Eksamen i Statistik 2. år

Skriftlig prøve (4 timer)

15. maj 2008 kl. 9.00–13.00

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

Opgave 1

Lad (X, Y) være en stokastisk variable på $\{1, 2\} \times \{1, 2, 3\}$ med sandsynlighedsfunktion $p(x, y)$ givet ved følgende tabel:

y	1	2	3
x			
1	0.1	0.2	0.2
2	0.2	0.1	0.2

(a) Udregn $P(X = 2 \text{ og } Y \leq 2)$.

(b) Opskriv sandsynlighedsfunktionen for den marginale fordeling af Y, og udregn middelværdi og varians i denne fordeling.

(c) Opskriv sandsynlighedsfunktionen for den betingede fordeling af Y, givet $X = 2$, og udregn middelværdi og varians i denne fordeling.

Opgave 2

Lad X_1 og X_2 være uafhængige, normeret eksponentialfordelte, altså med tætheder

$$p_1(x) = p_2(x) = \begin{cases} \exp(-x) \text{ for } x > 0, \\ 0 \text{ ellers.} \end{cases}$$

(a) Udregn middelværdi og varians for $X_1 + 3X_2$.

(b) Udregn sandsynligheden for at X_1 og X_2 enten begge to er < 1 eller begge to er > 1; altså

$$P\big((X_1, X_2) \in \big([0, 1[\times[0, 1[\big) \cup \big(]1, +\infty[\times]1, +\infty[\big)\big) .$$

(c) For hvilke værdier af $a > 0$ har den stokastiske variable $\exp(aX_1)$ en veldefineret middelværdi? Og hvad er middelværdien, når den er defineret?

Opgave 3

I nedenstående antalstabel er 5218 medarbejdere ved en større virksomhed klassificeret efter om de indtager en ledende stilling, samt efter deres svar på spørgsmålet "Hvor tilfreds er du med at være ansat i virksomheden" (her er antallet af svarkategorier reduceret til to, ved passende sammenlægninger).

	Tilfreds	Utilfreds	I alt
Ledere	216	179	395
Ikke–ledere	1899	2924	4823
I alt	2115	3103	5218

Vi betragter den model der fortolker antal tilfredse i de to grupper, 216 og 1899, som udfald af uafhængige, binomialfordelte stokastiske variable med hver sin sandsynlighedsparameter og antalsparametre 395 og 4823.

(a) Estimer de to sandsynlighedsparametre med angivelse af 99% sikkerhedsgrænser.

(b) Foretag et test for, om ledere og ikke–ledere kan antages at have samme grad af tilfredshed.

(c) I undersøgelsen blev endvidere medarbejdernes anciennitet registreret som en gruppering i fire grupper (0–5 år, 6–10 år, 11-20 år og over 20 år). Lad y betegne responsen "tilfredshed", kodet som 1 for tilfreds og 0 for utilfreds, og lad s og a betegne henholdsvis status (leder eller ikke–leder) og anciennitetsgruppe for en medarbejder. Kvotienttesstørrelsen $(-2\log q)$ for reduktion af den logistiske regressionsmodel

$$P(y=1) = \frac{\exp(\gamma + \alpha_s + \beta_a)}{1 + \exp(\gamma + \alpha_s + \beta_a)}$$

til modellen

$$P(y=1) = \frac{\exp(\gamma + \alpha_s)}{1 + \exp(\gamma + \alpha_s)}$$

er udregnet til 76.18. Hvad kan man slutte af dette? Og hvad er relationen til den model, vi så på i spørgsmål (a) og (b)?

Opgave 4

Følgende datasæt er taget fra en ejendomsmægleravis fra september 1999. For de 53 huse, som var udbudt til salg i området Lyngby–Holte–Birkerød–Hørsholm, noteredes blandt andet prisen (i 1000 kr.) og boligens areal i m^2.

PRIS	AREAL	PRIS	AREAL	PRIS	AREAL
1175	66	2271	177	2350	137
3176	228	2400	137	1669	121
1850	114	1999	140	1795	121
2159	145	1575	102	2496	160
1500	91	1673	106	2998	231
2032	106	3550	263	2299	154
3101	170	4203	200	1897	122
2700	198	1625	132	1668	135
1550	87	1895	152	2085	144
1800	126	2795	133	3997	357
3354	166	1295	125	2495	180
3384	158	3500	288	3248	217
2233	138	1975	152	2438	177
1771	118	2595	80	1875	109
2495	160	3400	208	1690	158
2898	148	2798	140	1495	127
1695	104	3493	283	2145	184
1700	124	1950	99		

(a) Idet prisen betegnes y og arealet x betragtes en simpel regressionsmodel med y som respons og x som forklarende variabel. Estimer parametrene i denne model. For hældningens vedkommende ønskes angivelse af 95% sikkerhedsgrænser. Ved udregningerne kan følgende mellemregningsstørrelser benyttes:

$$S_x = 8198 \qquad S_y = 124205$$

$$SS_x = 1429832 \qquad SS_y = 318811200$$

$$SP_{xy} = 20892204$$

(b) Estimer, med angivelse af 95% sikkerhedsgrænser, den forventede pris for et hus i området med et boligareal på 120 m^2.

(c) To andre variable, grundens areal i m^2 og antal soveværelser, er også registreret. En multipel regressionsmodel med disse tre forklarende variable giver anledning til variansanalyseskemaet

ANALYSIS OF VARIANCE TABLE
Square sums and F-tests for removal of terms, last first.

Effect	D.F.	S.S.	M.S.	F	P
CONSTANT	1	291073245.8	291073245.8	545.6716	0.000000
AREAL	1	17452399.6	17452399.6	86.5362	0.000000
GRAREAL	1	194189.7	194189.7	0.9622	0.331366
SOVEV	1	328388.6	328388.6	1.6482	0.205246
RESIDUAL	49	9762969.3	199244.3		
TOTAL	53	318811193.0			

Hvad kan man slutte af det?