\quad s_x^2 = s_y^2$

C. $\bar{x} < \bar{y}$ $\quad s_x^2 > s_y^2$

D. $\bar{x} < \bar{y}$ $\quad s_x^2 < s_y^2$

E. $\bar{x} > \bar{y}$ $\quad s_x^2 < s_y^2$

Uppgift 2

En återförsäljare av mobiltelefoner har kommit fram till att 80 % av kunderna vill ha en mobiltelefon med högupplösande (HD) kamera och att 32 % av kunderna vill ha extra minneskapacitet i mobiltelefonen. Dessutom kom man fram till att av alla de kunder som vill ha HD-kamera så vill 20 % också ha extra minne.

a) Vad är sannolikheten att en slumpmässigt vald kund vill ha en mobiltelefon med minst en av de två funktionerna? (4p)

A. 0,36
B. 0,64
C. 0,80
D. 0,96
E. 1,00

Anta att det under en förmiddag kommer in $n = 8$ kunder som vill köpa en mobiltelefon. Anta att de kommer in en och en och att deras önskemål är oberoende av varandra.

b) Vad är sannolikheten att fler än hälften dvs. 5 eller fler, vill ha HD-kamera i mobilen? (5p)

A. 0,990
B. 0,944
C. 0,800
D. 0,797
E. 0,203

OBS! Svarsalternativen för uppgift b) har avrundats till 3 decimaler.

Utgå ifrån följande bivariata fördelningen för de två slumpvariablerna X och Y:

$P(x,y)$	$Y=0$	$Y=1$
$X=0$	0,1	0,1
$X=1$	0,1	0,7

c) Ange korrelationskoefficienten ρ_{XY} mellan X och Y. (6p)

A. $\rho_{XY} = -0{,}375$
B. $\rho_{XY} = 0$
C. $\rho_{XY} = 0{,}375$
D. $\rho_{XY} = 2{,}375$
E. ρ_{XY} kan inte beräknas eftersom både X och Y är diskreta slumpvariabler

Uppgift 3

Aktier i två börsnoterade bolag som vi betecknar med X respektive Y, handlas dagligen i stor omfattning på börsen. Anta att förändringen av aktievärdet för respektive bolag från en dag till nästa är normalfördelade slumpvariabler som vi kan beteckna X och Y. Anta vidare att det finns ett linjärt beroende mellan aktiernas kursförändringar. Väntevärden och varianser samt kovariansen för de två bolagens förändringar i aktiekursen bedöms vara:

$$\mu_X = 1{,}35 \quad \sigma_X^2 = 9{,}0 \quad \mu_Y = 1{,}10 \quad \sigma_Y^2 = 16{,}0 \quad \sigma_{XY} = 5{,}76$$

a) Ange sannolikheten att värdet på X's aktie ökar, dvs. att kursförändringen är positiv. (4p)

A. 0,326
B. 0,560
C. 0,500
D. 0,950
E. 0,674

Anta att du har en aktieportfölj med 10 X aktier och 20 Y aktier.

b) Ange sannolikheten för att värdet på din portfölj ska öka i värde, dvs. att den totala förändringen är positiv. (7p)

A. 0,641
B. 0,359
C. 0,500
D. 0,663
E. 0,975

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

Definiera nu en slumpvariabel Y som har väntevärde μ_Y och varians σ_Y^2. Anta att du observerar ett stickprov av n oberoende observationer av Y och ur dessa beräknas stickprovsmedelvärdet $\bar{Y}$.

c) Ange vilket av följande alternativ som är ett sant påstående. (4p)

A. $\bar{Y}$ är en väntevärdesriktig skattning av μ_Y
B. Väntevärdet för $\bar{Y}$ är approximativt normalfördelat om stickprovet är stort
C. Standardavvikelsen för $\bar{Y}$ är lika med σ_Y/n
D. $\bar{Y}$ är t-fördelad med $n-1$ frihetsgrader
E. Centrala gränsvärdessatsen kan endast tillämpas om Y är normalfördelad

Uppgift 4

Roskildefestivalen har genom åren dragit hundratusentals besökare genom åren om inte mer. Anta nu att man vill få en uppskattning av hur stor andel som är förstagångsbesökare. För att besvara frågan genomförs därför en snabbintervju vid ingången till festivalområdet och vi antar att av $n = 400$ tillfrågade svarar 300 att det är första gången de besöker festivalen.

a) Beräkna ett 95 % konfidensintervall för andelen förstagångsbesökare. (5p)

A. 0,749 – 0,751

B. 0,714 – 0,786

C. 0,701 – 0,799

D. 0,722 – 0,778

E. 0,708 – 0,792

b) Man tycker att felmarginalen är för stor och vill dra ner den för att få bättre skattningar. Hur stort stickprov krävs för att felmarginalen ska vara högst lika med 0,02? TIPS: Vilket värde på andelen brukar man använda när man ska bestämma stickprovsstorleken? (5p)

A. $n = 49$

B. $n = 1225$

C. $n = 1801$

D. $n = 2401$

E. $n = 4802$

Uppgift 5

Universitetslärare är delade i sina meningar kring värdet av att introducera nya tekniska hjälpmedel i undervisningen såsom on-linekurser, appar till mobiler och läsplattor, datoriserade räknelabbar mm. För att studera om det finns skillnader mellan lärare i olika ämnen genomfördes en undersökning på olika universitet där man ställde frågan "Kan undervisningen med hjälp läsplattor förstärka inlärningen bland studenterna?". Svaren sammanställdes i följande tabell:

Ämne	Ja	Nej
Matematik och statistik	50	40
Historia och religion	70	30
Fysik och kemi	60	50

a) Vilken typ av test är lämplig i detta fall för att avgöra om det finns skillnader mellan lärare i olika ämnen och hur är testvariabeln fördelad? (4p)

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

B. Test för differenser av parade observationer – t-fördelad med 2 frihetsgrader

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

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

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

b) Givet att du testar på 5 % signifikansnivå, vilket av följande påståenden om observerat värde på testvariabeln och dess jämförelse mot det kritiska värdet samt slutsats är korrekt? (6p)

A. Observerat värde $|-2{,}000| < 3{,}182$ och $H_0: \mu_D = 0$ förkastas inte

B. Observerat värde $5{,}803 < 5{,}991$ och H_0: "samma fördelning för alla tre ämnen" förkastas inte

C. Observerat värde $6{,}271 < 7{,}815$ och H_0: "samma fördelning för alla tre ämnen" förkastas inte

D. Observerat värde $6{,}271 > 5{,}991$ och H_0: "samma fördelning för alla tre ämnen" förkastas

E. Observerat värde $|2{,}449| > 1{,}96$ och $H_0: \mu_X - \mu_Y = 0$ förkastas

Fullständig redovisning krävs för följande uppgifter.

Använd separata pappersark för uppgift 6 resp. uppgift 7.

Uppgift 6

Nedan redovisas resultaten från högskoleprovet för $n = 10$ gymnasieelever som deltog i högskoleprovet först gången under höstterminen i årskurs 3 och sedan igen under vårterminen.

Student	1	2	3	4	5	6	7	8	9	10
Resultat höst	0,70	0,77	0,90	1,08	1,10	1,25	1,33	1,40	1,42	1,45
Resultat vår	0,75	0,80	1,10	1,25	1,25	1,30	1,25	1,38	1,50	1,42
	0,05	0,03	0,20	0,17	0,15	0,05	-0,08	-0,02	0,08	-0,03

a) Uppskatta med 95 % konfidens den genomsnittliga förändringen i resultatet mellan vår- och höstterminen i den bakomliggande populationen av studenter som deltog under båda dessa terminer. Ange vilka förutsättningar och antaganden som du utgår ifrån. (10p)

b) Utgå ifrån de förutsättningar du angav i a) och testa om den genomsnittliga förändringen är noll mot att den är *större än* noll på 5 % signifikansnivå. Kan du använda resultatet i a) för att besvara frågan? Kommentera kortfattat dina slutsatser. (10p)

Uppgift 7

Utgå ifrån datamaterialet på följande sida där flera delberäkningar redan är gjorda. Du ska med en enkel linjär regressionsmodell förklara variationen i variabeln Y med variabeln X. Följande modell ska alltså behandlas och analyseras:

$$\text{Modell 1: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

En datorutskrift från Excel för den skattade modellen finns i bilagan på följande sida. Flera av uppgifterna har tappats bort och måste räknas om (av dig!).

a) Skatta modellens regressionskoefficienter och ange residualvariansen. (6p)

b) Beräkna förklaringsgraden för modellen och kommentera kortfattat. (4p)

c) Beräkna ett 95 % konfidensintervall för det betingade medelvärdet $\mu_{Y|x_{n+1}}$ för Y givet att $X = 16$, dvs. $x_{n+1} = 16$. (5p)

d) Åskådliggör datamaterialet i ett lämpligt diagram och rita även in den skattade regressionslinjen. Kommentera kortfattat lämpligheten av den valda modellen och t.ex. modellens förmåga att skatta medelvärdet för Y givet att x_{n+1} är lika med 16. (5p)

BILAGA till Uppgift 7

Obs	x	y	x^2	y^2	xy
1	4	3,10	16	9,6100	12,40
2	5	4,74	25	22,4676	23,70
3	6	6,13	36	37,5769	36,78
4	7	7,26	49	52,7076	50,82
5	8	8,14	64	66,2596	65,12
6	9	8,77	81	76,9129	78,93
7	10	9,14	100	83,5396	91,40
8	11	9,26	121	85,7476	101,86
9	12	9,13	144	83,3569	109,56
10	13	8,74	169	76,3876	113,62
11	14	8,10	196	65,6100	113,40
Summa	99	82,51	1001	660,1763	797,59

$\bar{x} = 9$

$s_x^2 = 11$

$\bar{y} = 7{,}501$

$s_y^2 = 4{,}12763$

$s_{xy} = 5{,}5$

Modell 1:

UTDATASAMMANFATTNING

Regressionsstatistik	
Multipel-R	[redacted]
R-kvadrat	[redacted]
Justerad R-kvadrat	[redacted]
Standardfel	[redacted]
Observationer	11

(standardavvikelsen för residualerna)

ANOVA

	f.g.	*KvS* (**SS**)	*MKv* (**MS**)	*F*	*p-värde*
Regression (**R**)	1	27,50	27,50	17,97	0,00218
Residual (**E**)	9	[redacted]	1,531		
Totalt (**T**)	10	[redacted]			

	Koefficient	*Standardfel*	*t-kvot*	*p-värde*
Konstant	[redacted]	1,125	[redacted]	[redacted]
1000 mil	[redacted]	0,118	[redacted]	[redacted]

STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2016

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2016-12-01

LÖSNINGSFÖRSLAG

Preliminär, med reservation för tryck- och slarvfel / 2016-12-19 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)	■	☐	☐	☐	☐
	c)	■	☐	☐	☐	☐
Uppgift 4	a)	☐	☐	☐	☐	■
	b)	☐	☐	☐	■	☐
Uppgift 5	a)	☐	☐	■	☐	☐
	b)	☐	☐	☐	■	☐

Uppgift 1

a) Rätt svar: **C**

En andel är en numerisk variabel; ex. 10 % är hälften av 20 %, alltså gäller kvotskala.

Utbildningsnivå är en kategorisk variabel på ordinalskala (A och B ej korrekta svar), medelvärden kan endast h en mening om det är numeriska variabler (D ej korrekt) och median fungerar för numeriska variabler och en del hävdar att det fungerar för alla typer av ordnade mängder (ordinalskala) men akademiska fakulteter kan inte ordnas på något meningsfullt sätt.

b) Rätt svar: **C**

Man kan resonera kring frågan och svarsalternativen utan att göra faktiska beräkningar.

Medelvärdet för Y torde vara centrerad i mitten då fördelningen är helt symmetrisk men fördelningen för X visar att fler observationer ligger till vänster och därmed är $\bar{x} < \bar{y}$ (och man kan ta bort A och E).

Variansen kan vara lite knepigare men vi har en koncentration av observationer mitt i fördelningen för Y med färre både till höger och vänster om mittpunkten. För X gäller däremot att observationerna mest trängs till vänster men att fördelningen är utdragen mot höger (höger sned, *right skewed*); detta medför att avståndet till medelvärdet lite oftare är större för X än för Y och att $s_x^2 > s_y^2$

Det går också att räkna fram värdena: $\bar{x} = 1 < \bar{y} = 1{,}5$ och $s_x^2 = 1{,}026 > s_y^2 = 0{,}769$.

Uppgift 2

a) Rätt svar: **D**

Beteckna händelsen "vill ha HD-kamera" med K och "vill ha minne" med M.

Givet: $P(K) = 0{,}80 \quad P(M) = 0{,}32 \quad P(M|K) = 0{,}20$

$P(K \cdot M) = P(M|K) \cdot P(K) = 0{,}20 \cdot 0{,}80 = 0{,}16$

Sökt: $P(K \cup M) = P(K) + P(M) - P(K \cap M) = 0{,}8 + 0{,}32 - 0{,}16 = \mathbf{0{,}96}$

b) Rätt svar: **B**

Låt X vara antalet av $n = 8$ kunder som vill ha HD-kamera där sannolikheten för varje enskild kund är $p = 0{,}80$. Eftersom de är oberoende följer att $X \sim Bin(8; 0{,}80)$.

Låt $Y = n - X$ vara antalet som inte vill ha HD-kamera, då följer att $Y \sim Bin(8; 0{,}20)$

Sökt: $P(X > 4) = P(Y \leq 3) = [\text{enl. tabell 7}] = 0{,}94372 \approx \mathbf{0{,}944}$

c) Rätt svar: **C**

De simultana sannolikheterna $P(X = x \cap Y = y) = P(x, y)$ är givna i tabellen. Summera sannolikheterna radvis och kolumnvis för att få marginalsannolikheterna, från dessa beräknas väntevärden och varianser för X och Y:

$P(x,y)$	$Y=0$	$Y=1$	Σ
$X=0$	0,1	0,1	0,2
$X=1$	0,1	0,7	0,8
Σ	0,2	0,8	1

Marginalfördelningarna är identiska:

$$\mu_X = \mu_Y = \sum_x xP(x) = 0 \cdot 0{,}2 + 1 \cdot 0{,}8 = 0{,}8$$

$$\sigma_X^2 = \sigma_Y^2 = \sum_x (x-\mu_X)^2 P(x) = (0-0{,}8)^2 \cdot 0{,}2 + (1-0{,}8)^2 \cdot 0{,}8 = 0{,}16$$

Kovarians och korrelation:

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \sum_x \sum_y xyP(x,y) - \mu_X\mu_Y = 1 \cdot 1 \cdot 0{,}7 - 0{,}8^2 = 0{,}06$$

$$\rho_{XY} = \sigma_{XY}/\sqrt{\sigma_X^2 \cdot \sigma_Y^2} = 0{,}06/0{,}16 = \mathbf{0{,}375}$$

Uppgift 3

a) Rätt svar: **E**

Låt X = förändringen i X-aktien, det är givet att $X \sim N(1{,}35; 3^2)$

Sökt: $P(X>0) = P\left(Z > \frac{0-1{,}35}{3}\right) = P(Z > -0{,}45) = P(Z < 0{,}45) =$

$= [\text{Tabell 1}] = 0{,}67364 \approx \mathbf{0{,}674}$

b) Rätt svar: **A**

Låt $W = 10X + 20Y$ dvs. en linjärkombination av X och Y. Då gäller

$$\mu_W = 10\mu_W + 20\mu_W = 10 \cdot 1{,}35 + 20 \cdot 1{,}1 = 35{,}5$$

$$\sigma_W^2 = 10^2 \cdot \sigma_X^2 + 20^2 \cdot \sigma_Y^2 + 2 \cdot 10 \cdot 20 \cdot \sigma_{XY} = 900 + 6400 + 2304 = 9604$$

$$\sigma_w = \sqrt{9604} = 98$$

Sökt: $P(X>0) = P\left(Z > \frac{0-35{,}5}{98}\right) = P(Z > -0{,}36224) \approx P(Z < 0{,}36) =$

$= [\text{enl. tabell 1}] = 0{,}64058 \approx \mathbf{0{,}641}$

c) Rätt svar: **A**

För stickprovsmedelvärdet gäller alltid att $E(\bar{Y}) = E(Y) = \mu_Y$, dvs. $\bar{Y}$ som en skattning för μ_Y är väntevärdesriktig (unbiased). Detta förutsatt att väntevärdet μ_Y existerar, det finns nämligen fördelningar som inte har väntevärden men här var det givet att den fanns.

Väntevärdet är en parameter, en konstant som inte ändras, alltså är det inte ens en slumpvariabel och kan per definition inte vara fördelad. Standardavvikelsen är $\sigma_Y/\sqrt{n}$. $\bar{Y}$ är inte t-fördelad men om Y är normalfördelad så är $(\bar{Y}-\mu_Y)/(s_y/\sqrt{n})$ där s_y är stickprovets standardavvikelse, t-fördelad med $n-1$ frihetsgrader. Centrala gränsvärdes satsen kan tillämpas på alla fördelningar så länge väntevärde och varians existerar för att approximera fördelningen för $\bar{Y}$; om Y är normalfördelad så är $\bar{Y}$ normalfördelad per automatik och CGS behövs inte.

Uppgift 4

a) Rätt svar: **E**

Ett 95 % KI för andelen P ges av

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

där $\hat{p} = 300/400 = 0{,}75$, $z_{\alpha/2} = z_{0{,}025} =$ [Tabell 2] $= 1{,}96$ och $n = 400$. Insättning ger

$$0{,}75 \pm 1{,}96 \cdot \sqrt{0{,}75 \cdot 0{,}25/400} \;=\; 0{,}75 \pm 0{,}042 \text{ eller } \mathbf{(0{,}708;\, 0{,}792)}$$

b) Rätt svar: **D**

Felmarginalen är $1{,}96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}$ vilket ska vara högst lika med 0,02. Använd $\hat{p} = 0{,}5$ för att maximera felmariginalen, kvadrera både leden och lös ut n:

$$1{,}96^2 \cdot \frac{0{,}5(1-0{,}5)}{n} \leq 0{,}02^2 \;\Leftrightarrow\; 1{,}96^2 \cdot \frac{0{,}25}{0{,}02^2} = \mathbf{2401} \leq n$$

Man kan också sätta in de olika svarsalternativen i formeln och se vilket som ger den sökta felmarginalen.

Uppgift 5

a) Rätt svar: **C**

Homogenitetstest, ett χ^2**-test**. Man vill testa om de tre lärarkategorierna skiljer sig åt med avseende på hur de svarar på frågan, Ja eller Nej. (Detta kan man också kalla oberoendetest eftersom de utförs på samma sätt, skillnaden är bara om lärarkategori ska betraktas som en slumpvariabel eller något fixt).

Man har $r = 3$ lärarkategorier och $c = 2$ svarkategorier, alltså är testvariabeln χ^2**-fördelad** med $(r-1)(c-1) =$ **2 frihetgrader**.

Övriga svarsalternativ är antingen fel (D, fel antal frihetsgrader) eller totalt fel (A, B och E).

b) Rätt svar: **D**

Testvariabel är $\chi^2 = \sum\sum(O_{ij} - E_{ij})^2/E_{ij}$ där $E_{ij} = R_i C_j/n$ och R_i och C_j är rad- respektive kolumnsumma och $n = 300$ är totala antalet observationer.

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

Observerat O_{ij}

50	40	90
70	30	100
60	50	110
180	120	300

Förväntat E_{ij}

54	36	90
60	40	100
66	44	110
180	120	300

$(O_{ij} - E_{ij})^2/E_{ij}$

0,2963	0,4444
1,6667	2,5000
0,5454	0,8182

$\chi^2_{obs} =$ 6,2710

Beslutsregeln är förkasta H_0 om $\chi^2_{obs} > \chi^2_{0{,}05;2} = \mathbf{5{,}991}$, alltså **förkastas nollhypotesen**.

Uppgift 6

a) Eftersom varje student mäts två gånger och vi är intresserade av *förändringen* låter vi D_i beteckna slumpvariabeln som definieras som differensen mellan $X_{\text{vår},i} - X_{\text{höst},i}$ för student nummer i.

Antaganden och förutsättningar för konfidensintervallet (och testet i b) nedan):

- Oberoende och lika fördelade differenser D_i, iid
- Differenserna är normalfördelade: $D_i \sim N(\mu_D, \sigma_D^2)$
- Variansen σ_D^2 är okänd och skattas med s_d^2

Baserat på antagandena används t-fördelningen med $n - 1 = 9$ frihetsgrader.

Ett 95 % konfidensintervall för μ_D ges av $\bar{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$

Från tabell 3: $t_{n-1;\alpha/2} = t_{9;0,025} = \mathbf{2,262}$

Beräkningar ger: $\bar{d} = 0{,}06 \quad s_d^2 = 0{,}008378 \quad s_d = 0{,}09153$

Insättning ger: $0{,}06 \pm 2{,}262 \cdot \frac{0{,}09153}{\sqrt{10}} \Leftrightarrow 0{,}06 \pm 0{,}0655 \Leftrightarrow (\mathbf{-0,0055; 0,1255})$

b) Utifrån antagandena i a) kan vi ställa upp testet enligt följande:

Hypoteser: $H_0: \mu_D = 0$ mot $H_1: \mu_D > 0$

Testvariabel & fördelning: $t = \dfrac{\bar{D} - 0}{S_d/\sqrt{n}}$ ~ t-fördelning med 9 frihetsgrader

Beslutsregel: Förkasta H_0 om $t_{obs} > t_{krit} = t_{9;0,05} = \mathbf{1,833}$

dvs. ett enkelsidigt test och beslutsregel

Beräkningar: $t_{obs} = \dfrac{\bar{d}}{s_d/\sqrt{n}} = \dfrac{0{,}06}{0{,}09153/\sqrt{10}} = \mathbf{2,073} > t_{krit}$

Slutsatser: H_0 förkastas, det observerade genomsnittliga förändringen $\bar{d} = 0{,}06$ är signifikant större än 0 på 5 % signifikansnivå.

Eftersom konfidensintervallet i a) är dubbelsidigt så motsvaras det av ett dubbelsidigt test med hypoteserna $H_0: \mu_D = 0$ mot $H_1: \mu_D \neq 0$ där risken för Typ I fel, dvs. signifikansnivån α, fördelas på båda sidor i fördelningen men i det enkelsidiga testet läggs all risk till höger.

Alltså, med konfidensintervallet i a) kan man inte förkasta påståendet att den genomsnittliga förändringen är lika med noll men med det enkelsidiga testet kan man hävda att det är större än noll!

Uppgift 7

a) Se formelsamlingen; samtliga siffervärden som används är givna i bilagan:

$$b_1 = \frac{Cov(x,y)}{s_x^2} = \frac{s_{xy}}{s_x^2} = \frac{5{,}5}{11} = \mathbf{0,5}$$

$$b_0 = \bar{y} - b_1\bar{x} = 7{,}501 - 0{,}5 \cdot 9 = 3{,}001 \approx \mathbf{3,0}$$

Residualvariansen = $s_e^2 = MSE$ = [enligt utskriften] = $\mathbf{1,531}$

b) Se formelsamlingen:

$$SST = (n-1)s_y^2 = (11-1) \cdot 4{,}12763 = 41{,}2763 \approx 41{,}28$$

$$SSR = \text{[givet i uppgiften]} = 27{,}50$$

$$R^2 = \frac{SSR}{SST} = \frac{27{,}50}{41{,}28} = \mathbf{0,666}$$

Alternativt:

$$SSE = (n-K-1)MSE = \text{[där } K = 1 \text{ förklaringsvariabel]} = 9 \cdot 1{,}531 \approx 13{,}78$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{13{,}78}{41{,}28} = \mathbf{0,666}$$

Kommentar: 66,6 % av variationen i Y kan förklaras av X vilket kan tolkas som relativt högt beroende på sammanhanget (det är i alla fall mer än hälften).

c) Ett 95 % konfidensinterval för det betingade medelvärdet $\mu_{Y|X_{n+1}=16}$ ges av

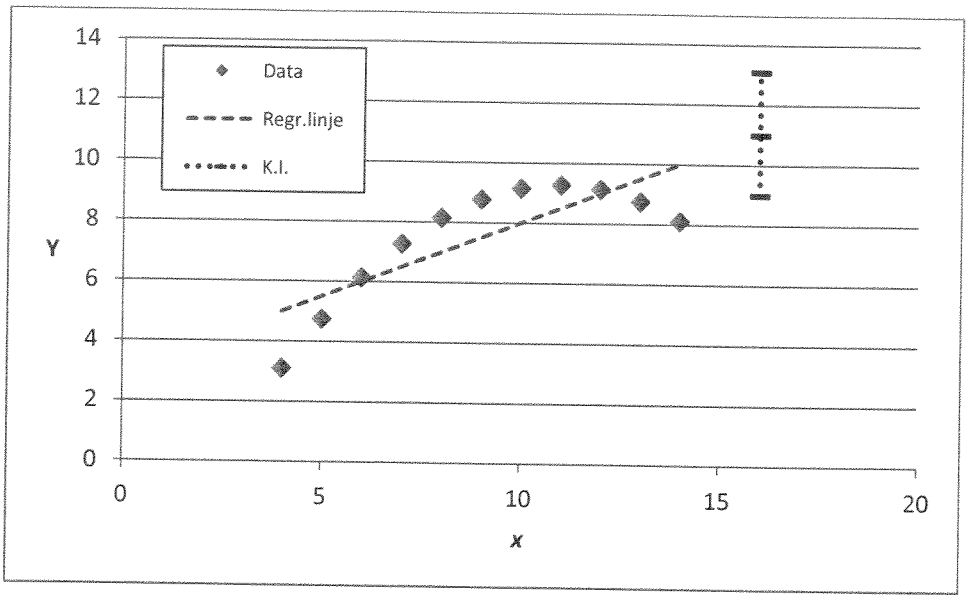
$$b_0 + b_1 x_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{(n-1)s_x^2}\right)}$$

$$3{,}001 + 0{,}5 \cdot 16 \pm 2{,}262 \cdot \sqrt{1{,}531 \cdot \left(\frac{1}{11} + \frac{(16-9)^2}{10 \cdot 11}\right)} \Rightarrow 11{,}001 \pm 2{,}050$$

eller avrundat $11{,}00 \pm 2{,}05$ dvs. $(\mathbf{8,9}\,;\mathbf{13,05})$.

Med 95 % konfidens inkluderar detta intervall $\mu_{Y|X_{n+1}=16}$ givet vissa förutsättningar.

d) Diagram:



Kommentar:

Datapunkterna följer tydligt en kurva som inte är linjär (faktiskt en andragradskurva). Detta indikerar starkt att modellen är helt olämplig för att beskriva datamaterialet. Den skattade linjära regressionslinjen är alltså bara en kompromiss som ger den lägsta residualvariansen (minsta-kvadrat-metoden).

Ytterligare kommentarer:

Överst till vänster i diagrammet har konfidensintervallet från c)-uppgiften lagts in. Detta efterfrågades inte i uppgiften och var således inget krav för full poäng men den har tagits med här för att illustrera hur fel det kan slå när man gör prognoser för y när x-värdet ligger utanför det observerade området för x-variabeln (s.k. extrapolering) och när den anpassade linjära modellen inte beskriver datamaterialet på ett bra sätt.

Om en andragradsmodell $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ istället hade använts så hade man fått den skattade modellen

$$\hat{Y} = -6{,}00 + 2{,}78X - 0{,}13X^2$$

med residualvariansen $s_e^2 = 0{,}0000$ och förklaringsgraden $R^2 = 100\,\%$, dvs. perfekt anpassning. Medelvärdet i c)-uppgiften hade skattats till $\hat{\mu}_{Y|x=16} = 5{,}2$ istället för 11,0. Observera att detta är en utvikning då andragradsmodeller inte ingick i kursen och för full poäng hade det räckt att notera att den linjära modellen inte är lämplig som beskrivning av datamaterialet.