Sandsynlighedsregning og Statistik (SS)

Institut for Matematiske Fag, Københavns Universitet

Januar 2011

Projekt

	Navn og hold	**Fødselsdag**
1		
2		
3		
4		

Denne projektopgave er en obligatorisk del af kurset *Sandsynlighedsregning og Statistik* (SS). Gruppens deltagere erklærer ved deres underskrift, at de alle har bidraget på lige fod ved udarbejdelsen af projektet.

Underskrifter:

SS Projekt

Den logaritmiske normalfordeling

Mathias Svensson, Ronni Elken Lindsgaard, Jacob Kirstejn & Philip Munksgaard

7 Januar 2011

Indhold

1 Sandsynlighedsregning

1. Lad X være normalfordelt med middelværdi μ og varians σ^2, og lad $Y = \exp(X)$. Lad f_X, F_X, f_Y og F_X være sandsynlighedstætheden og fordelingsfunktionerne for henholdsvis X og Y. Der gælder at $F_X(x) = F_Y(\exp(x))$. Substitueres x med $\ln y, y > 0$ fås $F_X(\ln y) = F_Y(y)$, hvorved $f_Y(y), y > 0$ kan udregnes:

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} F_X(\ln y) \\ &= \frac{1}{y} f_X(\ln y) \\ &= \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

Det ses også ud fra denne udregning at $f_Y(y) \cdot y = f_X(\ln y)$, hvilket benyttes i de følgende opgaver.

2. Middelværdien for Y findes, ved at substitueres med $x + \mu + \frac{\sigma^2}{2} = \ln y$, hvor det bruges at $\frac{\mathrm{d}y}{\mathrm{d}x} = \exp(x + \mu + \frac{\sigma^2}{2})$:

$$\begin{aligned} \mathrm{E}(Y) &= \int_0^\infty y \cdot f_Y(y)\ \mathrm{d}y \\ &= \int_0^\infty f_X(\ln y)\ \mathrm{d}y \\ &= \int_{-\infty}^\infty f_X\left(x + \mu + \frac{\sigma^2}{2}\right) \cdot \exp\left(x + \mu + \frac{\sigma^2}{2}\right)\ \mathrm{d}x \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \left(\int_{-\infty}^\infty f_X\left(x + \mu + \frac{\sigma^2}{2}\right) \cdot \exp(x)\ \mathrm{d}x\right) \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \left(\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x + \mu + \frac{\sigma^2}{2} - \mu)^2}{2\sigma^2}\right) \cdot \exp(x)\ \mathrm{d}x\right) \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \left(\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x + \frac{\sigma^2}{2})^2}{2\sigma^2} + x\right)\ \mathrm{d}x\right) \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \left(\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x - \frac{\sigma^2}{2})^2}{2\sigma^2}\right)\ \mathrm{d}x\right) \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \end{aligned}$$

Her udnyttes i det næstsidste trin, at $-\frac{(x+\frac{\sigma^2}{2})^2}{2\sigma^2} + x$ har dobbeltroden $\frac{\sigma^2}{2}$, og dermed kan omskrives. I det sidste trin bruges, at indholdet af integral blot er tæthedsfunktionen til normalfordelingen med middelværdi $\frac{\sigma^2}{2}$ og varians σ^2. Dette integral har netop værdien 1 og forsvinder derfor.

3. Vi skal finde variansen $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$. Vi ved at $(\mathrm{E}(Y))^2 = \exp(\mu + \sigma^2/2)^2 = \exp(2\mu + \sigma^2)$. Nu mangler altså blot at vise, at $\mathrm{E}(Y^2) = \exp(\sigma^2)\exp(2\mu+\sigma^2) = \exp(2\mu+2\sigma^2)$ for at $\mathrm{Var}(Y) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)$ er opfyldt.

Der laves substitution med $x + \sigma^2 + \mu = \ln y$, hvor $\frac{\mathrm{d}y}{\mathrm{d}x} = \exp\left(x + \sigma^2 + \mu\right)$:

$$
\begin{aligned}
\mathrm{E}(Y^2) &= \int_0^\infty y^2 \cdot f_Y(y)\ \mathrm{d}y \\
&= \int_0^\infty y \cdot f_X(\ln y)\ \mathrm{d}y \\
&= \int_{-\infty}^\infty \exp(x + \sigma^2 + \mu) \cdot f_X(x + \sigma^2 + \mu) \cdot \exp(x + \sigma^2 + \mu)\ \mathrm{d}x \\
&= \int_{-\infty}^\infty \exp(2x + 2\sigma^2 + 2\mu) \cdot f_X(x + \sigma^2 + \mu)\ \mathrm{d}x \\
&= \exp(2\sigma^2 + 2\mu) \cdot \left(\int_{-\infty}^\infty \exp(2x) \cdot f_X(x + \sigma^2 + \mu)\ \mathrm{d}x \right) \\
&= \exp(2\sigma^2 + 2\mu) \cdot \left(\int_{-\infty}^\infty \exp(2x) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x + \sigma^2)^2}{2\sigma^2} \right)\ \mathrm{d}x \right) \\
&= \exp(2\sigma^2 + 2\mu) \cdot \left(\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x + \sigma^2)^2 - 4\sigma^2 x}{2\sigma^2} \right)\ \mathrm{d}x \right) \\
&= \exp(2\sigma^2 + 2\mu) \cdot \left(\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x - \sigma^2)^2}{2\sigma^2} \right)\ \mathrm{d}x \right) \\
&= \exp(2\sigma^2 + 2\mu)
\end{aligned}
$$

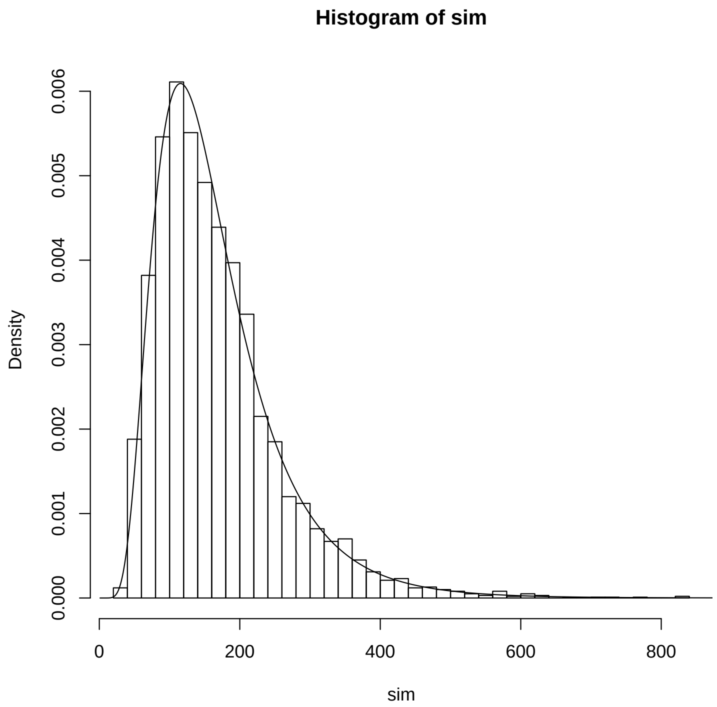
Igen bruges der i det sidste trin, at indholdet af integralet er tæthedsfunktionen til normalfordelingen - denne gang med med middelværdi σ^2 og varians σ^2. Da integralet har værdien 1 forkortes dette led.

4. Det ses direkte ud af formlen for normalfordelingen at den er symmetrisk omkring μ, altså med andre ord at $f_X(\mu - x) = f_X(\mu + x)$. Dette er i sig selv nok til at vise, at medianen for X er μ:

$$
\begin{aligned}
F_X(\mu) &= \int_{-\infty}^{\mu} f_X(x)\ \mathrm{d}x \\
&= \int_{-\infty}^{0} f_X(\mu + x)\ \mathrm{d}x \\
&= \frac{1}{2}\left(2\int_{-\infty}^{0} f_X(\mu + x)\ \mathrm{d}x\right) \\
&= \frac{1}{2}\left(\int_{-\infty}^{0} f_X(\mu + x)\ \mathrm{d}x + \int_{-\infty}^{0} f_X(\mu - x)\ \mathrm{d}x\right) \\
&= \frac{1}{2}\left(\int_{-\infty}^{0} f_X(\mu + x)\ \mathrm{d}x + \int_{0}^{\infty} f_X(\mu + x)\ \mathrm{d}x\right) \\
&= \frac{1}{2}\int_{-\infty}^{\infty} f_X(\mu + x)\ \mathrm{d}x \\
&= \frac{1}{2}\int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x \\
&= \frac{1}{2}
\end{aligned}
$$

Ved at bruge at $F_X(x) = F_Y(\exp(x))$ fås at $F_Y(\exp(\mu)) = F_X(\mu) = \frac{1}{2}$, hvorved medianen for Y er $\exp(\mu)$.

2 Simulation



Figur 1: Histogram for de observerede data samt tæthedsfunktionen for den logaritmiske normalfordeling

6. Histogrammet og tæthedsfunktionen ligner hinanden. Det tyder på at de observerede data er logaritmisk normalfordelte.

7. Stikprøverne stemmer nogenlunde overens med de beregnede resultater.

```
> c(mean(sim),var(sim),sd(sim),median(sim))
[1]  168.25094 7657.62393   87.50785  149.82703
```

Figur 2:

	Stikprøve	Udregning
Gennemsnit	$168,251$	$168,174$
Varians	$7657,624$	$8032,96$
Spredning	$87,51$	$89,63$
Median	$149,827$	$148,413$

3 Analyse af datasæt

8. Vi indlæser vores datasæt i R ved følgende kommando:

```
avit <- read.table("avit.txt", header=TRUE)
```

Figur 3:

9. Der laves først en variabel, avitM, hvori vi ligger udtaget af mændenes A-vitaminindtag på:

```
> avitM <- subset(avit, sex==1)$avit
```

Figur 4:

Herefter bestemmes antallet:

```
> length(avitM)
[]  1079
```

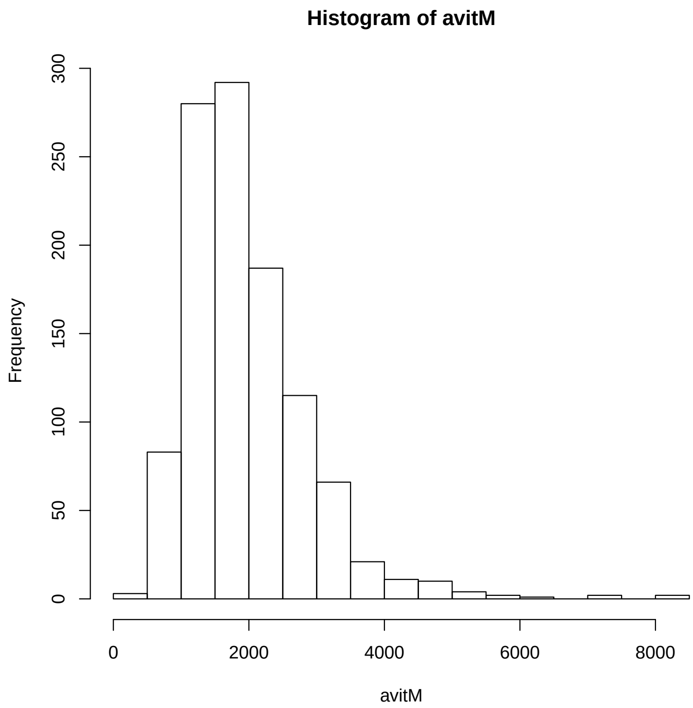
Figur 5:

Der laves nu endnu en variabel, der indeholder logaritmen til værdierne i avitM:

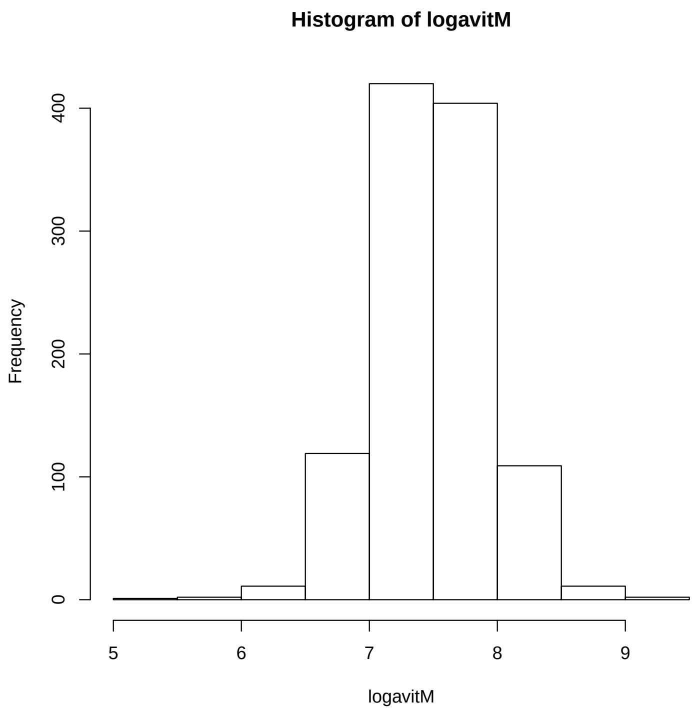
```
> logavitM <- log(avitM)
```

Figur 6:

10. Vi indtegner nu histogrammer henholdsvis for avitM og logavitM:



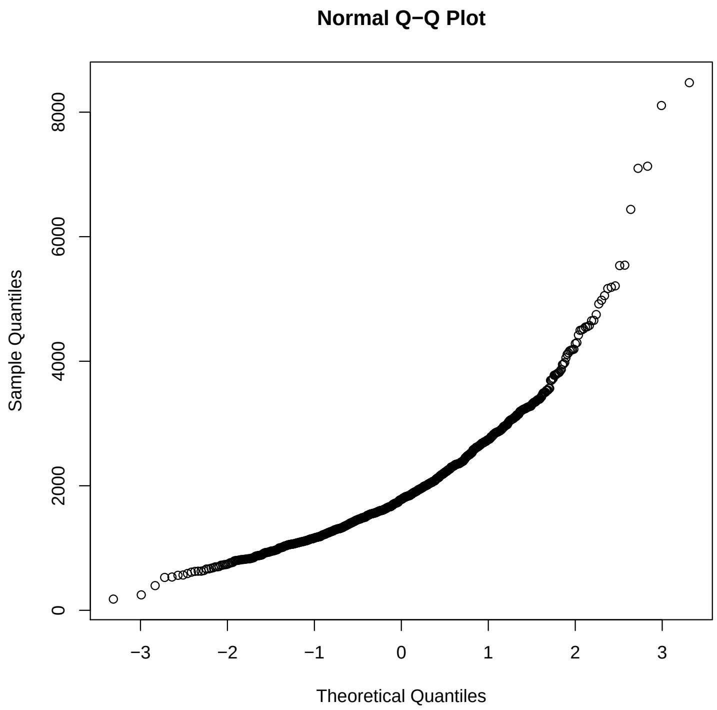
Figur 7: Histogram for avitM



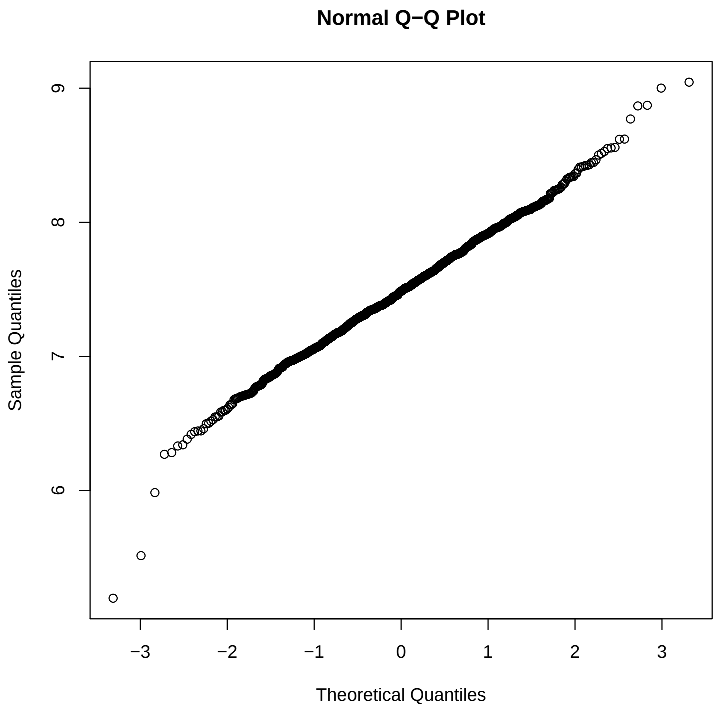
Figur 8: Histogram for logavitM

Som vi kan se ser grafen for logaritmen til indtaget af A-vitamin mere normalfordelt ud end grafen for de rene data.

Nedenfor følger QQ-plots for avitM og logavitM:



Figur 9: QQ-plot for avitM



Figur 10: Histogram for avitM

Vi ser her tydeligt, at grafen for de logaritmiske data er en meget bedre approksimation til en normalfordeling.

11. Gennemsnit for logavitM beregnes:

```
> mean(logavitM)
[]  7.484993
```

Figur 11:

Stikprøvevarians for logavitM beregnes:

```
> var(logavitM)
[]  0.1916541
```

Figur 12:

Stikprøvespredning for logavitM beregnes:

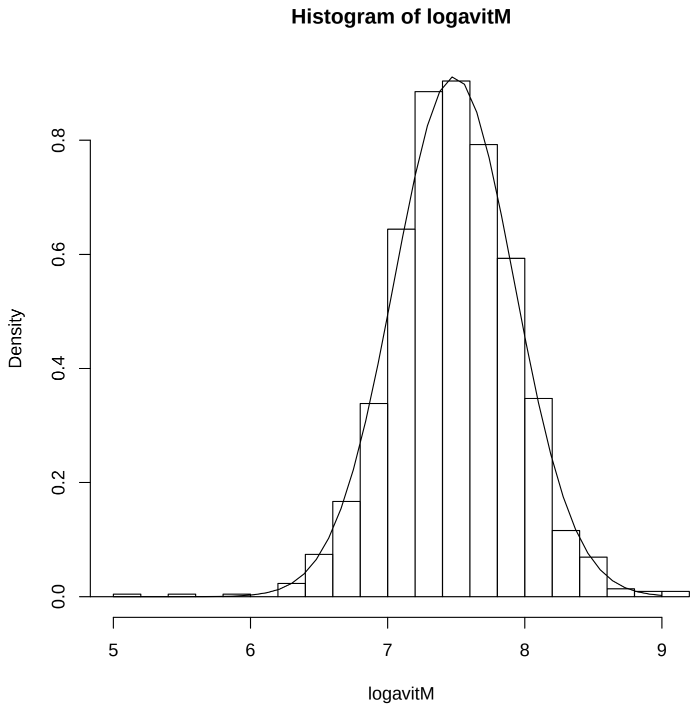
```
> sd(logavitM)
[]  0.4377832
```

Figur 13:

12. Histogrammet for logavitM indtegnes denne gang med tætheden for normalfordelingen med middelværdi og varians som fundet tidligere:

```
> hist(logavitM, prop=T, nclass=25)
> f = function(x) dnorm(x, mean(logavitM), sd(logavitM))
> plot(f, 0, 9, add=T)
```

Figur 14:

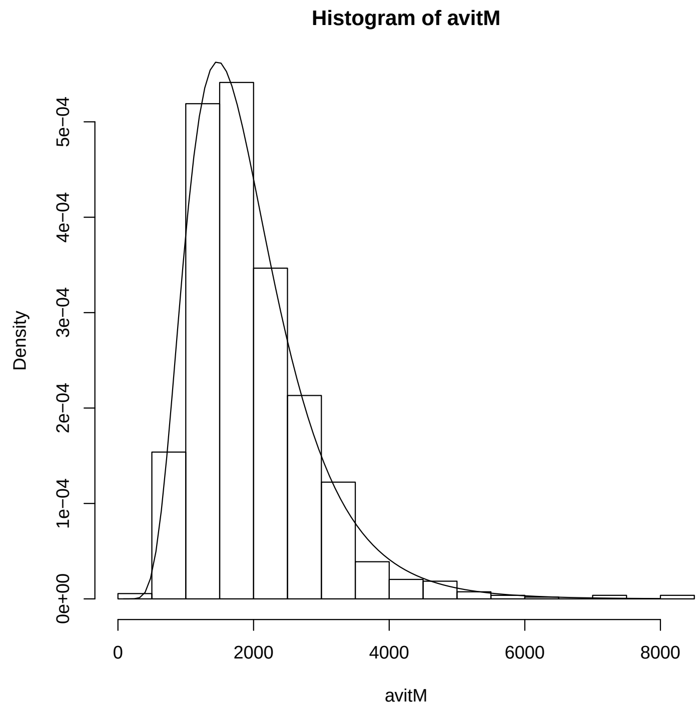


Figur 15: Histogram for logavitM med normalfordelingen indtegnet

Histogrammet for avitM indtegnes denne gang sammen med tætheden for den tilhørende logaritmiske normalfordeling:

```
> lnorm = function(y, mu, sigma) 1/(y*sqrt(2*pi*sigma^2)) *
> exp(-((log(y) - mu)^2) / (2*sigma^2))
> hist(avitM, prop=T, nclass=25)
> g = function(x) lnorm(x, mean(logavitM), sd(logavitM))
> plot(g, 0, 8000, add=T)
```

Figur 16:



Figur 17: Histogrammet for avitM med den tilhørende logaritmiske normalfordeling

Som vi kan se passer den tætheden for den logaritmiske normalfordeling godt på histogrammet over indtaget af A-vitaminer. Ligeledes passer passer tætheden af normalfordelingen med den fundne middelværdi og varians også godt med histogrammet for logavitM.

13. Den statistike model er givet ved udfaldsrummet $E = [0; \infty)$ samt familien

$$\mathcal{P} = \{N^n_{\mu,\sigma^2} : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

af fordelinger på $\mathbb{R}^2$ hvor N^n_{μ,σ^2} har tæthed

$$f_{\mu,\sigma^2}(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp(-\frac{(logy - \mu)^2}{2\sigma^2}), y > 0$$

For de 1079 $y_1, \dots, y_{1079}$ af logaritmen af nogle mænds indtag af a-vitaminer har vi fået at

$$\bar{y} = 7,485, \quad ssd = \Sigma_{i=1}^{1079}(y_i - \bar{y})^2 = 206,6031$$

Estimaterne er altså

$$\hat{\mu} = 7,485, \quad s^2 = \frac{206,6031}{1079 - 1} = 0,1917, \quad s = 0,4378$$

Den estimerede fordeling for $\hat{\mu}$ er $N(7,485; \frac{0,1917}{1079} = 1,7762 \cdot 10^{-4})$ og den estimerede fordeling for $\tilde{\sigma}^2$ er 0,1914χ^2_{1078}.

14. $n = 1079 \qquad \bar{y} = 7,485 \qquad \sigma^2 = 0,1917$

$$7,485 \pm 1,96 \cdot \frac{0,1917}{\sqrt{1079}} = (7,436; 7,496)$$

```
> t.test(logavitM)

        One Sample t-test

data:  logavitM
t = 561.6206, df = 1078, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.458842 7.511143
sample estimates:
mean of x
 7.484993
```

15. Eftersom logaritmefunktionen er en strengt voksende funktion (for $x > 0$) rykker medianen så at sige ikke plads i fordelingen selvom man tager logaritmen. Eftersom logaritmefordelingen er mere koncentreret ville et godt bud på medianen i fordelingen af A-vitamindtaget for mænd være e opløftet i middelværdien for logaritmen af fordelingen, altså $e^(7,485) = 1781.11$.