STK 1110 høsten 2004
Oblig II

Dette er oppgavesettet for andre obligatoriske innleveringsprosjekt for kurset STK 1110. Det legges ut **onsdag 3/xi/4**, og en bunke blir også tatt med på forelesningene onsdag og fredag den uken. Leveringsfrist er **fredag 12/xi/4**, til instituttkontoret ved Matematisk institutt, senest kl. 14:30.

Du kan levere håndskrevet eller tekstbehandlet besvarelse. Skriv navnet ditt øverst til høyre på første side. Der du bruker MATLAB eller MINITAB eller `R` må utskrifter og plott legges ved, eller klebes inn i besvarelsen.

Oppgave 1

ER DIN STATISTISKE VERKTØYKASSE i orden? La oss sjekke.

(a) Definer

$$f(x,\gamma) = \begin{cases} 1+\gamma(x-\frac{1}{2}) & \text{for } x \in [0,1], \\ 0 & \text{for } x \text{ utenfor } [0,1]. \end{cases}$$

For hvilke verdier av parameteren γ er dette en sannsynlighetstetthet? Og hva blir forventningen til en X som stammer fra f?

(b) La X være standard-eksponensialfordelt, det vil si med sannsynlighetstettheten

$$g(x) = \begin{cases} e^{-x} & \text{for } x \geq 0, \\ 0 & \text{ellers.} \end{cases}$$

Finn den momentgenerende funksjon for X, og bruk dette til å finne forventingsverdien.

(c) Visse uavhengige målinger $X_1, \ldots, X_{17}$ stammer fra $\mathrm{N}(\mu, \sigma^2)$-fordelingen. Man beregner

$$\sum_{i=1}^{17} X_i = 175.450 \quad \text{og} \quad \sum_{i=1}^{17} X_i^2 = 1821.645.$$

Finn empirisk gjennomsnitt og standardavvik. Lag dessuten 90%-konfidensintervall for hver av parametrene μ og σ.

(d) I situasjonen over, hva er forventningen til variablene

$$U = \sum_{i=1}^{17} \frac{(X_i-\mu)^2}{\sigma^2} \quad \text{og} \quad V = \sum_{i=1}^{17} \frac{(X_i-\widehat{\mu})^2}{\sigma^2},$$

der $\widehat{\mu}$ er maximum likelihood-estimatoren for μ?

(e) En stokastisk variabel Y har tettheten

$$h(y) = \begin{cases} e^{-y^3} 3y^2 & \text{for } y \geq 0, \\ 0 & \text{ellers.} \end{cases}$$

Finn den kumulative fordelingsfunksjonen $H(y)$ for Y. Vis at $W = Y^3$ har en standard-eksponensiell fordeling.

Oppgave 2

«If that dice is showing a one, I will go downstairs and rape Arlene», bestemmer hovedpersonen i Luke Rhineharts *The Dice Man* seg for. Riktig så dramatisk tilnærming behøver vi ikke ha til denne oppgaven, men vi skal teste om terningen jeg kaster med er «normal» eller ikke, ut fra antall kast jeg trenger for å få en éner første gang. Dette eksperimentet gjentar jeg 15 ganger, resulterende i antall nødvendige forsøk

```
2 2 2 10 7 1 3 1 2 12 2 1 2 5 1
```

Hver av disse modelleres altså som kommende fra fordelingen

$$P\{X_i = x\} = (1-p)^{x-1}p \quad \text{for } x = 1, 2, 3, \ldots,$$

der p er sannsynligheten for å få en éner med min terning. Dette er den geometriske fordelingen, med

$$\mathrm{E}X_i = \frac{1}{p} \quad \text{og} \quad \mathrm{Var}\, X_i = \frac{1-p}{p^2}.$$

Dette kan vises med noe analyse av rekker, hvilket er innenfor STK 1100–1110-pensum, men det skal ikke vises ved denne anledning.

(a) Finn maximum likelihood-estimatet $\widehat{p}$ for p, basert på dataene over.

(b) Finn et konfidensintervall for p med konfidensgrad tilnærmet 90%.

(c) Bruk den generaliserte likelihood ratio-test (GLR) til å teste om $p = 1/6$ eller ikke. Formuler en konklusjon.

Oppgave 3

Verden står ikke helt stille, og mange objekter er i vedvarende forandring. Det er mange eksempler fra anvendte vitenskaper der man studerer hvordan visse objekter forandrer seg over tid, med hensyn til geometriske karakteristika, utforming, størrelse, etc., når de blir utsatt for forskjellige prøvelser. Det kan vises til ulike anvendelser i kjemi, biologi, botanikk og geologi, der stokastisk modellering av geometriske objekters utvikling har spilt en betydelig rolle.

Her skal vi se på en ganske enkel modell for utvikling av *firkanter*, som utsettes for jevnlige «sjokk». Man starter med enhetskvadratet $F_0 = [0,1] \times [0,1]$ ved tid 0. Ved tid 1 utsettes firkanten for en prosess som bringer den til tilstanden $F_1 = [0, U_1] \times [0, V_1]$, der U_1 og V_1 er uavhengige og uniformt fordelte over $[0, c]$, der $c = 1.5$. Ved tid 2 kommer et nytt sjokk, representert ved nye skaleringer U_2 og V_2, slik at firkanten nå er blitt $F_2 = [0, U_1U_2] \times [0, V_1V_2]$. Slik fortsetter prosessen, slik at den opprinnelige firkanten etter n tidsenheter er blitt

$$F_n = [0, X_n] \times [0, Y_n], \quad \text{der } X_n = U_1 \cdots U_n \quad \text{og} \quad Y_n = V_1 \cdots V_n.$$

Det antas at alle $U_1, U_2, \ldots, V_1, V_2, \ldots$ er uavhengige fra den uniforme fordelingen over $[0, c]$, med den samme konstante skaleringsfaktor $c = 1.5$.

(a) Vis at U_i, og dermed V_i, har forventning $\frac{1}{2}c = 0.75$. Finn også deres varians. Bruk dette til å finne forventning for sidelengdene X_n og Y_n. Finn dessuten formler for forventning og varians for arealet A_n av firkanten etter n sjokk.

(b) Hva er sannsynligheten for at firkant nr. n er større enn nr. $n-1$, altså at F_n inneholder F_{n-1} som en delmengde?

(c) Hva er sannsynligheten for at firkant nr. n har et større areal enn nr. $n-1$? Finn en approksimasjon til den eksakte sannsynligheten ved simulering. (Det er fint om du også finner svaret ved en eksakt beregning.)

(d) Vi følger nå firkanten gjennom de $n = 4$ første sjokk, og spør først om sannsynligheten for at firkanten har blitt større hver av disse gangene, altså at

$$F_0 \subset F_1 \subset F_2 \subset F_3 \subset F_4.$$

Beregn denne. Det er imidlertid mer sannsynlig at firkantene avtar i størrelse; beregn sannsynligheten for at en firkant har blitt mindre ved hvert av de første $n = 4$ sjokk, altså at

$$F_0 \supset F_1 \supset F_2 \supset F_3 \supset F_4.$$

(e) Firkantene har altså en tendens til å bli mindre og mindre. Finn sannsynligheten

$$q_n = \mathrm{P}\{F_n \text{ er større enn } [0, \tfrac{1}{2}] \times [0, \tfrac{1}{2}]\},$$

igjen for $n = 4$, ved simulering. (Igjen kan du prøve å finne sannsynligheten eksakt, men det er ikke nødvendig.)

(f) Prøv å vise at arealet A_n til F_n virkelig går mot null, med sannsynlighet 1. – Parameteren c styrer hvor raskt firkantene sendes mot sin endelikt. Du må gjerne vise at om $c > e = 2.71828...$, så vil firkantprosessen tvert i mot eksplodere, dvs. at firkantenes areal vokser mot uendelig når n vokser.

Oppgave 4

Vi skal frem til noe så spektakulært som gutter & jenter, men vi trenger noen innledningsrunder før vi kommer så langt.

(a) Vi sier at X er Beta-fordelt med parametre (a, b) hvis dens tetthet er

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1} \quad \text{for } 0 < x < 1.$$

Vis at

$$\mathrm{E}X = \frac{a}{k} \quad \text{og} \quad \mathrm{Var}\, X = \frac{1}{k+1}\frac{a}{k}\frac{b}{k},$$

der $k = a + b$.

[Her er altså $\Gamma(\cdot)$ den såkalte gamma-funksjonen, definert ved $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}\,\mathrm{d}u$. Man har $\Gamma(a+1) = a\Gamma(a)$ for alle a, og for a heltall er $\Gamma(a) = (a-1)!$.]

(b) Vi skal se på «utblandede binomiske fordelinger». Anta at Y er binomisk (m, p) for gitt p men at p selv kommer fra en fordeling over $(0, 1)$ med forventning p_0 og varians σ_0^2. Vis at

$$\mathrm{E}Y = mp_0 \quad \text{og} \quad \mathrm{Var}\, Y = mp_0(1-p_0) + m(m-1)\sigma_0^2.$$

Skriv opp formler for forventning og varians for Y i det tilfelle at den underliggende p kommer fra en Beta-fordeling med parametre $(kp_0, k(1-p_0))$.
[Her passer det å anvende reglene for «dobbeltforventning», se teoremene i bokens Section 4.4.]

(c) I denne situasjonen, med en Beta-fordeling som over for p og en Y som for den gitte p er binomisk (m, p), vis at

$$\Pr\{Y = y\} = \binom{m}{y} \frac{\Gamma(k)}{\Gamma(kp_0)\,\Gamma(k(1-p_0))} \frac{\Gamma(kp_0+y)\,\Gamma(k(1-p_0)+m-y)}{\Gamma(k+m)}$$

for $y = 0, 1, \ldots, m$. Hva er grensen for denne fordelingen når $k \to \infty$?

(d) På slutten av 1800-tallet samlet hr. Geißler i Sachsen inn opplysninger om antallet piker og gutter i (svært mange) forskjellige familier, i tillegg til annen informasjon. Her henter jeg frem data for antallet jenter av de første åtte barn, for i alt 38,495 familier. Dataene er gitt under, på tabellform; det er altså 264 familier med 0 piker av 8, 1655 familier med 1 pike av 8, osv. – La nå Y_i være antallet jenter blant de første åtte barn, for familie nr. i. Finn gjennomsnitt $\bar{y}$ og empirisk standardavvik s for disse Y_i-ene. [Svarene skal bli $\bar{y} = 3.8750$ og $s = 1.4698$.]

8-barnsfamilier		**binomisk modell**		
# piker	*observert*	*predikert*	*obs.–pred.*	*residual*
0	264	192.3	71.7	5.17
1	1655	1445.4	209.6	5.51
2	4948	4752.4	195.6	2.84
3	8498	8928.9	−430.9	−4.56
4	10263	10485.0	−222.0	−2.17
5	7603	7879.8	−276.8	−3.12
6	3951	3701.2	249.8	4.11
7	1152	993.4	158.6	5.03
8	161	116.7	44.3	4.11
	38495	38495.0	$\chi^2 = 159.4\,(\mathrm{df} = 7)$	

TABELL *med analyse av pike-gutt-ratio for 38,495 familier med åtte eller flere barn. Her er «residual» lik (observed – predicted)/(predicted)$^{1/2}$, slik at Pearsons χ^2-observator er summen av de kvadrerte residualer.*

(e) Vi skal først se på den klassiske, rent binomiske forklaringsmodellen: alle fødsler er uavhengige av hverandre, og hver gang er sannsynligheten p_0 for at det blir en pike. Da er altså Y_i-ene uavhengige og binomisk $(8, p_0)$. Estimer p_0 fra data [svaret blir 0.4844, med en komplementær sannsynlighet 0.5156 for gutter]. Beregn så de teoretiske (eller forventede) antall familier med $0, 1, 2, \ldots, 8$ piker, under den binomiske forutsetningen; dette blir det jeg har kalt «predikert» i tabellen. Beregn også

$$\text{residualene} = \frac{O_j - E_j}{\sqrt{E_j}},$$

og endelig også Pearson-observatoren [se bokens Section 8.2 og 9.6]. Forklar hvorfor dette viser at den binomiske modellen blir for enkel. På hvilken eller hvilke måter ser den binomiske modellen ikke ut til å være god nok?

(f) Sammenlign den observerte varians s^2 med den binomiske variansen, og kommenter det du finner.

(g) Siden den binomiske forklaringen ikke er god nok, kaster vi oss ut i en videre modell. Jeg tenker meg at Skaperen deler ut en sannsynlighet p til hver kvinne (og hennes mann), og at kvinnens fødsler så er binomisk med denne p som sannsynlighet for å få jente. Anta spesifikt at p-ene kommer fra en Beta-fordeling med parametre $(kp_0, k(1-p_0))$, som i (a)–(c) over. Estimer p_0 og k ved hjelp av resultatene fra (b). [Jeg finner $\widehat{k} = 85.1961$, $\widehat{p}_0 = 0.4844$.]

(h) Hvis denne modellen stemmer, omtrent hvor mange kvinner har sin jente-sannsynlighet p utenfor $[0.40, 0.60]$?

(i) Til slutt skal du prøve å teste denne beta-binomiske teorien ved hjelp av Pearson-residualer og Pearson-testen. [Da vil det bli nødvendig å beregne gamma-funksjonen for ulike verdier. Denne ligger i MATLAB som `gamma(x)`. Det er noe mer praktisk å foreta beregningene logaritmisk, og `gammaln(x)` i MATLAB gir $\log(\Gamma(x))$. Beregningene kan også foretas i MINITAB eller SPLUS.]

Nils Lid Hjort