Rettevejledning til eksamen i Økonometri A 2012-I, 2. år

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

1. X er binomial fordelt, når udfaldet på ansøgningerne er uafhængige og ensfordelte Bernouilli fordelte. Kræver mange virksomheder.

$$P(k=13) = \binom{13}{13} 0,95^{13} 0,05^{0} = 0,513342.$$

2. Der er nu 3 udfald (afslag,samtale,kontrakt) med sandsynligheder $0,95, 0,04$ og $0,01$.Det er en multinomialfordeling.

$$P(13,2,0) = \binom{15}{13,2,0} 0,95^{13} 0,04^{2} 0,01^{0} = 0,086241$$

3. Y er negativ binomialfordelt. Udfaldet på ansøgningerne er uafhængige og ensfordelte. Kræver mange virksomheder. $E[Y] = \frac{2}{0,04} = 50$. Sandsynligheden for k ansøgninger er givet ved $f(k|0,04) = \binom{k-1}{1} 0,04^{2} 0,96^{k-2}$ og maksimum er for 24 og 25 ansøgninger, hvor sandsynligheden er $0,015017$.

Opgave 2

1. Middelværdi er 160. Variansen finder vi fra 25 pct. percentilen: 150 = 160-$0,6745\sigma$, så $\sigma = 14,8258$. Lad X_i være rengøring i time i, så dagens samlede rengøring $Y = \sum_{i=1}^{8} X_i$.Antag uafhængighed mellem timerne. Y er normalfordelt med middelværdi 1280 og varians $1758,434$.

$$P(1200 < Y < 1250) = \Phi(\frac{1250-1280}{\sqrt{1758,434}}) - \Phi(\frac{1200-1280}{\sqrt{1758,434}}) = 0,237176 - 0,02821 = 0,208966$$

ALTERNATIV: Uden uafhængighed kan opgaven også regnes. Det skal bare være konsekvent igennem hele opgaven. Fx med perfekt korrelation ($\rho = 1$) er $Y = 8X$. Her Bliver middelværdien den samme, men variansen bliver $Var(8X) = 64 \cdot Var(X) = 14067.47$.

$$P(1200 < Y < 1250) = \Phi(\frac{1250-1280}{\sqrt{14067.47}}) - \Phi(\frac{1200-1280}{\sqrt{14067.47}}) = 0,400158 - 0,249997 = 0,150162$$

2. Hver anden dag holder danskerne 4 pauser og derfor ændrer sandsynligheden for 1200 til 1250 kvadratmeter rengøring sig. Hvis $\widetilde{X_i}$ er rengøring i timer med pauser, så er $Y^{DK} = \sum_{i=1,3,5,7} X_i + \sum_{i=2,4,6,8} \widetilde{X_i}$. Denne er normalfordelt med middelværdien 1200 og varians (under uafhængighed) $1758,434$.

$$P(1200 < Y^{DK} < 1250) = \Phi(\frac{1250-1200}{\sqrt{1758,434}}) - \Phi(\frac{1200-1200}{\sqrt{1758,434}}) = 0,38344$$

Set på en tilfældig dag er sandsynligheden for en dansker derfor $0,5 \cdot 0,38344 + 0,5 \cdot 0,208966 = 0,296203$. Altså lidt højere end for polakker. For en tilfældig arbejder på en tilfældig dag er det $0,3 \cdot 0,208966 + 0,7 \cdot 0,296203 = 0,270032$.

ALTERNATIV: Tilsvarende uden uafhængighed. Fx med perfekt korrelation er $Y^{DK} = 4\widetilde{X} + 4X$. Middelværdien er 1200 og varansen er 14067.47.

$$P(1200 < Y^{DK} < 1250) = \Phi(\frac{1250-1200}{\sqrt{14067,47}}) - \Phi(\frac{1200-1200}{\sqrt{14067,47}}) = 0,163328$$

Set på en tilfældig dag er sandsynligheden for en dansker derfor $0,5 \cdot 0,163328 + 0,5 \cdot 0,150162 = 0,156745$. Altså lidt højere end for polakker. For en tilfældig arbejder på en tilfældig dag er det $0,3 \cdot 0,150162 + 0,7 \cdot 0,156745 = 0,15477$.

3. Lad D være enten DK for dansker eller PL for polak. Lad Z være præstationen.

$$P(D = DK | z \in (1200, 1250)) = \frac{P(1200 < Z < 1250 | D = DK) \cdot P(D = DK)}{P(1200 < Z < 1250)} = \frac{0,296203 \cdot 0,7}{0,270032} = 0,767843$$

Da rengøring på mellem 1200 og 1250 tyder på en præstation med pause, giver det god mening at den opdaterede sandsynlighed for at det er en dansker er større end $0,7$.

ALTERNATIV: Igen uden uafhængighed med fx perfekt korrelation:

$$P(D = DK | z \in (1200, 1250)) = \frac{P(1200 < Z < 1250 | D = DK) \cdot P(D = DK)}{P(1200 < Z < 1250)} = \frac{0,156745 \cdot 0,7}{0,15477} = 0,708932$$

Opgave 3

på Gåserød skole blev der gennemført en PISA lignende prøve i 2006. i alt 10 elever i 9. klasse deltog. Den samme prøve blev gentaget tre år senere dvs. i 2009 i alt 14 elever deltog. Resultaterne for de i alt 24 elever er vist i nedenstående tabel.

år	2006	2009
1	455,6	453,5
2	463,0	590,2
3	553,6	541,1
4	488,2	462,8
5	480,6	526,8
6	453,2	487,3
7	555,0	495,6
8	409,4	527,3
9	315,3	297,2
10	483,4	393,9
11		493,3
12		453,1
13		510,3
14		504,8
gennemsnit	465,7	481,2
spredning	69,0	70,6

Der opstilles nu følgende model:

$X_1, ..., X_{10}$ som er uafhængige og hvor $X_i \sim N(\mu_1, \sigma_1^2) \quad i = 1, ..., 10$

$Y_1, ..., Y_{15}$ som er uafhængige og hvor $Y_j \sim N(\mu_2, \sigma_2^2) \quad j = 1, ..., 15$

1. Antag at de to varianser er ens dvs. $\sigma_1^2 = \sigma_2^2 = \sigma^2$.Vis at den fælles varians estimeres til 69,9.

$$\frac{(n_1 - 1) * s_1^2 + (n_1 - 1) * s_1^2}{n_1 + n_2 - 2} = \frac{(10 - 1) * 69,0^2 + (14 - 1) * 70,6^2}{10 + 14 - 2} = 69,9^2$$

(s.498 i B&L)

2. Estimer $\mu_1$og μ_2 og angiv estimaternes egenskaber.

Det bliver de to gennemsnit dvs. $465,7$ og $481,2$. Estimaterne er middelrette (unbiased) og deres varians går mod nul når antallet af observationer går mod uendelig.

Dvs. at de er konsistente.

3. Udregn et 95 % konfidens interval for μ_1.

$$\overline{X}_1 +- t(n-1; 0.975)\frac{S}{\sqrt{n}} = 465,7 +- t(9; 0.975)\frac{69,0}{\sqrt{10}} = 465,7 +- 2,26\frac{69,0}{\sqrt{10}} = 416,4 - 515,0$$

(s. 386 i B&L)

4. Test om der er forskel på de to prøver dvs. test om $\mu_1 = \mu_2$. Alternativ hypotese skal angivers, det naturlige valg er dobbeltsiddet.

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{1/n_1+1/n_2}} = \frac{465,7 - 481,2}{69,9\sqrt{1/10}+1/14} = -0,54$$

dette er t-fordelt med 10+14-2 =22 frihedsgrader. Signifikanssandsynligheden bliver ca. 60%. Vi skal altså ikke forkaste, de to middelværdier er ens.

I den første prøve (i år 2006) er der 2 piger og dermed 8 drenge. I den anden prøve er der 4 piger og dermed 10 drenge. Lad Z_1og Z_2 være antallet af piger der har deltaget i prøverne (1 står for år 2006 og 2 står for 2009)

5. Argumenter for at Z_1og Z_2 er to uafhængige binomialfordelinger og angiv deres parametre.

Der er tale om to udfald. Samme "sandsynlighed" for at det bliver pige og uafhængighed.

6. Test om andelen af piger der har deltaget i en prøve er ens. Dvs. test om $P_1 = P_2$,

hvor P_1 og P_2 er sandsynlighedsparametrene i de to binomialfordelinger. Igen vil det være naturligt at vælge dobbeltsiddet alternativ.

$Z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1+1/n_2)}} = \frac{0,2-0,285}{\sqrt{0,25(1-0,25)(1/10+1/14)}} = -0,48$ som er $N(0,1)$. Signifikanssandsynlighed bliver ca 63%. Testet er dobbeltsiddet.

Hvis man betinger og tester i den hypergeometriske fordeling fås at et "værre" resultat dvs. mindre end 2 bliver ca. 51%. (Dette test er ensiddet)

Nedenstående SAS programmer kan bruges til løsning af opgave 4 og opgave 6.

```
data a;
*** opgave 4***;
input aar score @@;
cards;
2006 455.6 2006 463.0 2006 553.6 2006 488.2 2006 480.6 2006 453.2 2006
555.0 2006 409.4
2006 315.3 2006 483.4
2009 453.5 2009 590.2 2009 541.1 2009 462.8 2009 526.8 2009 487.3 2009
495.6 2009 527.3
2009 297.2 2009 393.9 2009 493.3 2009 453.1 2009 510.3 2009 504.8
;
```

```
proc ttest data=a;
class aar;
var score;
run;

*** Opgave 6***;
data a;
input sex $ aar antal;
cards;
p 1 2
d 1 8
p 2 4
d 2 10
;
proc freq data=a;
table sex*aar/norow nocol nopercent chisq;
weight antal;
run;
```