Projektopgave 2
“Sandsynlighedsregning og statistik”

Michael Andersen – michael@diku.dk
Henrik Jensen – henrikgjensen@gmail.com
Ulrik Bonde – bonde@diku.dk
Julie Nielsen – julie@diku.dk

8. januar 2010

Del 1

Opgave 1

For at vise at Y er givet ved

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp - \frac{(\log y - \mu)^2}{2\sigma^2}$$

Benytter vi først formlen for normalfordeling $N \sim (\mu, \sigma^2)$ givet ved M.S. 5.3.5

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y - \mu)^2}{2\sigma^2}$$

For at vise at ovenstående er lige med den logaritmiske normalfordeling benytter vi sætning 5.4.1 omkring transformation af kontinuerte funktioner på $\mathbb{R}$.

Som siger følgende

$$q(y) = \begin{cases} p(t^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} t^{-1}(y)) \right| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases}$$

Ovenstående er produktet af to led, derfor kan vi betragte de to led individuelt.

$$p(t^{-1}(y)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(\log y - \mu)^2}{2\sigma^2}$$

$$\frac{\mathrm{d}}{\mathrm{d}y} \left| t^{-1}(y) \right| = y^{-1} = \frac{1}{y}$$

Vi ganger nu disse to led sammen og får derfor

$$q(y) = \begin{cases} \frac{1}{y\sqrt{2\pi\sigma^2}} \exp -\frac{(\log y - \mu)^2}{2\sigma^2} & \text{hvis } y \in (0, \infty) \\ 0 & \text{ellers} \end{cases}$$

Det er hermed vist at den logaritmiske normalfordeling er givet ved ovenstående formel.

Opgave 2

Vi skal her vise at middelværdien af Y er givet ved $E(Y) = e^{\mu + \sigma 2/2}$. Vi benytter her sætning 5.2.3, der siger at hvis der gælder at

$$\int_I |t(x)| \, p(x) \, \mathrm{d}x < \infty$$

Da er den oprindelige stokastiske variable X er baseret på normalfordelingen og det vides at alle momenter for normalfordelingen eksistere, derfor vides det også at både middelværdi og varians må eksistere. Og derfor er middelværdien givet ved

$$E(t(X)) = \int_I t(x)p(x) \, \mathrm{d}x$$

Vores transformerede stokastiske variable er e^X, hvor $X \sim N(\mu, \sigma^2)$

$$E(e^X) = \int_{-\infty}^{\infty} e^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2} \, \mathrm{d}x$$

Vi foretager her en substitution $t = x - \mu$

$$= \int_{-\infty}^{\infty} e^{t+\mu} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-t^2}{2\sigma^2} \, \mathrm{d}t$$

$$= e^{\mu} \int_{-\infty}^{\infty} e^t \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-t^2}{2\sigma^2} \, \mathrm{d}t$$

Integralet i ovenstående identificeres som værende e^T, hvor $T \sim N(0, \sigma^2)$

$$E(e^(T)) = \int_{-\infty}^{\infty} e^t \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-t^2}{2\sigma^2} \, \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-t^2}{2\sigma^2} + t \, \mathrm{d}t$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-t^2}{2\sigma^2} + t \, \mathrm{d}t$$

Her er vi interesseret i at omskrive eksponenten til noget vi kan genkende.

$$-\frac{t^2}{2\sigma^2} + t = -\frac{1}{2\sigma^2} \left(t^2 - 2\sigma^2 t\right)$$

$$= -\frac{1}{2\sigma^2}\left(\left(t-\sigma^2\right)^2 - \left(\sigma^2\right)^2\right)$$

$$= -\frac{\left(t-\sigma^2\right)^2}{2\sigma^2} + \frac{\sigma^2}{2}$$

Det sætter vi nu ind igen, og får følgenede

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(t-\sigma^2)}{2\sigma^2} + \frac{\sigma^2}{2}$$

Dette kan vi nu omskrive til følgende

$$= \exp \frac{\sigma^2}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(t-\sigma^2)}{2\sigma^2}$$

Integralet genkender man som værende tæthedsfunktionen for $T \sim N(\sigma^2, \sigma^2)$ som i henhold til 5.1.3 giver 1.

Sammen med resultatet fra $X \sim N(\mu, \sigma^2)$, giver dette.

$$\exp \mu \exp \frac{\sigma^2}{2} = \exp \mu + \frac{\sigma^2}{2}$$

Hermed er det vist at middelværdien for Y er givet ved ovenstående.

Opgave 3

Der skal bestemmes medianen af fordelingen for Y. Hvis x er givet ved fordelingen F_X, da gælder at

$$q_x = F_X^{-1}$$

Da $Y = exp(X)$, gælder der

$$F_Y(y) = P(Y \leq y) = P(\exp X \leq y) = P(X \leq \log Y) = F_X(\log Y)$$

Og derfor gælder følgende

$$q_Y(z) = F_Y^{-1}(z) = (F_X \circ \log)^{-1}(z) = \exp q_X(z)$$

Derfor kan der nu indsættes $\frac{1}{2}$ som udgør medianen i fordeling.

$$q_Y(\frac{1}{2}) = \exp q_X(\frac{1}{2})$$

Der findes ikke noget eksplicit udtryk for værdien af fordelingsfunktion for normalfordelingen. Men det der skal løses er

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y$$

$$\frac{1}{2} = \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y$$

Da vi nu betragter $X \sim N(\mu, \sigma^2)$, og det vides at μ er middelværdien og findes der hvor $x = \mu$, må medianen være lige med

$$\exp \mu$$

Opgave 4

Vi benytter igen samme fremgangs måde som i opgave 2, og betragter derfor

$$(E(e^T))^2 = \int_{-\infty}^{\infty} (\exp t)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(t-\mu)^2}{2\sigma^2}\,\mathrm{d}t$$

Vi laver igen substitutionen $x = t - \mu$

$$\exp 2\mu \int_{-\infty}^{\infty} = \exp 2x \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-x^2}{2\sigma^2}\,\mathrm{d}x$$

$$\exp 2\mu \int_{-\infty}^{\infty} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-x^2}{2\sigma^2} + 2x\,\mathrm{d}x$$

Vi tager nu eksponenten ud igen og behandler den

$$\frac{-x^2}{2\sigma^2} + 2x = -\frac{1}{2\sigma^2}\left(x^2 - 2x2\sigma^2\right)$$

$$= \frac{1}{2\sigma^2}\left(-\left(x - 2\sigma^2\right)^2 - \left(2\sigma^2\right)^2\right)$$

$$= -\frac{(x-2\sigma^2)^2}{2\sigma^2} + \frac{(2\sigma^2)^2}{2\sigma^2}$$

$$= -\frac{(x-2\sigma^2)^2}{2\sigma^2} + \frac{4\sigma^4}{2\sigma^2} = -\frac{(x-2\sigma^2)^2}{2\sigma^2} + 2\sigma^2$$

Dette kan nu indsættes og der fås derved

$$\exp 2\mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\sigma^2)^2}{2\sigma^2} + 2\sigma^2\,\mathrm{d}x$$

$$\exp 2\mu + 2\sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\sigma^2)^2}{2\sigma^2}\,\mathrm{d}x$$

Det tilbageværende integrale genkendes som værende $X \sim N(\sigma^2, \sigma^2)$.

$$Var(Y) = E(Y^2) - (E(Y))^2$$

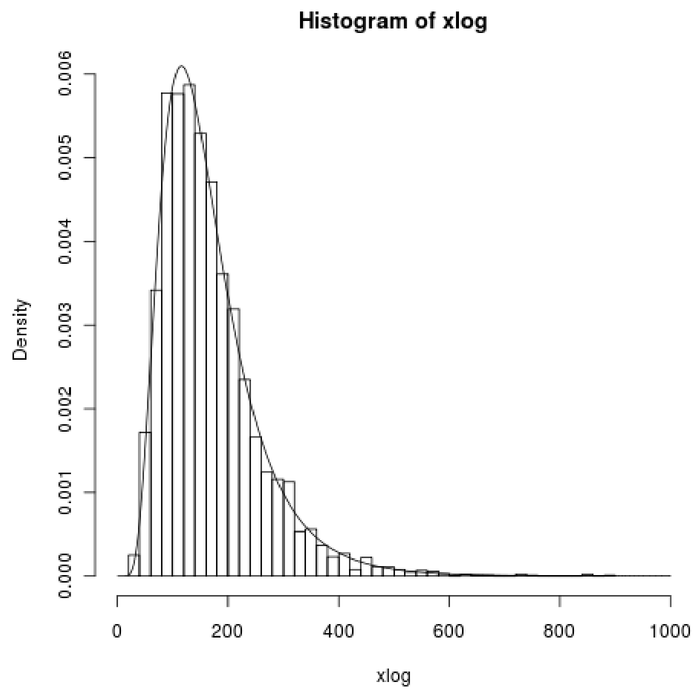
Med resultatet fra fra opgave 1 udregnes $E(Y^2)$ til

$$\exp \sigma^2 + 2\mu$$

Og vi har lige udregnet $(E(Y))^2$

$$(E(Y))^2 = \exp 2\mu + 2\sigma^2$$

$$Var(Y) = \exp \sigma^2 + 2\mu - \exp 2\mu + 2\sigma^2 = (\exp \sigma^2 - 1)\exp 2\mu + \sigma^2$$



Figur 1: Simulation af 5000 observationer fra den logaritmiske normalfordeling med parametre $(5, 0.25)$.

Del 2

Vi har simuleret 5000 observationer fra den logaritmiske normalfordeling med parametre $(5, 0.25)$. Plottet kan ses i figur 1.

Ved brug af **R** har vi fundet stikprøvens gennemsnit, varians, spredning og median. Disse ses i udskriften fra programmet i kodeboks 1.

Ved at bruge formlerne fra Del 1, har vi regnet de teoretiske værdier for gennemsnit, varians, spredning og median. De ses herunder.

$$\begin{aligned} E(S) &= e^{\mu+\sigma^2/2} \\ &= e^{5+0.125} \\ &= 168.1741 \end{aligned} \tag{1}$$

$$\begin{aligned} Var(S) &= (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} \\ &= (e^{0.25} - 1)e^{10+0.25} \\ &= 8032.96 \end{aligned} \tag{2}$$

```
[1] "Assignment 7"
[1] "Middelvaerdi"
[1] 167.1192
[1] "Varians"
[1] 7976.314
[1] "Standardafvigelse"
[1] 89.31021
[1] "Median"
[1] 147.5676
```

Kodeboks 1: Udskrift fra R-program

$$\begin{aligned}\sqrt{Var(S)} &= \sqrt{8032.96} \\ &= 89.62678 \end{aligned} \tag{3}$$

$$\begin{aligned}F^{-1}(0.5) &= e^{\mu} \\ &= e^{5} \\ &= 148.4132 \end{aligned} \tag{4}$$

Det ses, at de teoretiske værdier ligger meget tæt på dem som er blevet udregnet fra stikprøven. Dette kan kan også ses ved at kigge på plottet i figur 1, hvor tæthedsfunktionen følger de observerede værdier. Det var også at forvente, når man har taget så stor en stikprøve.

Del 3

I undersøgelsen indgår der 1079 mænd og 1145 kvinder. Vi ser i figur 2 og 3, at den originale data ikke er normalfordelt, men når vi transformerer denne logaritmisk, har vi en normalfordeling.

Vi siger derfor, at det logaritmisk transformerede indtag af A-vitamin for kvinder og mænd, er normalfordelt.

Vi estimerer $\mu = \bar{x} = 7.361356$ og $\sigma = s = 0.4486289$. Vi har den teoretiske fordeling

$$X \sim N(\mu, \sigma^2) \tag{5}$$

og får den estimerede fordeling, ved at indsætte vores estimater til μ og σ i (5).

$$X \sim N(7.361356, 0.2012679) \tag{6}$$

I figur 4 ses hvordan de observerede værdier forholder sig til den estimerede fordeling.

Vi undersøger nu, hvorvidt der er forskel på mænd og kvinders indtag af A-vitamin. Vi bruger `R` til at lave en t-test på vores data. Vi får outputet vist i kodeboks 2. Det 95 procents konfidensinterval aflæses direkte fra koden, mens estimatet for differensen Fra t-testen fås ved at trække de beregnede middelværdier fra hinanden.

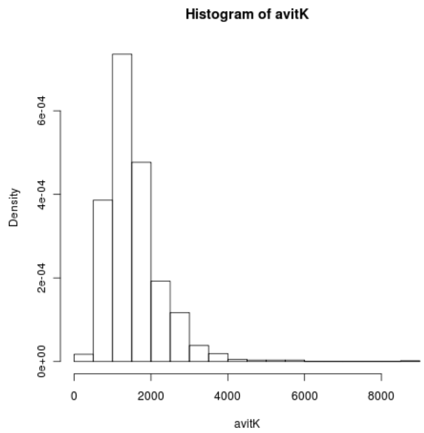
```
Welch Two Sample t-test

data:  logavitM and logavitK
t = 13.0815, df = 2206.488, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2041469 0.2761477
 sample estimates:
 mean of x mean of y
  7.484993  7.244845
```

Kodeboks 2: Udskrift fra T-test

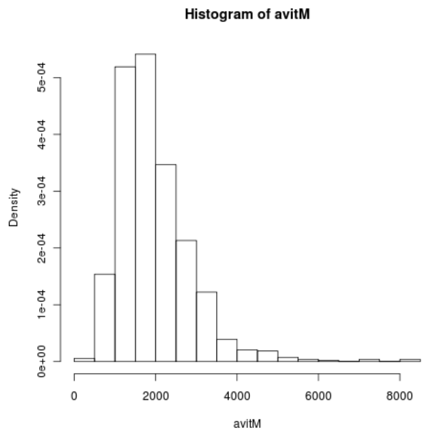
Vores 95-procents konfidensinterval for differensen mellem middelværdierne er $[0.2041469, 0.2761477]$. Dvs. at der er 95% sandsynlighed for at differensen mellem to middelværdier ligger i dette interval. Vores estimat af differens af middelværdierne er $7.484993 - 7.244845 = 0.240148$, hvor vi har at $0.240148 \in [0.2041469, 0.2761477]$. Mænd og kvinders logaritmiske indtag af A-vitamin kan derfor godt være det samme.

Ligesom vi kan finde medianen i den logaritmiske normalfordeling ved e^{μ}, kan vi bruge samme transformation for grænserne og differensen. På grund af regneregler for eksponentialfunktionen vil vi automatisk få et forhold når vi transformerer en differens, da $e^{a-b} = \frac{e^a}{e^b}$. Hvis vi transformerer estimatet og grænserne for konfidensintervalet for differensen, får vi et estimat og grænser for et konfidensinterval for forholdet, mellem medianerne i fordelingen af indtaget af A-vitamin, for mænd og kvinder. Konfidensintervallet for forholdet bliver $[1.226478, 1.318043]$ og differensen transformerer til 1.271437. Vi har at $1.271437 \in [1.226478, 1.318043]$.

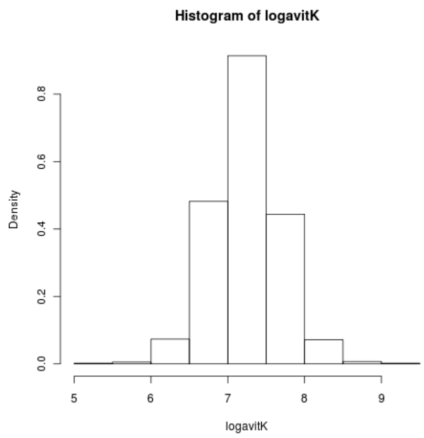
Da $1 \notin [1.226478, 1.318043]$ kan tæller og nævner i forholdet ikke være identiske, og der er derfor mindst 95% sandsynlighed for at kvinder og mænds indtag af A-vitamin er forskellige.



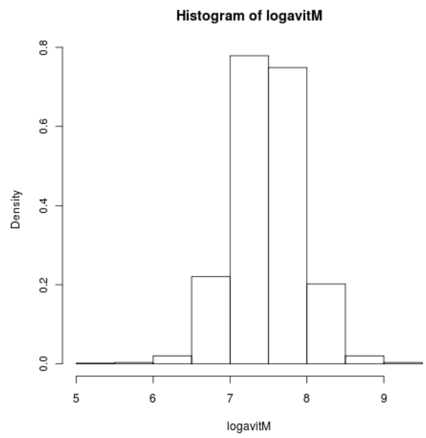
(a) A-vitaminindtag for kvinder



(b) A-vitaminindtag for mænd

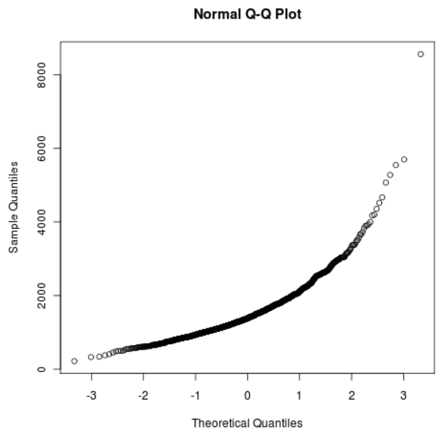


(c) Logaritmisk tranformeret A-vitaminindtag for kvinder

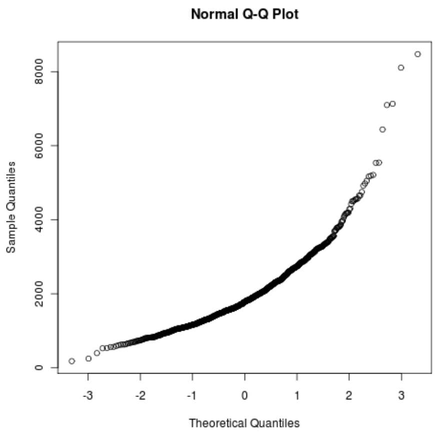


(d) Logaritmisk tranformeret A-vitaminindtag for mænd

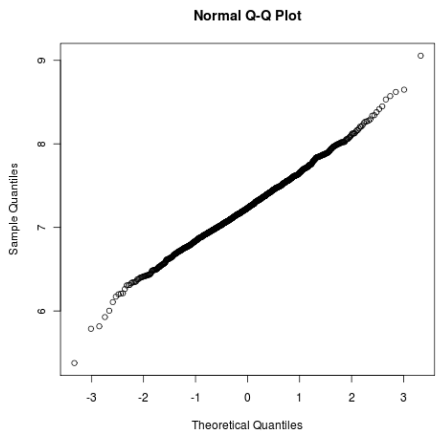
Figur 2: Kvinder og mænds indtag af A-vitamin. Den originale data ligner ikke en normalfordeling, men det gør den transformerede.



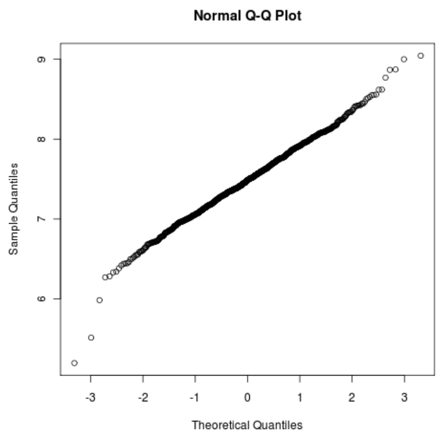
(a) QQ-plot for kvinders indtag af A-vitaminindtag



(b) QQ-plot for mænds A-vitaminindtag

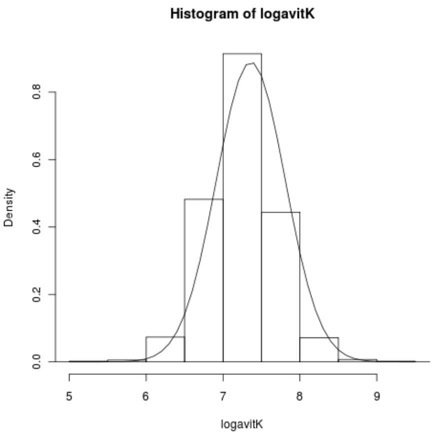


(c) QQ-plot af det logaritmisk tranformerede A-vitaminindtag for kvinder

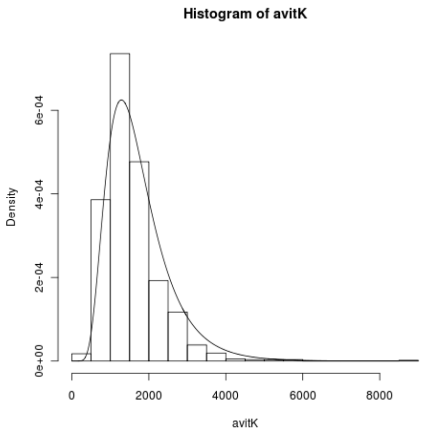


(d) QQ-plot af det logaritmisk tranformerede A-vitaminindtag for mænd

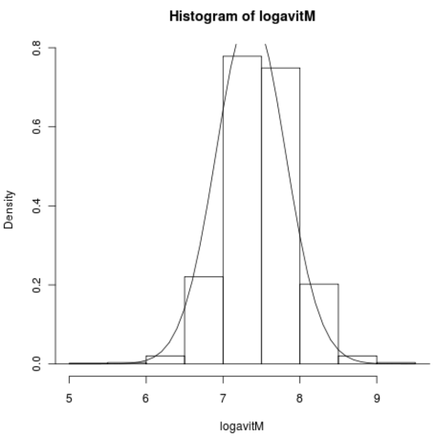
Figur 3: QQ-plots for indtag af A-vitamin. Det ses, at når data bliver logaritmisk transformeret, får vi en normalfordeling, da punktene, i disse qq-plots, følger en diagonal.



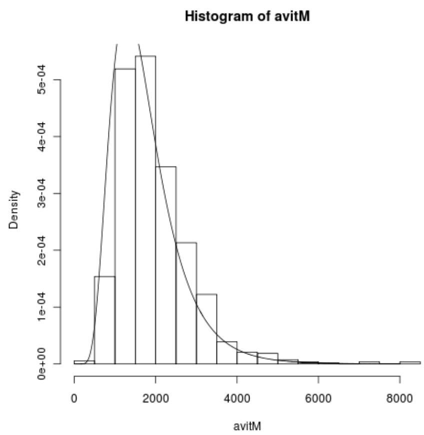
(a) Logaritmisk A-vitaminindtag for kvinder med estimeret tæthed



(b) A-vitaminindtag for kvinder med estimeret tæthed



(c) Logaritmisk A-vitaminindtag for mænd med estimeret tæthed



(d) A-vitaminindtag for mænd med estimeret tæthed

Figur 4: Det ses, at de observerede værdier nogenlunde følger den estimerede tæthed. Man kan også se, at der er flere kvinder end mænd i undersøgelsen, da værdierne for mænd ligger lidt under de estimerede værdier og vice versa.

A Source code

A.1 main.r

```r
#!/usr/bin/env Rscript

# R statistics for the course "Sandsynlighedsregning og statistik"
# at the Department of Mathematical Sciences, University of Copenhagen

#########################################
#
# Project 2
# Part 2
#
#########################################

f <- function(y, mu, sigma) {

  return (1/(sqrt(2 * pi * sigma^2) * y) * exp(-((log( y ) - mu)^2 / (2 * sigma^2))))

}

# Assignment 1
assignment1 <- function() {

    set.seed(42);

    mymu <- 5;
    mysigma <- sqrt(0.25);

    xlog <- rlnorm(5000, meanlog=mymu, sdlog=mysigma);

    # Set the filename for the plot and plot the graph
    name <- "plot_1.png";
    png(name);
    hist(xlog, density=NULL, prob=TRUE, nclass=50, xlim=range(0,1000,100));

    yval <- seq(1, 1000, 1);
    fval <- f(yval, mu=mymu, sigma=mysigma);

    points(yval, fval, type="l");

    dev.off();

    print("Assignment 7");
    print("Median");
    print(median(xlog));
    print("Middelvaerdi");
    print(mean(xlog));
    print("Standardafvigelse");
    print(sd(xlog));
    print("Varians");
    print(var(xlog));
    print("");

}

# Assignment 3
assignment2 <- function() {

   avitdata <- read.table("avit.txt", header=TRUE);
   attach(avitdata);

   avitM <- avit[sex==1];
   avitK <- avit[sex==2];

   logavitM <- log(avitM);
   logavitK <- log(avitK);

   print("Number of men in the experiment:");
   print(length(avitM));
   print("Number of females in the experiment:");
```

```
    print(length(avitK));

#    print(length(avitK)+length(avitM))

     name <- "plot_3_avitM.png";
     png(name);
     hist(avitM, prob=TRUE);

     name <- "plot_3_avitK.png";
     png(name);
     hist(avitK, prob=TRUE);

     name <- "plot_3_logavitK.png";
     png(name);
     hist(logavitK, prob=TRUE);

     name <- "plot_3_logavitM.png";
     png(name);
     hist(logavitM, prob=TRUE);

     name <- "plot_3_avitM_qq.png";
     png(name);
     qqnorm(avitM);

     name <- "plot_3_avitK_qq.png";
     png(name);
     qqnorm(avitK);

     name <- "plot_3_logavitM_qq.png";
     png(name);
     qqnorm(logavitM);

     name <- "plot_3_logavitK_qq.png";
     png(name);
     qqnorm(logavitK);

     logavit <- log(avit);
     mloga <- mean(logavit);
     print("Mean of the log");
     print(mloga);

     sdloga <- sd(logavit);
     print("log standard deviation:");
     print(sdloga);

     dev.off();

     # Exercise 11
     # logavitM og logavitK ~ N(mean(log(avit)), sd(log(avit)))

     # Exercise 12

     name <- "plot_3_logavitM_with_normal.png";
     png(name);
     hist(logavitM, prob=TRUE);

     yval <- seq(1,12,0.1);
     fval <- dnorm(yval, mean=mloga, sd=sdloga);

     points(yval, fval, type="l");

     dev.off();

     name <- "plot_3_avitM_with_lognormal.png";
     png(name);
     hist(avitM, prob=TRUE);

     yval <- seq(0,8000,1);
     fval <- f(yval, mloga, sdloga);

     points(yval, fval, type="l");

     dev.off();
```

```
    name <- "plot_3_logavitK_with_normal.png";
    png(name);
    hist(logavitK, prob=TRUE);

    yval <- seq(1,26,0.1);
    fval <- dnorm(yval, mean=mloga, sd=sdloga);

    points(yval, fval, type="l");

    dev.off();

    name <- "plot_3_avitK_with_lognormal.png";
    png(name);
    hist(avitK, prob=TRUE);

    yval <- seq(0,8000,1);
    fval <- f(yval, mloga, sdloga);

    points(yval, fval, type="l");

    dev.off();

    t.test(logavitM, logavitK, var.equal=TRUE);

}


########################################


########################################
# Main

assignment1();
assignment2();
```