Københavns Universitet
Institut for Matematiske Fag

Eksamen i Matematisk Statistik
24. juni 2021

Eksamen er en 8 timers individuel skriftlig opgave. Alle hjælpemidler er tilladte. Besvarelsen skal udarbejdes alene og afleveres i Digital Eksamen som en samlet pdf-fil inkl. R-kode og output.

Opgavesættet består af 3 opgaver med i alt 14 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Filen `covid19vaccine.txt` *indeholder data til opgave 2, og den er tilgængelig via Digital Eksamen.*

Opgave 1

Lad (X,Y) være stokastiske variable med værdier i $\mathbb{N}_0 \times \mathbb{R}_+$ bestemt ved at $X = x$ er trukket fra en Poissonfordeling med middelværdi λ, hvorefter Y er trukket fra en gammafordeling med formparameter $x+1$ og skalaparameter β, d.v.s. den simultane fordeling af (X,Y) har tætheden

$$f_{(\lambda,\beta)}(x,y) = \frac{\lambda^x}{x!}e^{-\lambda}\,\frac{y^x}{\beta^{x+1}x!}e^{-y/\beta}, \quad y>0,\, x=0,1,\dots \tag{1}$$

med hensyn til $m \times \nu$ hvor m er tællemålet på $\mathbb{N}_0$ og ν er standard lebesguemålet på $\mathbb{R}_+$.

1. Gør rede for, at familien af ovennævnte tætheder for ukendte $(\lambda,\beta) \in \mathbb{R}_+^2$, er en regulær to-dimensional eksponentiel familie som kan repræsenteres med kanoniske parametre $\theta_1 = \log(\lambda/\beta)$ og $\theta_2 = 1/\beta$, bestem de tilhørende kanoniske stikprøvefunktioner, og vis at kumulantfunktionen er givet som

$$\psi(\theta) = \frac{e^{\theta_1}}{\theta_2} - \log\theta_2.$$

2. Vis at middelværdier, varianser, og kovarians for X og Y er

$$\mathbf{E}(X) = \lambda, \quad \mathbf{E}(Y) = \beta(\lambda+1), \quad \mathbf{V}(X) = \lambda, \quad \mathbf{V}(Y) = \beta^2(2\lambda+1), \quad \mathbf{V}(X,Y) = \beta\lambda.$$

Betragt nu n indbyrdes uafhængige observationer $(X_1,Y_1),\dots,(X_n,Y_n)$ hvor (X_i,Y_i) har tæthed af formen (1) og $(\lambda,\beta) \in \mathbb{R}_+^2$ er ukendte.

3. Bestem maximaliseringsestimatoren for (λ,β) og gør rede for, at den er veldefineret med sandsynlighed 1.
4. Bestem den asymptotiske fordeling af maksimaliseringsestimatoren $(\hat{\lambda}_n,\hat{\beta}_n)$.

Betragt nu hypotesen $H_0 : \lambda = \beta$.

5. Gør rede for, at H_0 specificerer en en-dimensional regulær eksponentiel delfamilie af den første familie.

6. Bestem maksimaliseringsestimatoren $\hat{\hat{\beta}}_n$ for β under hypotesen og vis, at denne er veldefineret med sandsynlighed 1.

7. Tyder følgende sammenhørende observationer af X og Y på at H_0 er opfyldt?

X	2	2	4	1	4	2	2	1	4	1
Y	0.63	2.04	2.79	2.27	3.42	3.15	2.48	1.10	7.14	1.51

Opgave 2

To vacciner mod COVID-19 skal sammenlignes. Den ene vaccine er fra firmaet Kenson & Kenzon (KK) og den anden fra firmaet VaccZoo (VZ). I et forsøg lader et antal frivillige forsøgspersoner sig vaccinere og derefter måles antistofniveauet med en uges mellemrum i 5 uger. Udviklingen af antistofniveauet hen over de 5 uger betegnes som forsøgspersonens *antistofrespons*.

Forsøget er et randomiseret forsøg, hvor forsøgspersonerne inddeles tilfældigt i tre grupper. En gruppe vaccineres med vaccinen fra KK, en gruppe vaccineres med vaccinen fra VZ, og den sidste gruppe vaccineres med en placebovaccine (PL). Der er to formål med forsøget. Dels skal vi undersøge, om vaccinerne har en effekt på antistofniveauet i forhold til placebovaccinen. Dels skal vi undersøge, om antistofresponset er ens for de to vacciner.

Der deltager i alt 60 personer i forsøget, og forsøgsdesignet er beskrevet ved de to faktorer

- U (uge) med værdier i $\{1,2,3,4,5\}$
- T (vaccine) med værdier i $\{\mathrm{KK}, \mathrm{VZ}, \mathrm{PL}\}$.

Bemærk at U skal opfattes som kategorisk (en faktor) i hele opgaven.

Derudover indføres faktoren S med værdier i $\{\mathrm{VA}, \mathrm{PL}\}$, med $S = \mathrm{VA}$ netop hvis $T \in \{\mathrm{KK}, \mathrm{VZ}\}$. Dvs. S betegner om forsøgspersonen er vaccineret med en vaccine eller med placebo. Responsvariablen er antistofniveauet og betegnes X. Data er i filen `covid19vaccine.txt` som en tabel med 300 rækker og fem søjler betegnet `Id`, `U`, `T`, `S` og `X`. Søjlen `Id` indeholder et identifikationsnummer fra 1 til 60, der identificerer hver af de 60 forsøgspersoner.

Opgaven består af 5 delopgaver, hvor opgaverne 1 og 2 kan besvares uafhængigt af opgaverne 3, 4, og 5. De to første opgaver benytter kun data fra uge 5, og data kan f.eks. indlæses og filtreres på følgende måde.

```
library(tidyverse)
vaccine <- read_csv("covid19vaccine.txt", col_types = cols(U = col_factor()))
vaccine_uge5 <- filter(vaccine, U == 5)
```

1. Fit en lineær model med middelværdirum L_T (etsidet variansanalyse ud fra faktoren T) for X såvel som for $\log(X)$, hvor X er antistofniveauet i uge 5. Undersøg hvilken af de to modeller, der fitter data bedst.
2. Vælg den model fra delopgave 1, som fitter data bedst. Undersøg om der i uge 5 er forskel på antistofniveauet i de tre grupper. Er der forskel på antistofniveauet for de to vacciner?
3. Betragt nu designet
$$\mathbb{G} = \{T \times U, S \times U, U, T, S, 1\}.$$
Tegn faktorstrukturdiagrammet og gør rede for at designet er $\wedge$-stabilt.
4. Afgør om designet $\mathbb{G}$ er geometrisk ortogonalt, gør rede for at
$$L_T + L_{S\times U} \subseteq L_{T\times U},$$
og find $\dim(L_{T\times U})$ og $\dim(L_T + L_{S\times U})$.
5. Undersøg hvorvidt de to vacciner giver anledning til samme antistofrespons. Dvs. om antistofniveauet udvikler sig på samme måde for de to vacciner. Vælg på baggrund af undersøgelsen en model og redegør for modellens beskrivelse af antistofresponset.

Opgave 3

For at få en bedre forståelse for kontormiljøets indflydelse på migræneanfald lavede en stor virksomhed følgende forsøg med 500 medarbejdere. Medarbejderne blev tilfældigt inddelt i to grupper med 250 i hver, og den ene gruppe arbejdede fortsat på kontoret, mens den anden arbejdede hjemme den næste måned. Medarbejderne blev bedt om at indrapportere migræneanfald i løbet af den pågældende måned. Forsøget ledte til følgende tabel med antallet af migræneanfald i de to grupper.

	Migræne		
Gruppe	Nej	Ja	Total
Kontor	223	27	250
Hjemme	234	16	250
Total	457	43	500

1. Beregn, på basis af ovenstående tabel, et estimat for odds-ratio og test hypotesen om at migræne forekommer lige hyppigt i de to grupper. Forklar hvad testet siger om odds-ratio.
2. Find et estimat for standard error for odds-ratio ved deltametoden. Beregn to 95% konfidensintervaller for odds-ratio ved at: (1) bruge den estimerede standard error; (2) transformere et 95% konfidensinterval for log-odds-ratio. Kommenter eventuelle forskelle.