Handelshøjskolen i København
Erhvervsøkonomi/matematik–studiet

Eksamen i Statistik 2. år

Skriftlig prøve (4 timer)

16. juni 2005 kl. 9.00–13.00

Eksamenssættet er på 3 sider.
Alle skriftlige hjælpemidler samt lommeregner er tilladt.
Vægtfordeling: Opgaverne vægtes ens.

Bemærkning. I det originale eksamenssæt var der en fejl (opdaget af Nicolai Jespersen maj 2007) i opgave 3 (c), hvor teststørrelsen på 8.81 fejlagtigt var angivet til 13.7. Heldigvis uden afgørende konsekvenser for konklusionen.

Opgave 1

Lad X_1 og X_2 være uafhængige stokastiske variable, med samme fordeling på $\{-1, 0, 1\}$ givet ved sandsynlighedsfunktionen

$$p(x) = \begin{cases} 1/4 \text{ for } x = \pm 1\,, \\ 1/2 \text{ for } x = 0\,. \end{cases}$$

Vi sætter

$$S = X_1 + 2X_2\,.$$

(a) Beregn ES og var(S).

(b) Opskriv sandsynlighedsfunktionen for S.

(c) Udregn den betingede sandsynlighed

$$P(X_2 = 0 \mid S = 1)\,.$$

Opgave 2

Lad X_1 og X_2 være uafhængige, ligefordelte på enhedsintervallet. Med $X_{(1)}$ og $X_{(2)}$ betegnes som sædvanligt de ordnede variable, altså

$$(X_{(1)}, X_{(2)}) = \begin{cases} (X_1, X_2) \text{ for } X_1 \le X_2\,, \\ (X_2, X_1) \text{ ellers.} \end{cases}$$

(a) Opskriv fordelingsfunktion og tæthed for $X_{(2)}$.

(b) Hvad er sandsynligheden

$$P\left(X_{(1)} < \frac{1}{2} \text{ og } X_{(2)} > \frac{1}{2}\right)$$

for at de ordnede variable falder på hver sin side af $\frac{1}{2}$?

(c) For $Z = X_{(1)}/X_{(2)}$, opskriv tætheden for $(Z, X_{(2)})$, gør rede for at Z og $X_{(2)}$ er stokastisk uafhængige, og beskriv fordelingen af Z (Vink: Transformationen $(x_1, x_2) \to (z, y) = (x_{(1)}/x_{(2)}, x_{(2)})$ har den egenskab at hvert punkt $(z, y) \in]0, 1[^2$ er billede af netop to punkter).

Opgave 3

For at undersøge om de organiske opløsningsmidler, som malere arbejder med, kan være medvirkende årsag til barnløshed, foretog Malernes Fagforening i København en rundspørge blandt deres mandlige medlemmer i alderen 20–60, suppleret med en tilsvarende kontrolgruppe af jord– og betonarbejdere som normalt ikke er udsat for organiske opløsningsmidler. For de adspurgte, som hævdede at have forsøgt at få børn, blev det registreret om de faktisk havde fået børn. Resultatet var som følger:

Erhverv	Har børn	Ufrivilligt barnløs	I alt
Maler	1306	220	1526
J. og B.	535	54	589
I alt	1841	274	2115

Vi betragter binomialmodellen, hvor 1306 og 535 fortolkes som observationer af uafhængige binomialfordelte variable med hver sin sandsynlighedsparameter og antalsparametre 1526 og 589.

(a) Estimer, med angivelse af approksimative 99% sikkerhedsgrænser, de to sandsynlighedsparametre p_1 og p_2.

(b) Foretag et test for hypotesen $p_1 = p_2$, og drag de relevante konklusioner vedrørende organiske opløsningsmidlers indflydelse på forplantningsevnen.

(c) De adspurgte blev også spurgt om deres alder. Med opdeling i tre passende aldersgrupper kommer tabellens tal til at se sådan ud:

Erhverv	Alder	Har børn	Barnløs
Maler	51-60	250	18
J. og B.	51-60	151	15
Maler	31-50	901	167
J. og B.	31-50	302	30
Maler	21-30	155	35
J. og B.	21-30	82	9

Her ligger det lige for at opfatte de seks antal $250, \ldots, 82$ som observationer af uafhængige binomialfordelte variable med sandsynlighedsparametre p_{ea} ($e = 1, 2$ for erhverv, $a = 1, 2, 3$ for aldersgruppe). Kvotienttestet for reduktion af den fulde model (hvor de seks sandsynlighedsparametre varierer frit) til den logistiske regressionsmodel

$$p_{ea} = \frac{\exp(\gamma + \alpha_e + \beta_a)}{1 + \exp(\gamma + \alpha_e + \beta_a)}$$

fører til en kvotienttesteststørrelse på 5.51. Kvotienttestet for yderligere reduktion af denne model til

$$p_{ea} = \frac{\exp(\gamma + \beta_a)}{1 + \exp(\gamma + \beta_a)}$$

fører til en kvotienttesteststørrelse på 8.81. Hvilke konklusioner kan man drage af dette? Og i hvilke henseender er disse konklusioner mere troværdige end dem vi kunne drage af testet i spørgsmål (b)?

Opgave 4

Nedenstående ses 12 samhørende målinger af vindstyrken x og effekten y (elektricitetsmængde pr. tidsenhed) for en vindmølle. Bemærk at talparrene er ordnet efter voksende værdier af x, den tidsmæssige rækkefølge fremgår ikke af tabellen, og vi skal heller ikke bruge den.

X	Y
2.45	0.123
2.70	0.500
2.90	0.653
3.05	0.558
3.40	1.057
3.60	1.137
3.95	1.144
4.10	1.194
4.60	1.562
5.00	1.582
5.45	1.501
5.80	1.737

(a) Estimer parametrene i en simpel regressionsmodel med x som forklarende variabel og y som respons. For hældningens vedkommende ønskes angivelse af 95% konfidensgrænser. Følgende størrelser kan benyttes:

$2.45 + \cdots + 5.80 = 47.00$,
$2.45^2 + \cdots + 5.80^2 = 197.4400$,
$0.123 + \cdots + 1.737 = 12.748$,
$0.123^2 + \cdots + 1.737^2 = 16.360030$
$2.45 \times 0.123 + \cdots + 5.80 \times 1.737 = 55.69840$.

(b) Estimer, med angivelse af 95% konfidensgrænser, den forventede effekt ved vindstyrke 4.00.

(c) Undersøg ved hjælp af en tegning, om sammenhængen rent faktisk ser ud til at kunne beskrives ved en linie. Suppler med en vurdering baseret på følgende udskrift af parameterestimaterne i en multipel regression, hvor både x og $\log(x)$ er med som forklarende variable.

	Estimate	Std.dev.	T	P
CONSTANT	-2.054	0.3439	-5.971	0.000210
X	-0.650	0.2469	-2.633	0.027243
LOGX	4.260	0.9654	4.413	0.001688