Eksamen i Statistik 2. år

Skriftlig prøve (4 timer)

20. maj 2010 kl. 9.00–13.00

Eksamenssættet er på 4 sider.
Alle skriftlige hjælpemidler samt lommeregner er tilladt.
Vægtfordeling: Opgaverne vægtes ens.

Opgave 1

Lad (X_1, X_2, X_3) betegne resultatet af tre kast med en terning. Altså: X_1, X_2 og X_3 er stokastisk uafhængige, ligefordelte på $\{1, 2, \ldots, 6\}$.

(a) Sæt $Y = X_1 - X_2$. Opskriv sandsynlighedsfunktionen for Y, og angiv $\mathrm{E}Y$ og $\mathrm{var}(Y)$.

(b) Hvad er sandsynligheden for at der er både lige og ulige tal blandt de tre tal X_1, X_2 og X_3?

(c) Sæt $S = X_1 + X_2 + X_3$. Hvad er den betingede sandsynlighed for at $X_1 = 3$, givet at $S = 5$?

Opgave 2

Lad $X_1, \ldots, X_{10}$ være uafhængige, rektangulært fordelte stokastiske variable på enhedsintervallet [0,1]. Vi sætter $S = X_1 + \cdots + X_{10}$.

(a) Beregn middelværdi og varians for S.

(b) Hvad er, sådan cirka, sandsynligheden for at $S \geq 6$? Bemærk: Det kræves naturligvis ikke, at man beregner denne sandsynlighed eksakt, det er alt for besværligt. Men ifølge den centrale grænseværdisætning er S approksimativt normalfordelt, og det kan man bruge til at finde en udmærket approksimation til $P(S \geq 6)$.

(c) Ved 10000 uafhængige computer simulationer af en stokastisk variabel som S ovenfor fandt man 1387 værdier som var ≥ 6. Hvordan stemmer det med resultatet fra spørgsmål (b)?

Opgave 3

I en undersøgelse af 8635 firmaers risiko for at gå fallit observerede man om virksomhederne var gået fallit i et givet kalenderår, og hvor mange ansatte der var på virksomheden ved årets begyndelse. Antallet af ansatte er i nedenstående tabel repræsenteret ved en faktor på tre niveauer:

1: Højst 10 ansatte,
2: over 10 men højst 100,
3: over 100.

ANSATTE / FALLIT	-10	11-100	101-	SUM
Ja	224	126	5	355
Nej	5008	2402	870	8280
SUM	5232	2528	875	8635

Til analyse af disse tal betragter vi modellen hvor de tre søjlesummer er givne, og antal fallitter indenfor de tre grupper er uafhængige, binomialfordelte med hver sin sandsynlighedsparameter og den tilsvarende søjlesum som antalsparameter.

(a) For hver af de tre kategorier af firmaer, givet ved faktoren `ANSATTE`, estimer med angivelse af 95% konfidensgrænser sandsynligheden for at et firma går fallit i løbet af det givne kalenderår.

(b) Undersøg ved et statistisk test, om fallitsandsynligheden afhænger af firmaets størrelse, målt i antal ansatte.

(c) Der foreligger også oplysninger om firmaernes egenkapital ved indgangen til det pågældende kalenderår. Lad `LOGEK` betegne $\log(1+E)$, hvor E er firmaets egenkapital. `LOGEK` er altså, stort set, den naturlige logaritme til firmaets egenkapital (1 er lagt til for at undgå at tage logaritmen til 0; bemærk at `LOGEK`=0 svarer til en egenkapital på 0). En model, som er forsøgt anvendt, er den logistiske regressionsmodel

$$P(\text{fallit}) = \frac{\exp(\alpha_a + \beta x)}{1 + \exp(\alpha_a + \beta x)}$$

hvor a er niveauet af faktoren `ANSATTE` og x er værdien af `LOGEK`. Parameterestimaterne i denne model så sådan ud:

	Estimate	Std.dev.	U	P
ANSATTE[1]	-1.575	0.1003	-15.705	0.000000
ANSATTE[2]	-0.932	0.1326	-7.028	0.000000
ANSATTE[3]	-2.346	0.4697	-4.994	0.000001
LOGEK	-0.3063	0.01742	-17.582	0.000000

Hvad er, ifølge denne model, estimatet for fallitsandsynligheden for et firma som har 11–100 ansatte og en egenkapital på 0?

Opgave 4

For 47 schweiziske amter har man omkring 1888 registreret følgende demografiske oplysninger. Alle oplysningerne er, i passende forstand, angivet i procent:

`FER`: Fertiliteten (altså et passende mål for antal børn pr. indbygger eller pr. kvinde i den fødedygtige alder; hvad tallet helt præcist dækker over er ikke opgivet).

`AGRI`: Andel af den mandlige befolkning, som er beskæftiget ved landbruget.

`EXAM`: Andelen af værnepligtige, som har fået højeste karakter ved en bestemt militær prøve.

`EDU`: Andelen af værnepligtige, som har modtaget uddannelse ud over almindelig skolegang.

`CATH`: Andelen af katolikker i befolkningen.

`INFM`: Børnedødeligheden, opgjort som den andel af de levendefødte børn der dør inden de er fyldt et år.

I denne opgave betragtes nogle statistiske modeller til forklaring af børnedødeligheden som funktion af de øvrige variable. I første omgang ser vi kun på én forklarende variabel, nemlig `FER`, som i det følgende betegnes x, medens `INFM` betegnes y. De 47 samhørende værdier af x og y ser sådan ud:

X	Y	X	Y	X	Y	X	Y
80	22	67	19	73	21	92	16
83	22	69	23	74	24	79	18
93	20	62	19	72	18	70	20
86	20	68	21	61	16	66	21
77	21	72	20	58	21	73	19
76	27	56	20	65	23	64	23
84	24	54	11	76	15	78	20
92	25	65	20	69	20	68	20
82	21	66	18	77	18	35	18
83	24	65	22	71	19	45	18
87	25	57	17	79	20	43	19
64	17	57	15	65	18		

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

$S_x = 3298 \quad S_y = 938 \quad SS_x = 238560 \quad SS_y = 19126 \quad SP_{xy} = 66505$

(b) For det første af de 47 amter ($x_1 = 80$, $y_1 = 22$), estimer den forventede børnedødelighed med angivelse af 95% sikkerhedsgrænser.

(c) I en multipel regressionsmodel, som inddrager alle de nævnte variable, fås variansanalyseskema og parameterestimater, der ser ud som følger:

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	18720.08511	18720.08511	2121.4396	0.000000
FER	1	65.79567	65.79567	8.7052	0.005022
AGRI	1	21.04927	21.04927	2.9027	0.095485
EXAM	1	4.30554	4.30554	0.5882	0.447314
EDU	1	4.20224	4.20224	0.5683	0.455139
CATH	1	0.04168	0.04168	0.0055	0.941228
RESIDUAL	41	310.52050	7.57367		
TOTAL	47	19126.00000			

	Estimate	Std.dev.	T	P
CONSTANT	8.39	5.618	1.493	0.143047
FER	0.1515	0.05515	2.747	0.008905
AGRI	-0.0120	0.02888	-0.416	0.679552
EXAM	0.0589	0.09870	0.597	0.554007
EDU	0.0486	0.08687	0.559	0.578990
CATH	0.0011	0.01483	0.074	0.941228

Hvad kan man slutte af det?