FORELESNING 21

APR 12 2012

MET 3431

EIVIND ERIKSEN

STATISTIKK

PLAN:

Krysstabeller og χ^2-fordeling (Kji-kvadrat fordeling) [T] 11.1, 11.3

Husk: Oppgaver (m/løsning) fra studentveiledning II ligger ute på It's Learning.

1. 11-1: Kji-kvadrat fordelingen
2. 11-3: Krysstabeller og kji-kvadrattesten
3. Kji-kvadrattesten i JMP

Kapittel 11

Samvariasjon mellom to kategoriske variabler

- Korrelasjon og regresjon handler om samvariasjon mellom to kontinuerlige variable
- I dette kapitlet handler det om samvariasjon mellom to *kategoriske* variable

Example

For filmer:

- Sammenheng mellom budsjett og inntekter $\leftrightarrow$ korrelasjon/regresjon
- Sammenheng mellom filmsjanger og filmens nasjonalitet $\leftrightarrow$ krysstabell og kji-kvadrattesten

Krysstabeller

Når vi har samlet inn data om to kategoriske data, kan dette presenteres i en *krysstabell*

Example

Er det en sammenheng mellom farge på hjelmen og risiko for trafikkulykke?
To kategoriske variabler:

Farge Svart, hvit eller gul

Ulykke Involvert eller ikke involvert

Dataene vises i krysstabellen

	Svart	Hvit	Gul	Total
Ikke involvert	491	377	31	899
Involvert	213	112	8	333
Total	704	489	39	1232

$P(\text{svart}) = \frac{704}{1232}$

$P(\text{ikke involvert}) = \frac{899}{1232}$

$P(\text{svart og ikke involvert}) = \frac{704}{1232} \cdot \frac{899}{1232} \approx 0.417$

↑ hvis de to variablene er uavhengige

⇒ Forventet antall: $0.417 \cdot 1232$

Notasjon for krysstabeller

- O står for den observerte hyppigheten (frekvensen) i cellen
- E står den forventede hyppigheten
- r står for antall rader, og c for antall søyler i krysstabellen

Example

	Svart	Hvit	Gul	Total
Ikke involvert	491 (513.714)	377 (356.827)	31(28.459)	899
Involvert	213 (190.286)	112 (132.173)	8 (10.541)	333
Total	704	489	39	1232

- $r = 2$ og $c = 3$
- E er forventet verdi. Selv om 491 syklister med svart hjelm ikke hadde ulykke, så ville vi forventet 513.714 dersom farge og ulykker er uavhengige:

$$E = \frac{899 \cdot 704}{1232} = 513.714$$

$\frac{899}{1232} \cdot \frac{704}{1232} \cdot 1232 =$

Hypotesetest:

H_0: ingen sammenheng

H_1: det er en sammenheng

	sort	hvit	gul	
ikke inv.	491	377	31	899
inv.	213 (190.3)	112	8	333
	704	489	39	1232

O = observert, E = forventet

$$p(\text{svart}) = \frac{704}{1232}$$

$$p(\text{involvert}) = \frac{333}{1232}$$

Hvis ingen sammenheng (Ho sann), så

$$p(\text{svart og involvert}) = \frac{704}{1232} \cdot \frac{333}{1232}$$

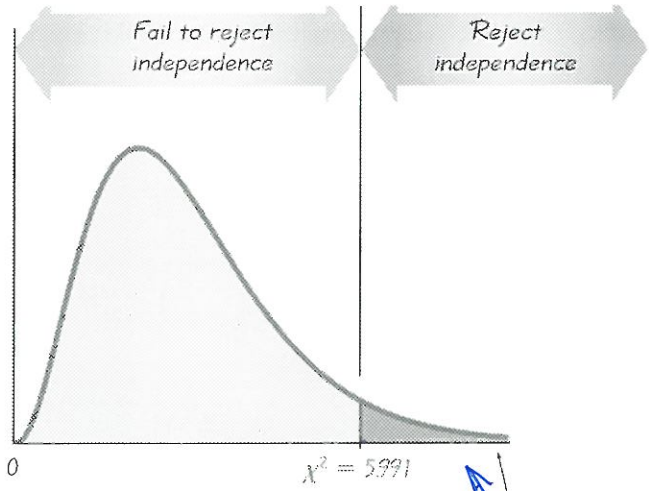
Forventet antall (svart og involvert):

$$\frac{704}{1232} \cdot \frac{333}{1232} \cdot 1232 \simeq 190.3$$

χ = kji (gresk x)

Kji-kvadratfordelingen

- Brukes til å teste om det er *uavhengighet* mellom to kategoriske variable.
- Fordelingen antar bare positive verdier
- Testen er alltid høyresidig
- Fordelingen har frihetsgrader



kritisk verdi for χ^2

forkastnings-område

Kji-kvadrattesten

når vi skal undersøke om det er sammenheng mellom to kategoriske variable.

Hypotesene

- H_0 : Ingen sammenheng mellom variablene
- H_1 : Det er en sammenheng mellom variablene

Testobservatoren

Testobservatoren er

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \cdots$$

med $(r-1)(c-1)$ frihetsgrader. Kritisk verdi finnes i Tabell A4

antall frihetsgrader

r = antall rader

c = antall kolonner = k

$r=2$, $c=3$ } $df = (2-1)\cdot(3-1) = 1\cdot 2 = 2$

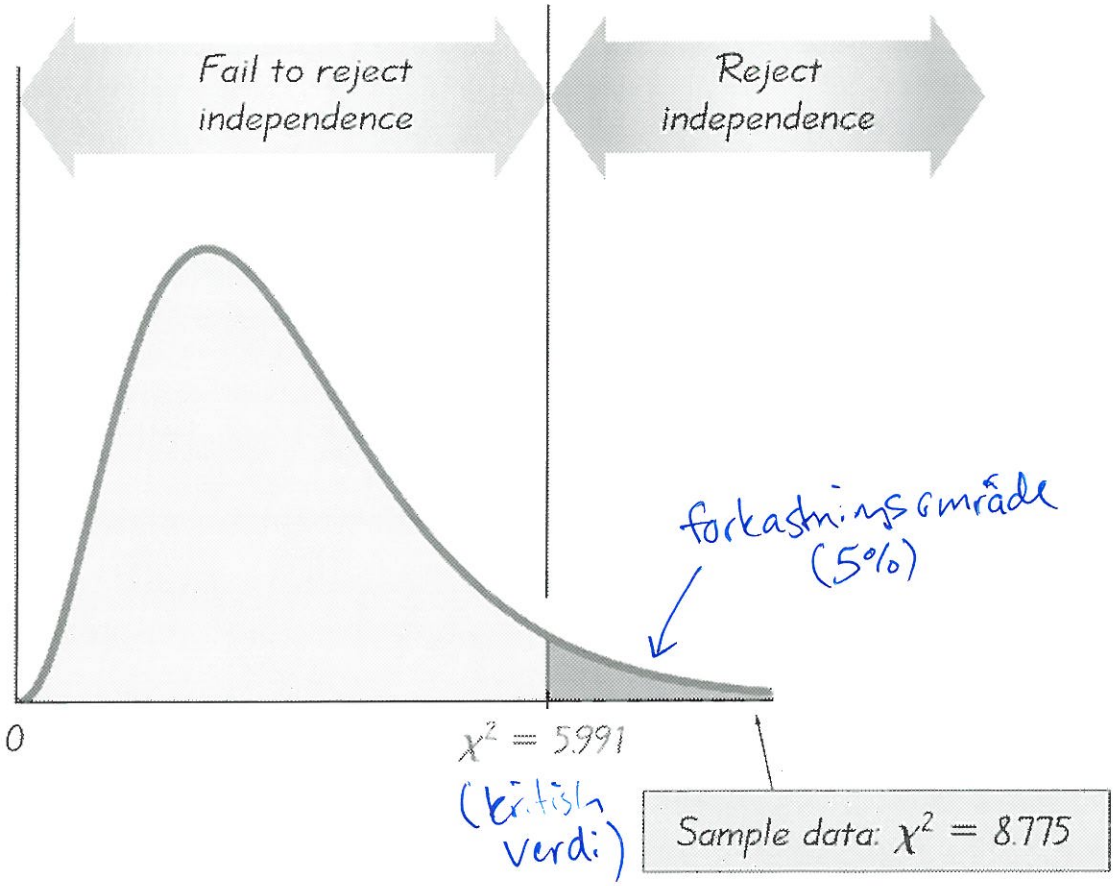
Example

- H_0 : Ingen sammenheng mellom hjelmfarge og utsatthet for ulykker
- H_1 : Det er en sammenheng mellom farge og ulykker

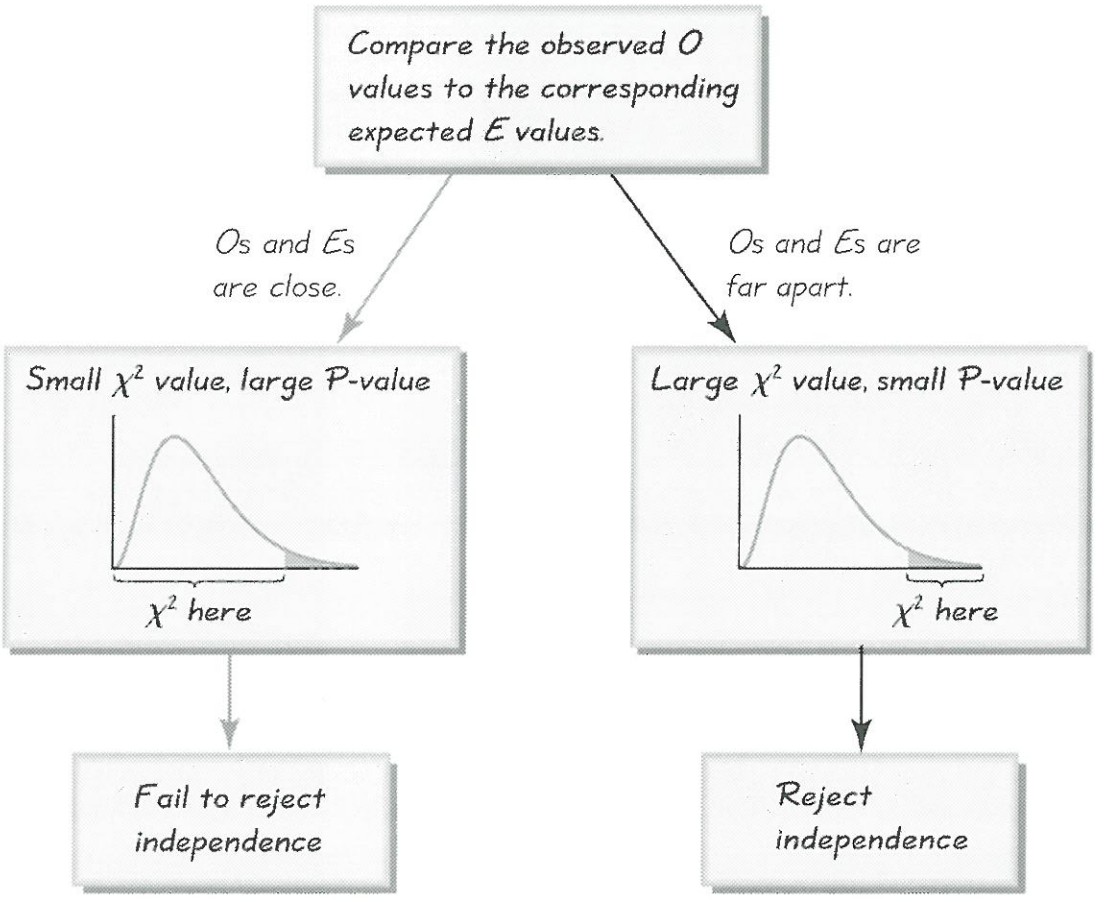
Testobservatoren er

$$\chi^2 = \frac{(491 - 513.714)^2}{513.714} + \frac{(377 - 356.827)^2}{356.827} + \cdots + \frac{(8 - 10.541)^2}{10.541} = 8.775$$

- Hvor 'uvanlig' er denne testobservatoren?
- Sjekk med $(2-1)\cdot(3-1) = 2$ frihetsgrader i Tabell A4
- Kritisk verdi er $\chi_{0.05,2} = 5.991$
- Vi har en *uvanlig* testobservator: Forkast H_0
- Det ser ut til å være en sammenheng mellom farge og ulykker



Figur: Kji-kvadrattesten er alltid høyrehalet



Figur: Logikken bak kji-kvadrattesten

Krav til testen

For at testen skal virke må det forventede antallet E være minst 5 i hver celle

Example

	Svart	Hvit	Gul	Total
Ikke involvert	491 (513.714)	377 (356.827)	31(28.459)	899
Involvert	213 (190.286)	112 (132.173)	8 (10.541)	333
Total	704	489	39	1232

Her er alle forventede verdier høyere enn 5

Nytt eksempel

Example

Krysstabell for ulydighet og plass i søskenrekka:

	Eldst	Midten	Yngst	Enebarn	Total
Ulydig	24	29	35	23	111
Lydig	450	312	211	70	1043
Total	474	341	246	93	1154

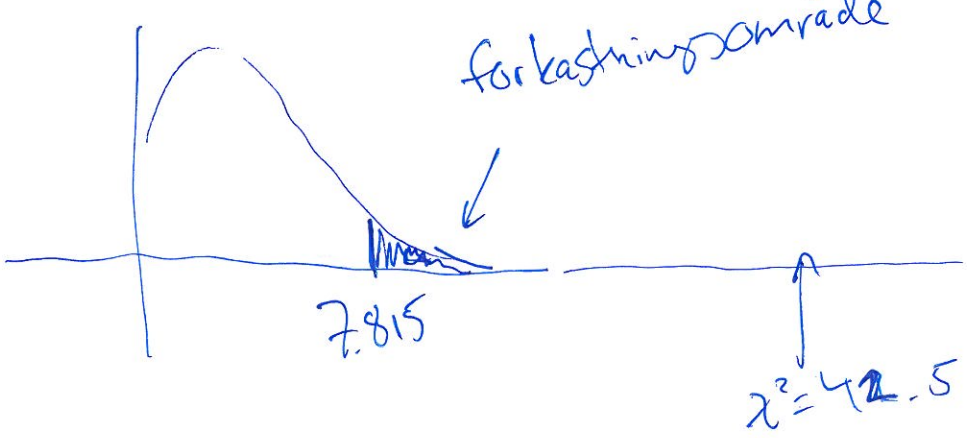
- Setter $\alpha = 0.05$
- H_0 : Ingen sammenheng mellom lydighet og plass i søskenrekka
- H_1 : Det er en sammenheng

Kritisk χ^2-verdi:

$df = (2-1) \cdot (4-1) = 1 \cdot 3 = 3$

$\alpha = 0.05$

$\chi^2_{0.05} = 7.815$



Dersom plass i rekka ikke har noen betydning forventer vi at

$$\frac{341 \cdot 111}{1154} = 32.800$$

er ulydige blant de i midten.

	Eldst	Midten	Yngst	Enebarn	
Ulydig	24(45.593)	29 (32.800)	35 (23.662)	23 (8.945)	111
Lydig	450 (428.407)	312 (308.200)	211 (222.338)	70 (84.055)	1043
Total	474	341	246	93	1154

- Testobservatoren blir

$$\chi^2 = \frac{(24-45.6)^2}{45.6} + \frac{(29-32.8)^2}{32.8} + \cdots + \frac{(70-84.1)^2}{84.1} = 42.5$$

- Frihetsgrader $\nu = (2-1)\cdot(4-1) = 3$
- Kritisk verdi fra tabell A4 er $\chi^2_{0.05,3} = 7.815$

Konklusjonen i Kji-kvadrattesten

Er kritisk verdi større enn testobservatoren?

1. Har regnet ut forventede verdier E
2. Regnet ut testobservator χ^2 fra E og O
3. Så fant vi kritisk verdi χ^2_α i Tabell A4
4. : Forkast H_0 dersom testobservatoren er større enn kritisk verdi χ^2_α

Example

- Testobservatoren hadde verdi $\chi^2 = 42.5$
- Kritisk verdi var $\chi^2_\alpha = 7.81$
- Konklusjon: Forkast H_0!
- Det er en sammenheng mellom lydighet og plass i søskenrekka

Kji-kvadrattesten

Kji-kvadrattesten for uavhengighet mellom to kategoriske variable

1. Vi har to kategoriske variabler (r og k verdier i hver variabel)
   - H_0: Det er ingen sammenheng mellom variablene
   - H_1: Det er en sammenheng mellom variablene
2. Observasjonene er ordnet i en $r \times k$ krysstabell
3. Finn for hver celle i tabellen den forventede (dersom H_0 er sann) hyppigheten
4. Regn ut testobservatoren

$$\chi^2 = \Sigma \frac{(\text{Observert-Forventet})^2}{\text{Forventet}}$$

5. Finn kritisk verdi χ^2_α i Tabell A4 med frihetsgrad $(r-1)\cdot(k-1)$
6. Forkast H_0 dersom $\chi^2 > \chi^2_\alpha$

Nytt eksempel

Example

- Studentene velger studium fra $r = 4$ kategorier
- Jobbsektoren som de havner i etter studiet deles i $k = 5$ kategorier
- En markedsundersøkelse av 121 BI studenter gir krysstabell

Studie / **Jobb**	Organ.	Finans	Markedsf.	Regnsk	Sum
Finans	7	7	0	3	17
Markedsf.	8	1	20	0	29
Ledelse	7	3	3	1	14
Regnsk	6	10	1	24	41
Salg	10	3	6	1	20
Total	38	24	30	29	121

Hypotesene: Sammenheng eller ikke sammenheng?

Example

- H_0 : Det er ingen sammenheng mellom studieretning og jobbtype
- H_1 : Det er en sammenheng mellom studieretning og jobbtype
- Test på 1% nivået

Example

- Regner ut de *forventede verdiene* E
- F.eks. kan vi forvente at $24 \cdot \frac{29}{121} = 5.752$ finansstudenter havner i en markedsføringsjobb

Studie / **Jobb**	Organ.	Finans	Markedsf.	Regnsk	Sum
Finans	7(5.3)	7(3.4)	0(4.2)	3(4.1)	17
Markedsf.	8(9.1)	1(5.8)	20(7.2)	0(7.0)	29
Ledelse	7(4.4)	3(2.8)	3(3.5)	1(3.4)	14
Regnsk	6(12.9)	10(8.1)	1(10.2)	24(9.8)	41
Salg	10(6.3)	3(4.0)	6(5.0)	1(4.8)	20
Total	38	24	30	29	121

Bruk tre desimaler

Vi har avrundet til en desimal for å forenkle tabellen. I utregningen benyttes tre desimaler.

$$\frac{38 \cdot 17}{121} = 5.3$$

Example

- Testobservatoren

$$\chi^2 = \frac{(7-5.339)^2}{5.339} + \frac{(7-3.371)^2}{3.371} + \frac{(0-4.215)^2}{4.215} + \cdots + \frac{(1-4.794)^2}{4.794} = 84.5$$

- Kritisk verdi på 1% nivået, frihetsgrad $(4-1)\cdot(5-1) = 12$, fra tabell A4 er

$$\chi^2_{0.01,12} = 26.217$$

- H_0 forkastes siden testobservatoren er større enn kritisk verdi
- Det er altså en sammenheng mellom studieretning og jobb etter endt studium

Krav får å bruke kji-kvadrattesten

- Forventede verdier E bør være minst 5
- Maksimalt 20% av cellene bør ha lavere verdi enn 5
- Det er altså helt i grenseland å stole på testen for dette eksempelet!

Legge inn dataene i JMP

JobbStudier

JobbStudier

Columns (3/0)
Jobb *
Studie *
Antall

	Jobb	Studie	Antall
1	Finans	Organisasjon	7
2	Finans	Finans	7
3	Finans	Ledelse	0
4	Finans	Regnskap	3
5	Markedsførin	Organisasjon	8
6	Markedsførin	Finans	1
7	Markedsførin	Ledelse	20
8	Markedsførin	Regnskap	0
9	Ledelse	Organisasjon	7
10	Ledelse	Finans	3

Jobb nominal. *Data type: numeric* og *Modeling type: nominal*

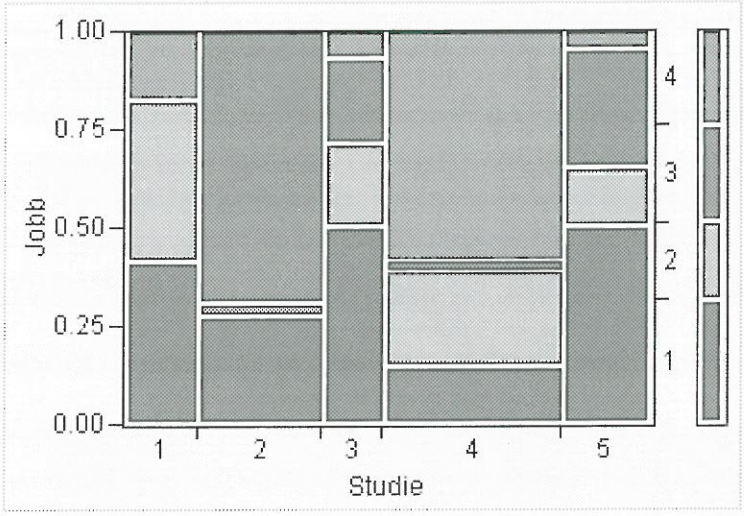
Studie nominal. *Data type: numeric* og *Modeling type: nominal*

Antall kontinuerlig.

For *Jobb* og *Studie* er det lurt å legge til *Value labels*. Legg inn inn tall i tabellen, men når resultatene vises, så kommer navnet på studiet opp.

Contingency Analysis of Jobb By Studie

Mosaic Plot



Analyze > Fit Y by X

1. Legg inn *Jobb* på Y
2. Legg inn *Studie* på X
3. Legg inn *Antall* på Freq

JMP utskrift for kji-kvadrattest

Tests Frihetsgrader, degress of freedom

N	DF	-LogLike	RSquare (U)
121	12	45.591227	0.2460

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	91.182	<.0001*
Pearson	84.496	<.0001*

Testobservator

p-verdi

Warning: 20% of cells have expected count less than 5. ChiSquare suspect.

- Forutsetningene til kji-kvadrattesten er ikke møtt, så vi får advarsel
- DF = frihetsgrader
- Pearson = Kji-kvadrat testobservator
- Vi ser at JMP gir p-verdi $0,0001$. Altså skal vi forkaste på ethvert rimelig α nivå

Oppgave 5

(a) I hypotesetesting, hva vil det si å begå en type I feil? Anta at du tester på signifikansnivå $\alpha = 0.05$, og at H_0 er sann. Hva er sannsynligheten for å begå en type I feil?

(b) I en stikkprøve på 157 studenter ved BI Bergen våren 2011 hadde 13 studenter mobiltelefon av merket Samsung. Test påstanden om at mindre enn ti prosent av studentene ved BI Bergen har Samsung mobiltelefon. Skriv opp nullhypotesen og alternativhypotesen. Foreta testen på et $\alpha = 0.05$ signifikansnivå og formuler konklusjonen i et lettfattelig språk.

(c) Du har tilgang til stikkprøver av ukentlige timer på jobb for 60 mannlige studenter og for 99 kvinnelige studenter. Er de to stikkprøvene med menn og kvinner uavhengige eller relaterte? Test hypotesen om at mannlige studenter gjennomsnittlig jobber mer enn kvinnelige studenter basert på de to stikkprøvene vist i Figur 2. Skriv opp nullhypotesen og alternativhypotesen og bruk signifikansnivå $\alpha = 0.1$. Skisser p-verdien som arealet under en graf.

Oppgave 6

(a) Figur 3 viser et scatterplott av timer på jobb vs studietimer per uke for en tilfeldig valgt stikkprøve av åtte BI studenter våren 2011. Hvor mange studenter i stikkprøven oppga at de studerte 30 timer i uka? Hvor mye jobbet hver av disse studentene i uka?

(b) Korrelasjonskoeffisienten for dataene i Figur 3 er $r = 0.73$. Test på $\alpha = 0.05$ nivået om det er en korrelasjon mellom timer på jobb og timer brukt på studier. Skriv opp H_0 og H_1.

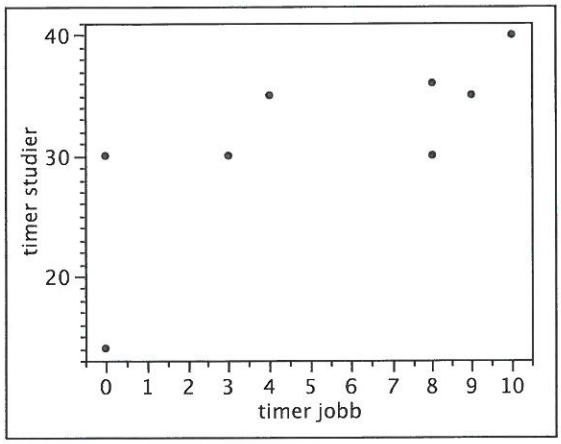
(c) Gi et eksempel på to kontinuerlige variable som du tror er signifikant negativt korrelerte.

Oppgave 7

Figur 4 gir krysstabellen for sammenhengen mellom kjønn og studium i en tilfeldig stikkprøve av 291 studenter fra BI Trondheim. I hver celle oppgis det faktiske og det forventede antallet.

(a) Hvilken hypotesetest kan du bruke for å avgjøre om det er en sammenheng mellom kjønn og studium i populasjonen av studenter ved BI Trondheim?

(b) Foreta en test på 1% nivået og konkluder i et lettfattelig språk.



Figur 3: Oppgave 6

		kjønn		
	Count Expected	Mann	Kvinne	
Studium	Økonomi	92 79,6701	69 81,3299	161
	Markedsføring	52 64,3299	78 65,6701	130
		144	147	291

Figur 4: Oppgave 7