Institut for Matematiske Fag
Københavns Universitet

Sandsynlighedsregning og Statistik (SS)
Ernst Hansen, Helle Sørensen
December 2015

Obligatorisk opgave i SS

Formalia og praktisk information

Projektet skal afleveres senest tirsdag den 15. december 2015 ved forelæsningerne. For sen aflevering accepteres ikke! Projektet skal laves alene eller i grupper på op til 3 studerende, hvor gruppen afleverer en fælles besvarelse. Brug den officielle forside der ligger på Absalon. I får projektet tilbage ved øvelserne tirsdag den 5. januar. Hvis projektet ikke er bestået, skal det revideres og genafleveres senest tirsdag den 12. januar ved forelæsningerne.

Opgave 1 samt spørgsmål 6–8 i opgave 2 bruger kun stof der er gennemgået i kursusuge 1–3 (når tætheden i spørgsmål 5 tages for givet). Spørgsmål 1–5 i opgave 2 kræver stof der gennemgås tirsdag den 8. december.

Der er vink til nogle af spørgsmålene, herunder hjælp til R, sidst i opgaven, men prøv først om I kan løse spørgsmålene uden hjælp.

Opgave 1

Data til denne opgave stammer fra en større undersøgelse af danskernes kostvaner fra 1986 (Haraldsdottir, J., Holm, L., Jensen, J.H. and Møller, A, 1986, *Danskernes kostvaner 1985*, Levnedsmiddelstyrelsen, publ. nr. 138).

Filen `avit.txt` på Absalon indeholder data over det daglige indtag af A-vitamin for 2224 personer. Der er tre variable: `person` der blot er en nummerering af personerne, `sex` der har værdien 1 hvis personen er en mand og 2 hvis personen er en kvinde, samt `avit` der angiver det daglige indtag af A-vitamin (målt i RE, dvs. mikrogram retinol).

1. Indlæs datasættet i R.

   Hvor mange mænd henholdsvis kvinder indgår i undersøgelsen?

   Lav en ny variabel, `avitM`, der indeholder indtaget af A-vitamin for mændene. Kontrollér at den har den rigtige længde (antallet af mænd i undersøgelsen, som I bestemte lige før). Lav også en variabel, `logavitM`, der indeholder den naturlige logaritme til værdierne i `avitM`.

Lad $x_1, \ldots, x_n$ være de observerede værdier af uafhængige identisk fordelte stokastiske variable $X_1, \ldots, X_n$. Gennemsnittet er

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

og vi kan også beregne stikprøvevariansen, s_n^2, og stikprøvespredningen, s_n:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_n = \sqrt{s_n^2}.$$

Bemærk at der divideres med $n-1$ i stedet for n.

De engelske betegnelser er „sample variance“ og „sample standard deviation“. Stikprøvevarians og stikprøvespredning kaldes også empirisk varians og empirisk spredning.

Under passende antagelser vil $\bar{x}_n$ nærme sig $E(X_i)$ og s_n^2 vil nærme sig $\mathrm{Var}(X_i)$ for $n \to \infty$. I BH afsnit 6.3 er der nogle formelle middelværdiberegninger der understøtter dette, og der bliver gjort rede for hvorfor man dividerer med $n-1$ i stedet for n.

2. Bestem median, gennemsnit, stikprøvevarians og stikprøvespredning for variablene `avitM` og `logavitM`, dvs. udfyld følgende skema:

	Median	Gennemsnit	Stikprøvevarians	Stikprøvespredning
`avitM`	*	*	*	*
`logavitM`	*	*	*	*

3. Tegn et histogram for `avitM` på „sandsynlighedsskala“, dvs. således at det samlede areal under histogrammet er 1. Tegn tætheden for normalfordelingen med middelværdi og varians lig gennemsnit og stikprøvevarians for `avitM` i samme figur.

   Lav den tilsvarende figur for `logavitM`.

   Er det mest fornuftigt at antage at A-vitaminindtaget eller logaritmen til A-vitaminindtaget er normalfordelt? I skal argumentere ud fra figurerne, men også inddrage en sammenligning mellem median og gennemsnit i jeres svar.

4. Brug en normalfordeling til at bestemme et fornuftigt estimat for sandsynligheden for at en tilfældig mand har et A-vitaminindtag på højst 2000.

5. Lad X være en stokastisk variabel med $X \sim N(\mu, \sigma^2)$, og definer $Y = e^X$. Bestem fordelingsfunktionen for Y, og vis derefter at tætheden for Y er givet ved

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right), \quad y > 0$$

   Fordelingen med denne tæthed kaldes den logaritmiske normalfordeling — fordi logaritmen til en stokastisk variabel med denne tæthed er normalfordelt.

6. Tegn histogrammet over `avitM` igen. Indtegn derefter grafen for tætheden f i samme graf i samme graf. Overvej nøje hvilke af de værdier I beregnede i spørgsmål 2 som det er fornuftigt at bruge som μ og σ^2.

Opgave 2

I denne opgave tages udgangspunkt i funktionen

$$f(x,y) = \begin{cases} \frac{4x^3}{y^3} & \text{hvis } 0 < x < 1,\ x < y \\ 0 & \text{ellers} \end{cases}$$

Lad A være det område af $\mathbb{R}^2$ hvor f er positiv.

1. Lav en skitse af området A, og vis at f er en sandsynlighedstæthed.

Lad (X, Y) være en todimensional stokastisk variabel med simultan tæthed f.

2. Lav en skitse af området $A \cap B$ hvor $B = \{(x,y) \in \mathbb{R}^2 | x + x \leq 1\}$, og bestem derefter $P(X + Y \leq 1)$.
3. Bestem den marginale tæthed for X.
4. Bestem middelværdi og varians for X.
5. Vis at den marginale tæthed for Y er givet ved

$$g(y) = \begin{cases} y & \text{hvis } 0 < y \leq 1 \\ y^{-3} & \text{hvis } y > 1 \\ 0 & \text{ellers} \end{cases}$$

6. Bestem middelværdien for Y, og vis at $\int_{-\infty}^{\infty} y^2 g(y)\, dy = \infty$. Vi skriver i så fald $E(Y^2) = \infty$ og siger at „Y ikke har varians“.
7. Bestem fordelingsfunktionen G for Y og bestem derefter den inverse funktion G^{-1} (fraktilfunktionen).
8. Lav 10000 simulerede udfald af Y. Beregn gennemsnittet af de simulerede tal, og sammenlign med middelværdien fra spørgsmål 6.

   Lav et histogram på „sandsynlighedsskala“, dvs. således at det samlede areal under histogrammet er 1. Kommenter figuren i relation til beregningen af $\mathrm{E}(Y^2)$ i spørgsmål 6.

   Lav et nyt histogram hvor I zoomer ind på intervallet $(0, 6)$. I kan bruge `xlim` i `hist`-kommandoen som beskrevet i vinket nedenfor. Tegn tætheden for g i samme figur og kommenter grafen.

R-hjælp og andre vink

1.1. Se opgave SS.1 for hjælp til indlæsning af data.

Brug fx. kommandoen `table(myData$sex)` hvis I har kaldt datasættet `myData`.

Afsnittene om „Subsets of data frames“ og „Transformation of vectors in data frames“ på side 10 i *Getting started with R and RStudio* Husk også afsnittet om „Logical expressions“ på side 5–6. Længden af en vektor `x` kan beregnes med kommandoen `length(x)`. Den naturlige logaritme hedder `log` i R.

1.2. Se afsnit 4 i *Getting started with R and RStudio*.

1.3. Husk at `hist(x, prob=TRUE)` laver et histogram på sandsynlighedsskala for vektoren `x`. Prøv også at eksperimentere med breakpoints vha. `breaks`. Tætheden for den normalfordelingen kan udregnes og indtegnes i et allerede eksisterende plot på følgende måde, hvor `**` skal erstattes af de relevante tal:

```
x <- seq(**,**,**)
f1 <- dnorm(x, mean=**, sd=**)
lines(x, f1)
```

Se også side 230–231 i BH og evt. R-programmet der blev brugt ved forelæsningerne fredag den 4. december. Det ligger på Absalon i mappen „R-programmer“.

1.4. Brug funktionen `pnorm` med passende argumenter, se evt. afsnit 6 i *Getting started with R and RStudio* eller side 229 i BH.

1.6. Hvilken variabel skal I tænke på som X, hvilken variabel skal I tænke på som Y?

Se desuden vinket til spørgsmål 1.3. Lav vektorer `y` og `f2` ved at skrive noget passende efter `**` i de følgende linier:

```
y <- ** Passende vektor, fx. defineret vha. seq
f2 <- ** Funktionsudtrykket for tætheden, evalueret i y
```

2.8. Til simulationen kan I lade jer inspirere af følgende kode:

```
U <- **
Y <- rep(NA,10000)
Y[U<0.5] <- **
Y[U>0.5] <- **
```

Til histogrammet hvor I zoomer ind kan I lade jer inspirere af følgende kode:

```
hist(Y, breaks=**, prob=TRUE, xlim=c(0,6))
y1 <- seq(0,1,0.1)
g1 <- **
lines(y1,g1)
y2 <- seq(**,**,**)
g2 <- **
lines(y2,g2)
```

Overvej også at bruge `ylim` i `hist`-kommandoen.