Rettevejledning til eksamen på Økonomistudiet vinter 2013-2014 Økonometri A

Målbeskrivelse:

Kurset har som mål at introducere studerende til sandsynlighedsteori og statistik. Målet er, at de studerende efter at have gennemført faget kan:

- Forstå og benytte de vigtigste sandsynlighedsteoretiske begreber som: sandsynlighed, simultane-, marginale- og betingede- sandsynligheder, fordeling, tæthedsfunktion, uafhængighed, middelværdi, varians og kovarians samt at selvstændigt kunne anvende disse begreber på konkrete problemstillinger

- Kende resultatet fra den centrale grænseværdi sætning

- Kende og genkende de mest anvendte diskrete og kontinuerte fordelinger som: Bernoulli, Binomial, Poisson, multinomial, negative binomial fordeling, hypergeometrisk, geometrisk, lige-, normal-, Chi-i-anden-, eksponential, gamma-, t-, F-fordeling samt at selvstændigt kunne arbejde med disse fordelinger i konkrete problemstillinger

- Forstå de vigtigste statistiske begreber som: tilfældige udvælgelse, likelihood funtionen, sufficiens, stikprøvefunktion, egenskaber ved stikprøvefunktionen, estimation heruden maksimum likelihood og moment estimation, konsistens, konfidensinterval, hypoteseprøvning, teststørrelser, hypoteser, testsandsynlighed, signifikansniveau og type I og II fejl

- Være i stand til selvstændigt at gennemføre en simpel statistisk analyse, som involverer estimation, inferens og hypoteseprøvning, f.eks. sammenligning af middelværdien i to populationer eller uafhængighedstest for diskrete stokastiske variable.

- Indlæse og kombinere datasæt, lave nye variable, udtrække en stikprøve og udføre simple statistiske analyser ved hjælp af statistik-pakken SAS

- Beskrive resultatet af egne analyser og overvejelser i et klart og tydeligt sprog

Opgave 1

Forsknings og udviklingsafdelingen i en virksomhed har udviklet et nyt produkt, der skal testes. De udfører 40 test uafhængigt af hinanden. Hvert test har to udfald: enten går testet godt, "succes", eller også fejler testet, "fiasko". I første omgang kender afdelingen ikke sandsynligheden for om det går godt.

1. Lad X være antallet af succeser i de 40 test. Hvilken fordeling følger X? Hvad er sandsynligheden for at alle test går godt, hvis sandsynligheden for succes er $0,9$?

Svar: X er binomial fordelt med $n = 40$, da forsøgene er uafhængige og hvert udfald er Bernouilli. $\Pr(X = 40) = \binom{40}{40} 0,9^{40} = 0,9^{40} \simeq 0,0148$.

De 40 test har været udført af to forskellige personer i virksomheden. En statistikker analyserer de to personers forsøg og finder at person A er sandsynligheden for succes $0,85$ og for person B er sandsynligheden for succes $0,95$. Person A er til gengæld hurtigere til at udføre testene. Af 100 test vil de 55 være gennemført af person B og 45 af person A.

2. Find sandsynligheden for en succes uafhængigt af om det er person A eller person B, der gennemfører testet.

Svar: Der er en fejl i teksten, idet det er person A, der er den hurtigste og kan gennemføre 55 test. Men besvarelser, hvor person B er den hurtigste, er ok. Loven om den totale sandsynlighed anvendes: $p = p^A \cdot \Pr(A) + p^B \cdot \Pr(B) = 0,85 \cdot 0,55 + 0,95 \cdot 0,45 = 0,895$

De to personer bliver bedt om lave et test hver og møde op på chefens kontor. Den første person har haft en succes, men det har den anden person ikke. Lad D være en stokastisk variabel for den første persons type, så når $D = A$ er den første person type A. Lad X_1 og X_2 være udfaldet af testet for henholdsvis den første og den anden person.

3. Hvad er sandsynligheden for, at den første person er person A? (Hint: Brug Bayes teorem og at der er uafhængighed mellem udfaldet af testene betinget på type. Dvs. $\Pr(X_1, X_2|D) = \Pr(X_1|D) \cdot \Pr(X_2|D)$)

Svar: Find $\Pr(D = A|X_1 = succes, X_2 = fiasko) = \frac{\Pr(X_1=succes, X_2=fiasko|D=A)\cdot\Pr(D=A)}{\Pr(X_1=succes, X_2=fiasko)} = \frac{\Pr(X_1=succes|D=A)\cdot\Pr(X_2=fiasko|D=A)\cdot\Pr(D=A)}{\Pr(X_1=succes, X_2=fiasko)} = \frac{0,85\cdot0,05\cdot0,50}{0,85\cdot0,05\cdot0,50+0,15\cdot0,95\cdot0,50} \simeq 0,230$.Bemærk at sandsynligheden i nævneren beregnes vha. loven om den totale sandsynlighed. Initialt er sandsynligheden $0,5$ for at den første er person A. Men fordi vedkommende har en succes og den anden ikke har en succes, falder sandsynligheden for at det er person A, da denne har mindre sandsynlighed for succes end person B.

Opgave 2

En virksomhed kan enten være ejet af en dansker eller en udlænding og en virksomhed har enten elever eller også har den ikke elever. Antallet af virksomheder i hele økonomien fordelt på ejerskabet og elevstatus er givet ved følgende tabel,

	Elevstatus		
Ejerskab	Har elever	Har ikke elever	Total
Dansk ejerskab	63	117	180
Udenlandsk ejerskab	7	13	20
Total	70	130	200

Lad Y være en stokastisk variabel for ejerskab og Z en stokastisk variabel for elevstatus.

1. Find den simultane fordeling for Y og Z. Find de marginale fordeling for henholdsvis Y og Z. Er der uafhængighed mellem Y og Z?

Svar:

simultate fordeling:

	Elevstatus		
Ejerskab	Har elever	Har ikke elever	Total
Dansk ejerskab	$0,315$	$0,585$	$0,9$
Udenlandsk ejerskab	$0,035$	$0,065$	$0,1$
Total	$0,35$	$0,65$	1

De marginale fordelinger:

Y	
Dansk	$0,9$
Udenlandsk	$0,1$

Z	
har elever	$0,35$
Har ikke elever	$0,65$

Der er uafhængighed, da fx $0,9 \cdot 0,35 = 0,315$. Hvis de øvrige celler undersøges findes samme svar.

Den følgende tabeller viser antallet af virksomheder fordelt på ejerskab og elevstatus i branche A.

	Elevstatus		
Ejerskab	Har elever	Har ikke elever	Total
Dansk ejerskab	40	45	85
Udenlandsk ejerskab	5	10	15
Total	45	55	100

2. For virksomhederne i branche A find fordelingen af ejerskab betinget på elevstatus. Find også den marginale fordeling af ejerskab. Er der uafhængighed? Kommenter i fht. forrige spørgsmål.

Svar:

Den betingede fordeling

f(Y|Z)	har elever	har ikke elever
dansk	$\frac{40}{45}$	$\frac{45}{55}$
udenlandsk	$\frac{5}{45}$	$\frac{10}{55}$

Den marginale fordeling

	Y
dansk	$\frac{85}{100}$
udenlandsk	$\frac{15}{100}$

Der er afhængighed. Kan ses på mange måder: fx $\frac{40}{45} \neq \frac{45}{55}$. I denne branche er sandsynligheden for at virksomheden er dansk betinget på, at den har elever, større end blandt virksomheder, som ikke har elever. I en analyse kan det derfor være vigtigt at kontrolle for fx branche, da uafhængighed ændres, når vi betinger på en tredje variabel.

Regeringen fører nu en kampagne, der er målrettet virksomheder, der ikke har elever. I branche A er sandsynligheden for at en virksomhed, der ikke i forvejen har elev, 20 pct. efter kampagnen.

3. Hvad er det forventede antal virksomheder, der har elever, i branche A efter kampagnen?

Svar: Først antager vi at kampagnen ikke har indflydelse på de virksomheder der har elever dvs. der er 5 udenlandske og 40 danske. Af de 45 danske og 10 udenlandske, der ikke har elever, er det nu 20 pct. af dem som har, dvs. 11. Det forventede antal virksomheder er altså 56.

Opgave 3

I forbindelse med en større international undersøgelse kaldet European Social Survey (ESS) er der udtrukket 1.496 tilfældige personer i Danmark ud af de i alt 4.079.910 stemmeberettigede personer. Blandt de 1.496 svarede 1.405 personer at de havde stemt ved det sidste folketingsvalg i 2011. Dermed var der 91 personer der har svaret, at de ikke har deltaget i det sidste folketingsvalg i 2011.

1. Opstil en model for antallet af personer der har stemt ved det sidste folketingsvalg i 2011 blandt de 1.496 udspurgte personer. Argumenter for modellen.

Svar: Oprindeligt bliver der tale om en hypergeometrisk fordeling. Her er stikprøven stor; nemlig 1.496 og denne stikprøve kan ubesværet approximeres til en binomialfordeling. Så X =antallet af personer der har stemt ved sidste folketingsvalg er binomialfordelt med antalsparameter 1.496 og sandsynlighedsparameter p. Her udtrykker p andelen af befolkningen, der har stemt ved sidste folketingsvalg.

2. Estimer andelen af personer (p) i Danmark, der har stemt ved det sidste folketingsvalg i 2011. Redegør for estimatorens egenskaber.

Svar: $p = \frac{x}{n} = 1405/1496 = 93,9\%$
Estimatoren er middelret, dvs. $E(\hat{p}) = p$ og endvidere går variansen mod nul når n går mod uendelig.
Dette skyldes at $Var(\widehat{p}) = \frac{p(1-p)}{n}$. Så samlet er estmatoren konsistent

3. Udregn et 95% konfidensinterval for estimatoren for p.

Svar: $\hat{p} \pm 1,96\sqrt{((\hat{p}(1-\hat{p}))/n)}$
$0,939 \pm 1,96\sqrt{((0,939 * (1-0,939))/1496)}$
$0,939 \pm 0,012$
$92,7\%$ til $95,1\%$

Ved det sidste folketingsvalg var stemmeprocenten på $87,7\%$.

4. Test om stemmeprocenten i ESS undersøgelsen er mindre end stemmeprocenten ved det sidste folketingsvalg.

Her burde der nok være spurgt om stemmeprocenten i ESS var større end den ved valget. Men nu er der spurgt om stemmeprocenten var mindre end valget.
$H_0 : p = 0,877$ $H_A : p < 0,877$

$U = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0{,}939-0{,}877}{\sqrt{\frac{0{,}877*(1-0{,}877)}{1496}}} = 7{,}3$ en lidt "uhyggelig" stor teststørrelse. Sandsynligheden for at falde til venstre for $7{,}3$ i en standardiseret normalfordeling er i praksis lig 1. Dermed er signifikanssandsynligheden meget stor og nulhypotesen kan ikke afvises! Det er i orden hvis der kommer en del kritik af dette spørgsmål.

Man ønsker at sammenligne Danmark med Norge og Sverige. I nedenstående tabel er vist de tilsvarende oplysninger for Norge og Sverige.

ESS	Danmark	Norge	Sverige	i alt
stemt	1.405	1.221	1.515	4.141
ikke stemt	91	188	159	438
I alt	1.496	1.409	1.674	4.579

5. Opstil en statistisk model for ovenstående tabel

svar: I spørgsmålet skal der opstilles en model. Og her er det oplagt at argumentere for, at der er tre uafhængige binomialfordelinger. En for Danmark en for Norge og en for Sverige. Argumenterne fra sp. 3.1 kan bruges for hvert enkelt land. Så samlet 3 uafhængige binomialfordelinger. Med hver sin p i første omgang

$X_{dk} \sim bin(1496, p_{dk})$, $X_{no} \sim bin(1409, p_{no})$, $X_{se} \sim bin(1674, p_{se})$

Alternativ model: (X_{ij}) er multinomisk$(4579, p_{ij})$. Her er $i = 1, 2$ (stemt, ikke stemt) og $j = 1, 2, 3$ angiver de 3 lande. Bemærk at den multinomiske har 6 celler.

6. Test om stemmeprocenten er ens for de tre nordiske lande.

svar: $H_0 : p_{dk} = p_{no} = p_{se}$ H_A : at mindst to af lande er forskellige

Testet bliver et chi-i-anden test. For hver celle skal der udregnes det forventede antal personer, givet at de tre lande har samme p.

De forventede værdier bliver:

ESS	Danmark	Norge	Sverige	I alt
stemt	$1352,9$	$1274,2$	$1513,9$	4141
ikke stemt	$143,1$	$134,8$	$160,1$	438
I alt	1496	1409	1674	4579

I næste tabel er vist testbidragene, samt den samlede teststørrelse

ESS	Danmark	Norge	Sverige	I alt
stemt	$2,01$	$2,22$	$0,00$	$4,23$
ikke stemt	$18,97$	$21,02$	$0,01$	$39,99$
I alt	$20,97$	$23,34$	$0,01$	$44,22$

Så teststørrelsen bliver $44,2$ som er chi-i-anden fordelt med frihedsgrader= 2.

Som er udregnet ved antallet af rækker-1 ganget med antallet af søjler -1 hvilket giver $(2-1)*(3-1)=2$.

Signifikanssandsynligheden (p-værdien) bliver sandsynligheden for at observere en størrelse der er større end $44,2$ i chi-i-anden fordelingen med $df=2$.

Dette bliver i praksis nul.

Her kommer et SAS program der regner det hele ud

```
data a;
input land $ stemt $ antal;
cards;
DK stemt 1405
DK ikke_stemt 91
NO stemt 1221
NO ikke_stemt 188
SE stemt 1515
SE ikke_stemt 159
;
proc freq data=a;
table stemt*land/norow nocol nopercent cellchi2 chisq;
weight antal;
run;
```