UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I:	**ST 110 – Statistiske metodar og dataanalyse**
TID FOR EKSAMEN:	**Tirsdag 1. juni 2004, kl. 9:00–15:00**
HJELPEMIDLER:	**Formelsamlingar (ST 110 og Rottmann), kalkulator**

Dette oppgåvesettet inneheld fire oppgåver og er på til saman seks sider, inkludert eit ein-sides appendiks med opplysningar du kan få bruk for.

Oppgåve 1

AV OG TIL KRYSSER VERDA ORIGO, og vi skal her sjå på eit par aspekt ved den lineære regresjonsmodellen der ein veit at regresjonslina går gjennom null. Forutsett derfor at

$$Y_i = bx_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

der støyledda ε_i er uavhengige og normale $\mathrm{N}(0, \sigma^2)$. Vi tenkjer oss at x_i-ane er kjende, positive tal som er under kontroll av eksperimentator. Spesielt er dermed $Y_i \sim \mathrm{N}(bx_i, \sigma^2)$. For å forenkla eit par av problema under (ved dette høve) skal vi forutsetta at σ *er eit kjend tal.*

(a) Vis at minste-kvadratsum-estimatoren for b vert

$$\widehat{b} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

(b) Finn venteverdi og varians for $\widehat{b}$.

(c) Vis at $\widehat{b}$ er ein såkalla Markov-estimator («uniformly minimum variance unbiased»-eigenskapen).

(d) Ein er interessert i å teste hypotesen $H_0: b = 4.0$ mot alternativet $b > 4.0$. Finn fram til sannsynskvotetesten («the generalised likelihood ratio test») for H_0 med nivå 5%. (Kom i hug at σ er eit kjend tal. Dersom du ikkje kjem i mål med sannsynskvotetesten kan du setja opp ein rimeleg test «direkte».)

(e) Finn eit uttrykk for styrkefunksjonen for testen i (d).

(f) Korleis må eksperimentet planleggast, med omsyn til $x_1, \ldots, x_n$, for å oppnå at sjansen er minst 95% for å forkasta $b = 4.0$, om den verkelege verdien er $b = 5.5$?

(g) Du er statistikar A, og har gjort analysen over, og du har nytta dei eksisterande omstende som verkeleg sikrar at regresjonslina går gjennom origo. Ein annan statistikar B vel likevel, grunna konservativitet eller ignorans, å nytta den vanlege regresjonsmodellen med konstantledd, altså den der Y_i-ane er uavhengige og normale med same varians og $\mathrm{E}\, Y_i = a + bx_i$. Diskuter kort korleis B's analyse kjem til å verka samanlikna med din. Relater gjerne diskusjonen til punkta over.

Oppgåve 2

Den østerrikske legen Karl Landsteiner fikk Nobelprisen i medisin for sitt banebrytande arbeid om blodtypa hos menneske, framlagt i år 1900. Dette var første gang AB0-systemet for blodtypar hos mennesket vart brukt. Dette systemet er betinga av visse blodsubstansar eller antigener «a» og «b». Det er fire blodgrupper, A, B, AB, 0, der blodgruppe A svarer til at «a» er til stades, blodgruppe B svarer til at «b» er til stades, blodgruppe AB svarer til at både «a» og «b» er til stades, medan til slutt blodgruppe 0 svarer til at verken «a» eller «b» er til stades.

Landsteiner og andre var tidleg klar over at blodgruppene hos mennesket følgde sannsynslovane til arvelighetslæra, men det var uklart nøyaktig kva for mekanismar som låg under. Inntil 1924 var det to ulike teorier om dette. Hypotesa til den første teorien handla om to alleler ved kvart av to gen-loci. Den konkurrerande teorien bygde i staden på at det var tre alleler for eit enkelt gen-locus som kontrollerte blodgruppa.

La θ_{A}, θ_{B}, θ_{AB}, θ_0 vere sannsyna for at eit tilfeldig individ i ein stor og genetisk stabil populasjon skal ha blodgruppe respektive A, B, AB, 0. Det er i og for seg interessant å gå djupare med ein-locus- og to-loci-teoriane og deretter utarbeide formlar for desse frekvensane basert på dei to teoriane, og desse sannsynsutrekningane er heller ikkje så vanskelege for ein student med ST 110 bak seg. Men sidan dette er ein eksamen og tempus fugit skal det her berre settas fast at to-loci-teorien fører til formlane

$$\theta_{\mathrm{A}} = a(1-b), \qquad \theta_{\mathrm{B}} = (1-a)b, \qquad \theta_{\mathrm{AB}} = ab, \qquad \theta_0 = (1-a)(1-b),$$

uttrykt ved visse sannsynsparametre a og b, der $0 < a < 1$ og $0 < b < 1$; mens ein-locus-teorien fører med seg formlane

$$\theta_{\mathrm{A}} = p(2-p-2q), \qquad \theta_{\mathrm{B}} = q(2-2p-q), \qquad \theta_{\mathrm{AB}} = 2pq, \qquad \theta_0 = (1-p-q)^2,$$

uttrykt ved visse andre sannsynsparametre p og q, der $p > 0$, $q > 0$, og $p + q < 1$.

Disputten vart i praksis avgjort ved F. Bernsteins artikkel «*Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstrukturen des Menschen*» frå 1924. Han brukte eit datamateriale samla inn av Kirihara om blodgruppene for 502 japanarar som budde i Korea. Av desse hadde 212 type A, 103 type B, 39 type AB, medan 148 hadde type 0.

(a) Presiser dei forutsetningar som må gjerast for å kunne sjå desse frekvensane 212, 103, 39, 148 som utfall av ein multinomisk fordelt vektor [sjå eventuelt appendikset under]. Gå ut frå i punktene under at desse forutsetningane er oppfylt.

(b) I dette og det neste punktet skal du sjå på truverdet av to-loci-teorien i lys av dette materialet. Vis at sannsynsmaksimeringsestimata («maximum likelihood»-estimata) for a og b vert 0.500 og 0.283.

(c) Finn dei «forventa frekvensane» under to-loci-hypotesen (antall enkeltforsøk i den multinomiske modell, multiplisert med estimerte sannsyn). Test så to-loci-teorien.

(d) Så skal du vurdere kor god ein-locus-teorien er. Skriv opp eit eksplisitt uttrykk for likelihoodfunksjonen $L(p, q)$. Ved bruk av passande numeriske algoritme kan ein berekne at denne maksimerast for $(p, q) = (0.295, 0.154)$. Finn dei forventa frekvensane under ein-locus-hypotesen og test kor god denne teorien er.

Oppgåve 3

HELD VI UT Å LEVA MED MINDRE RISIKO og lågare samfunnsmessige omkostningar? Av og til ser det som kjend ikkje slik ut – det kan verka betre for dei fleste å halda fram med å leve under dei eksisterande tilhøva, så lenge risikoen og omkostningane er tynt spredd på «resten av samfunnet», eller over ei lang framtid. Eit døme er trafikkulykker, der Noreg ser ut til å leva godt med fleire hundre dødsfall og eit svært høgt antall millioner kroner i etterutgifter i året, sjølv om ein rasjonelt kan forstå at desse tala kunne bli dramatisk reduserte dersom alle bilhastigheter ble t.d. halvert.

La desse tankar vera bakgrunn for fylgjande forenkla og stiliserte statistiske beslutningsproblem. Ein bestemt type ulykker skjer med ein viss rate θ pr. månad. Desse ulykkene er av ein slik art at dess fleire som skjer, dess større kostnad pr. ulykke. Ein kan investera ein passande stor sum c, som ein kan vente leier til lågare ulykkesrate, samt arten og vidare samfunnskostnader av desse. Altså skal mitt tenkte byråd velge mellom beslutningene d_0, som er å halda fram nøyaktig som før, og d_1, som er denne profylaktiske investeringa. Ein tenker seg at tapet assosiert med dei to ulike beslutningane, samla over eitt komande kalenderår, er av forma

$$L(\theta, d_0) = k\theta^2,$$
$$L(\theta, d_1) = c + \tfrac{1}{10}\, k\theta = 10\, k + \tfrac{1}{10}\, k\theta,$$

der kostnaden k, på same vis som prislappen $c = 10\, k$, er berekna på førehand.

(a) Du er den engasjerte statistiker for byrådet, og oversett den relevante fagkunnskapen til at θ har en Gamma-fordeling med parametre $(3.0, 1.5)$ [sjå eventuelt appendikset under]. Finn formlar for risikoen (forventa tap) for beslutningane d_0 og d_1. Kva for ein beslutning bør ein ta?

(b) Uakta visdomen i råda dine under punkt (a) fell det seg slik at byrådet ikkje nytta seg av denne investeringa. Byen lever altså vidare med tapsfunksjon $k\theta^2$ – og ulykker i antall $X_1, \ldots, X_{24}$ kjem i tillegg, i ein periode på to år. Desse er for den gitte θ uavhengige og Poisson-fordelte med θ som parameter, og ein observerar $\sum_{i=1}^{24} X_i = 90$. Kva vert no din (justerte) oppfatning av θ?

(c) For å visa kva θ betyr vel du å leggja fram for byrådet eit truverdighetsintervall $[L, H]$, som du med ein garanti på 90 prosent meiner dekkjer den underligjande rateparameter. Forklar korleis du kan berekna $[L, H]$.

(d) Saka kjem opp for byrådet att, i lys av ulykkene $X_1, \ldots, X_{24}$. Kva vert di anbefaling til byrådet?

Oppgåve 4

KANDIDATANE VERT FØRST GJORT MERKSAME PÅ at svar på denne oppgåva kan gjerast mykje kortare enn oppgåveteksten! Men for meg som forfattar den forklarende teksten er det altså naudsynt med nokre utfyllande opplysningar, for å klarlegge nødvendige faktorar ved saka.

Juliet B. Schor er professor i Women's Studies ved Harvard-universitetet, og har publisert fleire bøker som har nådd ut til ålmenta, m.a. *The Overworked American: The Unexpected Decline of Leisure* frå 1992. Denne oppgåva skal nærma seg hennes hovudtese og hovudbudskap fra 1998-boka *The Overspent American: Why We Want What We Don't Need*. Denne hovudbudskapen er kort sagt at fleire og fleire amerikanarar, på alle samfunnsnivå, meir og meir lar seg lokka til å kjøpe enda meir og enda meir, dette for å henge med på naboers og bekjentes si materielle utvikling (altså «to keep up with the Joneses»). Professor Schor er ikkje snauare enn at ho også lanserer sitt eige tipunktsprogram («Stopping the upward creep of desire») for å trekke USA ut or krisa.

Mykje av bakgrunnen for heile boka og bodskapen hennar (og med det også for planen hennar for å redda USA frå det grusomme grepet til kjøpeeskaleringsspøkelset) er eit par statistiske undersøkjingar. I ein av desse registrerte ein

$$Y = \text{annual household saving},$$

saman med ei rekke kovariatar, for $n = 834$ amerikanarar i eit spesielt populasjonssegment. Eg skal ikkje her gi den presise definisjonen av Y, men den er altså det årlege sparebeløpet for husstanden, i US dollar, med visse modifikasjonar som m.a. tek omsyn til alder og gjeld.

Ein av Schors viktigaste bodskap er knytta til spesialkovariaten x_{13}, som er den sjølopplevde finansielle status for individet sammenlikna med den eigne (og like sjølopplevde) referansegruppa. Kvar person vert, etter å ha peika ut si naturlege referansegruppe (den gruppe av menneske ein mest naturleg sammenliknar seg med), stilt spørsmålet: «Korleis er din eigen finansielle status, sammenlikna med dei fleste av di referansegruppe?» Skalaen for x_{13} er da

$$x_{13} = \begin{cases} 1 & \text{much worse,} \\ 2 & \text{worse,} \\ 3 & \text{same,} \\ 4 & \text{better,} \\ 5 & \text{much better,} \end{cases}$$

der det dessutan var flest forekomster av «much worse» og «worse».

Alt dette resulterte i ein multippel regresjonsanalyse som (mellom anna) gav desse resultata. Tabellen under gjev for kvar av tretten veldefinerte kovariatar den estimerte regresjonskoeffisient b_j, standard error (estimert standardavvik) SE = SE(b_j) for denne, og dei t-brøkane $t_j = b_j/\mathrm{SE}(b_j)$ som høyrar til. (Tabellen gir også b_0, SE og t_0 for konstantleddet i regresjonslikninga. Skalaen for b_0 er US dollar.)

	b	SE	tratio	covariate
0	19995.000	15263.360	1.31	constant term
1	0.112	0.025	4.48	household income
2	0.025	0.050	0.50	permanent household income
3	0.016	0.019	0.84	household net worth
4	-3763.000	2187.791	-1.72	sex (1 = male, 2 = female)
5	-9916.000	6655.034	-1.49	age
6	1140.000	942.149	1.21	age-squared
7	-204.000	1700.000	-0.12	race
8	-174.000	424.390	-0.41	occupation
9	-1448.000	1366.038	-1.06	educational level
10	-1232.000	367.761	-3.35	number of dependents
11	629.000	767.073	0.82	satisfaction with income
12	-208.000	72.474	-2.87	hours per week watching TV
13	2963.000	773.629	3.83	relative financial status

(a) Rekn ut eit tilnærma 95%-grads konfidensintervall for regresjonskoeffisienten β_{12}, og kommenter kort eventuelle forutsetningar du har gjort.

(b) Har antall timar familien ser på fjernsyn nokon signifikant konsekvens for Y (annual household saving)? Sett opp den relevante nullhypotesa med alternativ, gjer ein test, og formuler ein konklusjon.

Appendiks

1. Ein seier at U er Gamma-fordelt med parametre (a, b) dersom dens tettleik er på forma

$$g(u) = \frac{b^a}{\Gamma(a)} u^{a-1} e^{-bu} \quad \text{for } u > 0.$$

Dens venteverdi er a/b medan variansen er a/b^2.

2. Korrekt bøying av *locus*:

	singularis	*pluralis*
nominativ	locus	loci
vokativ	loce	loci
akkusativ	locum	locos
genitiv	loci	locorum
dativ	loco	locis
ablativ	loco	locis

3. Nokre normalkvantilar: dersom $\Phi(u) = \Pr\{\mathrm{N}(0,1) \leq u\}$, har ein $\Phi(1.282) = 0.90$, $\Phi(1.645) = 0.95$, $= \Phi(1.960) = 0.975$, og $\Phi(2.326) = 0.99$.

4. Den multinomiske modellen:

$$\Pr\{Y_1 = y_1, \ldots, Y_k = y_k\} = \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k},$$

der $p_1, \ldots, p_k$ er sannsyn med sum 1 og $y_1, \ldots, y_k$ er ikkjenegative tal med sum n. I denne modellen har ein

$$\mathrm{E}\, Y_i = np_i, \quad \mathrm{Var}\, Y_i = np_i(1-p_i), \quad \mathrm{cov}(Y_i, Y_j) = -np_ip_j \text{ for } i \neq j.$$