Institut for Matematiske Fag
Københavns Universitet

Sandsynlighedsregning og Statistik (SS)
Michael Sørensen, Helle Sørensen
December 2010

Projekt: Den logaritmiske normalfordeling

Projektet skal afleveres senest 7. januar 2011 ved forelæsningerne. Det skal laves i grupper på 3–4 studerende. Gruppen afleverer en fælles besvarelse. Husk at udfylde og vedlægge forside med navn, holdnummer og underskrift.

Vi skal undersøge den logaritmiske normalfordeling, dels sandsynlighedsteoretisk og ved simulation, og dels se hvordan man ved transformation kan analysere data der kan beskrives ved en logaritmisk normalfordeling. Der er vink til nogle af spørgsmålene, herunder hjælp til R, sidst i opgaven, men prøv først om I kan løse spørgsmålene uden hjælp.

Del 1: Sandsynlighedsregning

Lad X være normalfordelt med middelværdi μ og varians σ^2. Fordelingen af $Y = \exp(X)$ kaldes den logaritmiske normalfordeling med parametre (μ, σ^2) — naturligvis fordi $\log(Y)$ er $N(\mu, \sigma^2)$-fordelt.

1. Vis at sandsynlighedstætheden for Y er givet ved
$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right), \quad y > 0$$
2. Vis at middelværdien af Y er givet ved $\mathrm{E}(Y) = e^{\mu+\sigma^2/2}$.
3. Vis at variansen for Y er givet ved $\mathrm{Var}(Y) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$.

Medianen for en kontinuert fordeling med fordelingfunktion F er det tal z hvor $F(z) = 1/2$.

4. Bestem medianen for X's fordeling. Bestem derefter medianen for Y's fordeling.

Del 2: Simulation

5. Simulér 5000 observationer fra den logaritmiske normalfordeling med parametre $(5, 0.25)$.
6. Tegn et histogram på sandsynlighedsskala over de simulerede værdier (således at det totale areal af rektanglerne er 1). Indtegn tæthedsfunktionen for den logaritmiske normalfordeling i samme tegning. Sammenlign histogrammet og tæthedsfunktionen.

7. Beregn gennemsnit, stikprøvevarians og stikprøvespredning ($\bar{y}$, s^2 og s) samt stikprøvemedian for de simulerede værdier. Sammenlign med resultaterne fra spørgsmål 2–4.

Del 3: Analyse af datasæt

Vi skal nu analysere et datasæt over det daglige indtag af A-vitamin for 2224 personer. Datasættet er tilgængeligt i filen `avit.txt`. Der er to variable: `avit` der angiver det daglige indtag af A-vitamin (målt i RE, dvs. mikrogram retinol), samt `sex` der har værdien 1 hvis observationen er for en mand og 2 hvis observationen er for en kvinde.

Data 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.)

8. Indlæs datasættet i R.

9. Lav en ny variabel, `avitM`, der indeholder indtaget af A-vitamin for mændene. Hvor mange mænd indgår i undersøgelsen?

   Lav en ny variabel mere, `logavitM`, der indeholder den naturlige logaritme til værdierne i `avitM`.

10. Tegn histogrammer og QQ-plots for variablene `avitM` og `logavitM`. Diskuter figurerne.

11. Beregn gennemsnit, stikprøvevarians og stikprøvespredning for `logavitM`.

12. Tegn histogrammet for `logavitM` igen, denne gang sammen med tætheden for normalfordelingen med middelværdi og varians lig med de værdier I beregnede i spørgsmål 11.

    Tegn også histogrammet for `avitM` igen, denne gang sammen med tætheden for den tilhørende logaritmiske normalfordeling. Diskuter figurerne.

13. Opstil en statistisk model for `logavitM`. Angiv estimaterne for parametrene i modellen. Angiv også den teoretiske samt den estimerede fordeling for estimatorerne.

14. Beregn et 95% konfidensinterval for det gennemsnitlige logaritmiske A-vitaminindtag for mænd.

    I skal både beregne konfidensintervallet 'i hånden' (dvs. sætte tal ind i de rette formler) og bruge `t.test` funktionen. Kontrollér at I får det samme.

15. Foreslå et estimat for medianen i fordelingen af A-vitaminindtaget for mænd. Bestem også et 95% konfidensinterval for denne median.

R-hjælp og andre hints

2. Kig evt. først på specialtilfældet hvor $\mu = 0$. Brug for eksempel MS, sætning 4.2.3 med $t(x) = e^x$ og $X \sim N(0, \sigma^2)$, og omskriv integranden til noget der har med tætheden for $N(\sigma^2, \sigma^2)$ at gøre.

4. Hvad sker der med medianen ved transformation af en stokastisk variabel med en voksende funktion?

5. Kommandoen `sim <- rnorm(5000,mean=5,sd=0.5)` laver en vektor med 5000 simulerede udfald fra normalfordelingen med middelværdi 5 og varians 0.25. Hvordan kan I bruge disse værdier til at simulere udfald fra den logaritmiske normalfordeling?

6. Husk at `hist(x, prob=T)` laver et histogram på sandsynlighedsskala for vektoren `x`. Prøv evt. at eksperimentere med `nclass`, for eksempel `hist(x, nclass=25)`.

   Tætheden for den logaritmiske normalfordeling med parametre (μ, σ^2) kan defineres og indtegnes i et allerede eksisterende plot på følgende måde:

   ```
   f <- function(y,mu,sigma) her skrives funktionsudtrykket

   yval <- seq(0,1000,1)          ## definerer y-værdier
   fval <- f(yval, 5, 0.5)        ## tætheden i y-værdierne
   lines(yval, fval)              ## tegn oveni plot
   ```

7. Se afsnit 4 i *Introduction to R*.

8. Brug `read.table` og `attach` som beskrevet i afsnit 3 i *Introduction to R*.

   Husk at angive den fulde sti til filen, eller at skifte 'arbejdskatalog' (working directory) til det katalog hvor datafilen ligger. `attach`-kommandoen gør at variablene i datasættet kan bruges direkte.

9. Se afsnit 2 i *Introduction to R*. Kommandoen `avit[sex==1]` kan for eksempel være nyttig.

   Længden af en vektor `x` kan beregnes med kommandoen `length(x)`. Den naturlige logaritme hedder `log` i R.

10. Et QQ-plot der sammenligner variablen `x` med en normalfordeling kan laves med kommandoen `qqnorm(x)`. Virker det rimeligt at antage at `avitM` og/eller `logavitM` er normalfordelt?

12. Se vink til spørgsmål 6.

15. Se vink til spørgsmål 4.