Handelshøjskolen i København
Erhvervsøkonomi/matematik–studiet

Reeksamen i Statistik 2. år

Skriftlig prøve (4 timer)

28. juli 2006 kl. 9.00–13.00

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

Opgave 1

Lad (R, S) betegne en stokastisk variabel på $\{1, 2\} \times \{1, 2\}$ med sandsynlighedsfunktion $p(r, s)$ givet ved følgende tabel:

$s =$	1	2
$r = 1$	0.40	0.00
2	0.35	0.25

(a) Udregn $P(R = 1)$, samt den betingede sandsynlighed for hændelsen $R = 1$, givet $S = 1$. Er R og S stokastisk uafhængige?

(b) Udregn middelværdi og varians for R, og opskriv sandsynlighedsfunktionen for R.

(c) Opskriv sandsynlighedsfunktionen for $R + S$, og beskriv den betingede fordeling af R, givet $R + S = 3$.

Opgave 2

Lad X være en stokastisk variabel som er ligefordelt på intervallet [2,4].

(a) Opskriv tætheden for X, og udregn sandsynligheden for hændelsen $X \in [3, 3.5]$.

(b) Udregn middelværdi og varians for X.

(c) Opskriv tætheden for $Y = X^2 - 4$.

Opgave 3

I en kreditforening var der på et givet tidspunkt 8021 låntagere, som var i restance med seneste termin. Nedenstående tabel viser, med opdeling i fire aldersgrupper, hvor mange af disse der fik sat deres ejendom på tvangsauktion i løbet af det følgende år. For hver aldersgruppe g anvender vi betegnelserne

m_g = det samlede antal personer i aldersgruppen,
y_g = antallet af tvangsauktioner blandt disse.

g	y_g	$m_g - y_g$	m_g
1 (–30)	167	1244	1411
2 (30–40)	261	2011	2272
3 (40–50)	262	2612	2874
4 (50–)	93	1371	1464
I alt	783	7238	8021

Vi betragter den statistiske model, der fortolker $y_1, \ldots, y_4$ som observationer af uafhængige binomialfordelte variable med antalsparametre $m_1, \ldots, m_4$ og sandsynlighedsparametre $p_1, \ldots, p_4$, som i første omgang varierer frit.

(a) Estimer, med angivelse af approksimative 95% sikkerhedsgrænser, sandsynlighederne for tvangsauktion i hver af de fire aldersgrupper.

(b) foretag et test for hypotesen $p_1 = p_2 = p_3 = p_4$.

(c) Den model vi betragtede ovenfor kan fortolkes som en logistisk regressionsmodel, når de fire sandsynlighedsparametre skrives på formen

$$p_g = \frac{\exp(\alpha_g)}{1 + \exp(\alpha_g)}$$

hvor de fire parametre $\alpha_g = \text{logit}(p_g)$ varierer frit på hele den reelle akse.

I det oprindelige datasæt forelå data på individniveau, og for hver låntager var den faktiske alder i år oplyst. Det er nærliggende at benytte den faktiske alder i år i stedet for den grovere gruppering der er benyttet ovenfor. Til vurdering af, om sandsynligheden for tvangsauktion kan antages at afhænge logit–lineært af alderen, er følgende to logistiske regressionsmodeller estimeret i det oprindelige datasæt:

$$\text{Model 1:} \qquad p_i = \frac{\exp(\alpha_g + \beta_g a_i)}{1 + \exp(\alpha_g + \beta_g a_i)}$$

$$\text{Model 2:} \qquad p_i = \frac{\exp(\alpha + \beta a_i)}{1 + \exp(\alpha + \beta a_i)}$$

Her betegner p_i sandsynligheden for at person nr. i går på tvangsauktion, a_i er person nr. i's alder, og g den tilsvarende aldersgruppe.

Det oplyses, at testet for reduktion af model 1 til model 2 førte til kvotienttesteststørrelsen $-2 \log q = 7.45$. Hvad kan man konkludere af det?

Opgave 4

Nedenstående datasæt indeholder højderne i cm for i alt 42 kvinder, fordelt på tre aldersgrupper. De 42 kvinder er udtaget tilfældigt blandt dem der indgik i et meget stort datasæt, som er indsamlet i forbindelse med et studium af befolkningens sundhedstilstand (Østerbrundersøgelsen). Datasættet kan ikke opfattes som repræsentativt når det gælder alder (der er for få i gruppe 1), men fordelingen af kvindernes højder inden for aldersgrupperne kan antages at være omtrent som i befolkningen.

Gruppe 1	Gruppe 2	Gruppe 3
163	172 163	162
155	169 159	152
159	164 159	152
172	164 166	162
165	163 162	162
	162 156	164
	169 153	151
	165 166	159
	161 158	156
	157 165	141
	162 151	151
	159 151	
	155 156	

Summer og kvadratsummer af observationer i grupperne er som følger:

	sum	kvadr.sum
Gruppe 1 (5 kvinder under 40)	814	132684
Gruppe 2 (26 kvinder mellem 40 og 60)	4187	675015
Gruppe 3 (11 kvinder over 60)	1712	266936

(a) Estimer parametrene i den ensidede variansanalysemodel, hvor de 42 højder antages at være uafhængige, normalfordelte med samme varians og middelværdi der kun afhænger af aldersgruppen. For middelværdiparametrenes vedkommende ønskes angivelse af 95% sikkerhedsgrænser.

(b) Foretag Bartletts test og test for homogenitet.

(c) Kan det antages at middelværdierne i gruppe 1 og 2 er ens?