DATALOGI A
Projekt E

Dette projekt er det sidste af fem obligatoriske projekter der tilsammen udgør eksamen i Datalogi A. Projektet besvares i grupper på 2; eventuelt efter foregående aftale med instruktor i grupper på 1 eller 3.

Projektet afleveres senest fredag den **18. december 1997** i E 401 (Matematik-Økonomi sekretariatet) eller i postkassen overfor. Projektkvitteringsark skal vedlægges. Senest den **10. januar 1998** modtager hver enkelt studerende en email om hvorvidt projektet er godkendt. Samtidig oplyses om det registrerede pointtal for den studerende, og om hvorvidt kurset er bestået eller ej. Det påhviler den studerende at kontrollere disse oplysninger og at gøre eventuelle indsigelser til instruktor eller Søren Eilers inden den **27. januar 1998**. Dette er også sidste afleveringsdato for eventuel genaflevering af ikke godkendte projekter.

Baggrund

Vi er interesserede i at undersøge om en bestemt genetisk konfiguration **A** disponerer for rygning. Til det formål indføres betegnelserne:

p:	Sandsynligheden for at man har konfiguration **A**
q:	Sandsynligheden for at ryge
p_1:	Sandsynligheden for at ryge hvis man ikke har **A**
p_2:	Sandsynligheden for at ryge hvis man har **A**

Disse størrelser er delvist sammenvævede: Hvis man kender tre af dem, kan man udregne den fjerde. For eksempel er

$$q = (1-p)\,p_1 + p\,p_2. \tag{1}$$

Det spørgsmål vi interesserer os for er hvordan p_1 og p_2 forholder sig til hinanden – og til q. Man kan fx se af (1) at hvis $p_1 = p_2$, så gælder der også at $p_1 = p_2 = q$.

I medicinsk litteratur udregner man ofte *odds-ratio* (OR) ved

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Vi vil studere logaritmen af OR, og betegner denne med $\ln OR$. Hvis denne størrelse er tæt ved 0, så er $p_1 \simeq p_2$. Hvis $\ln OR$ er positiv, så er p_1 større end p_2. Og hvis $\ln OR$ er negativ, så er p_1 mindre end p_2.

Grunden til at man foretrækker $\ln OR$ fremfor p_1 og p_2 i sig selv, er at dette ene tal på en symmetrisk måde opsummerer situationen. Specielt er det sådan at fortolkningen af $\ln OR$ ikke afhænger af hvor store sandsynlighederne p og q er. En god tommelfingerregel er at hvis $\ln OR$ er større end 3 (eller mindre end -3), så er forskellen mellem **A**-gruppen og ikke-**A**-gruppen så markant at eksperterne på området godt er klar over den, også selv om p måske er meget lille.

At spørge om **A** har noget med rygning at gøre er altså at spørge om $\ln OR$ er nul eller ej. Og vi undersøger dette spørgsmål ved at teste 1000 tilfældigt udvalgte mennesker for konfiguration **A**, og sammenholde det med oplysninger om deres rygevaner. Vi får dermed en tabel som herunder:

	Ikke-ryger	ryger	I alt
Ikke **A**	n_{11}	n_{12}	$n_{1\bullet}$
A	n_{21}	n_{22}	$n_{2\bullet}$
Ialt	$n_{\bullet 1}$	$n_{\bullet 2}$	1000

Tabel 1: Undersøgelse af 1000 personer

Herudfra estimerer man

$$\widehat{p} = \frac{n_{2\bullet}}{1000}, \quad \widehat{q} = \frac{n_{\bullet 2}}{1000}, \quad \widehat{p}_1 = \frac{n_{12}}{n_{1\bullet}}, \quad \widehat{p}_2 = \frac{n_{22}}{n_{2\bullet}}$$

og dermed

$$\widehat{OR} = \frac{n_{12} \cdot n_{21}}{n_{11} \cdot n_{22}}.$$

Der må bag sammenhængen mellem **A** og rygning ligge en *sand* værdi af $\ln OR$. Imidlertid må vi jo forvente, at hvis vi laver en ny undersøgelse på 1000 andre personer, så vil celletallene ændre sig en smule. Man kan altså ikke stole alt for meget på de konkrete celletal, og vores undersøgelse bør derfor, snarere end blot eksperimentets konkrete $\ln \widehat{OR}$, resultere i et helt *interval* af værdier der med rimelighed kan tænkes at ligge bag den fundne værdi. Udfordringen ligger i at bestemme intervallet sådan at det bliver forholdsvis lille, samtidig med at man holder styr på risikoen for at den sande værdi falder udenfor det.

Simulation

En undersøgelse af 1000 personer efter retninglinjerne beskrevet ovenfor tænkes udført. Ud fra undersøgelsen beregnes tallene

$$p_{\text{emp}}, \quad q_{\text{emp}}, \quad \ln OR_{\text{emp}}.$$

Vi tænker os endvidere at man ved sammenligning med andre undersøgelser kan konstatere at de fundne sandsynligheder for at have konfiguration **A** og for at ryge falder sammen med de sande værdier, altså at $p = p_{\text{emp}}$ og at $q = q_{\text{emp}}$. Tilbage er at vurdere hvilke bagvedliggende $\ln OR$-værdier der med rimelighed kan give anledning til den observerede værdi $\ln OR_{\text{emp}}$.

Vi søger en løsning på denne opgave ved hjælp af *simulation*. Vi fastholder en værdi af p_1, beregner p_2 ud fra (1), og udfører et computereksperiment hvori en tabel af samme form som Tabel 1 genereres ved lodtrækning ud fra p, q, p_1 og p_2. Der skal altså for hver af 1000 personer trækkes lod om hvorvidt vedkommende har konfiguration **A** eller ej, hvorefter der trækkes lod om hvorvidt personen er ryger eller ej. Ud fra tabellen udregnes en $\ln OR$-værdi, som vi vil kalde en *simuleret* $\ln OR$*-værdi*.

For at studere den eksperimentelt fundne værdi af $\ln OR_{\text{emp}}$ udregnes nu 500 simulerede $\ln OR$-værdier, og af disse beregnes *90%-variationsområdet* som intervallet fra den 25. største simulerede $\ln OR$-værdi til den 475. største. Vi siger at $\ln OR_{\text{emp}}$ er *kompatibel* med den fastholdte værdi af p_1 (og dermed med den tilhørende værdi af $\ln OR$) hvis $\ln OR_{\text{emp}}$ ligger i 90%-variationsområdet.

Opgavetekst

1. Bestem en formel for $\ln OR$ ud fra givne værdier af p, q og p_1.

2. I filen `L:/dat-a/projekter/projekte/edata.txt` findes resultatet af det ovenfor beskrevne (fiktive) eksperiment. Filen består af i alt 1000 linjer med to kolonner. Hver linje svarer til en person i undersøgelsen. I første kolonne står om vedkommende har konfiguration **A** eller ej, i anden kolonne står om vedkommende ryger eller ej.

   Undersøg hvordan filen er organiseret, og optæl ved hjælp af C observationerne i en tabel af samme form som Tabel 1. Beregn p_{emp}, q_{emp} og $\ln OR_{\text{emp}}$.

3. Udfør ved hjælp af C en simulation som beskrevet herover for parametrene

   (2) $$p = p_{\text{emp}}, \qquad q = q_{\text{emp}}, \qquad p_1 = p_2 = q_{\text{emp}}.$$

   Der skal altså genereres 500 $\ln OR$-værdier ved simulation af i alt 5000000 personers forhold angående **A** og rygning. Bestem det tilhørende 90%-variationsområde, og optegn ved hjælp af Maple et søjlediagram for de 500 simulerede $\ln OR$-værdier.

   Hvilken værdi af $\ln OR$ giver formlen i Spørgsmål 1 for parametrene i (2)? Er $\ln OR_{\text{emp}}$ kompatibel med denne værdi?

4. Bestem ved hjælp af C 90%-variationsområder for hver af parameterværdierne i tabellen nedenfor:

p	p_{emp}	p_{emp}	p_{emp}	p_{emp}	p_{emp}	p_{emp}	p_{emp}	p_{emp}	p_{emp}
q	q_{emp}	q_{emp}	q_{emp}	q_{emp}	q_{emp}	q_{emp}	q_{emp}	q_{emp}	q_{emp}
p_1	0.12	0.16	0.20	0.24	0.28	0.32	0.35	0.38	0.41

Præsenter ved hjælp af Maple resultatet af alle disse simulationseksperimenter på én graf: på x-aksen afsættes $\ln OR$ værdien udregnet ud fra formlen i Opgave 1, og på y-aksen afsættes det tilhørende 90%-variationsområde. Afsæt også en vandret linje gennem $\ln OR_{\text{emp}}$.

Vurder på grundlag af grafen hvilke værdier af $\ln OR$ der er kompatible med $\ln OR_{\text{emp}}$.

Rapport

Projektet besvares med en skriftlig rapport skrevet i $\LaTeX\,2_\varepsilon$. Rapporten skal ud over de to figurer beskrevet i Spørgsmål 3 og 4 indeholde en konklusion på resultatet af de udførte simulationseksperimenter.

Der må kun afleveres ét C program som bilag til rapporten. Dette program skal indeholde tre funktioner `opg2`, `opg3` og `opg4` der hver især kan benyttes til at besvare spørgsmålene 2, 3 og 4. Det skal være muligt at afprøve funktionerne separat. Der ønskes en oversigt over de C funktioner der blev defineret i programmet, der ud over at for hver funktion kort beskrive funktionens brug og effekt også opregner hvilke andre funktioner, der kalder funktionen. Diskutér med udgangspunkt i oversigten om det udformede program udnytter mulighederne for *funktionsgenbrug* optimalt.

Programmet skal udformes i overensstemmelse med Deitel & Deitels "Good programming practices"; specielt skal det kommenteres forsvarligt. En fil indeholdende programteksten skal opbevares på `M:` drevet indtil projektet er godkendt, sådan at instruktor kan rekvirere en kopi til prøvekørsel af programmet.

Tak

Tak til Ernst Hansen fra Afdeling for Teoretisk Statistik der har foreslået opgaven og ydet uvurderlig hjælp ved dens endelige udformning.