Math 152 Assignment 3 Solutions

9.15

With $\sigma = 1$

The p-value for the test of H_0 is, for $0 \le x \le 3$,

$$P(X \ge x \mid H_0) = 1 - \Phi(x)$$

where Φ is the c.d.f of $N(0,1)$.

On the other hand

$$P(H_0 \mid x) = \frac{P(x \mid H_0)\, P(H_0)}{P(x \mid H_0) P(H_0) + P(x \mid H_1) P(H_1)}$$

$$= \frac{\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{2}}{\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \cdot \frac{1}{2}}$$

$$= \frac{e^{-x^2/2}}{e^{-x^2/2} + e^{-(x-1)^2/2}}$$

With $\sigma = 2$

$$P(X \ge x \mid H_0) = P\left(\frac{X}{\sigma} \ge \frac{x}{\sigma} \mid H_0\right)$$

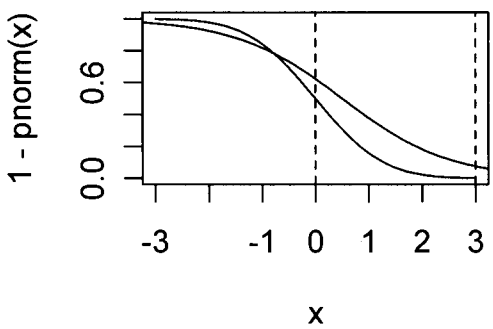
$$= 1 - \Phi\left(\frac{x}{\sigma}\right) = 1 - \Phi\left(\frac{x}{2}\right)$$

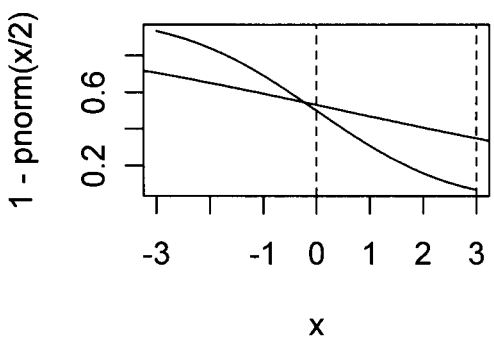
while $P(H_0 \mid x) = \dfrac{\frac{1}{2} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \cdot \frac{1}{2} + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-1)^2}{2\sigma^2}}}$

$$= \frac{e^{-x^2/8}}{e^{-x^2/8} + e^{-\frac{(x-1)^2}{8}}}$$

In both cases ($\sigma=1$, $\sigma=2$) it is clear that $P(X \geq x \mid H_0) \neq P(H_0 \mid x)$, and this is illustrated in the attached plots.

i.e. The p-value is not the posterior probability that the Null hypothesis is true!


1 - pnorm(x)
0.0
0.6
-3
-1
0
1
2
3
x


1 - pnorm(x/2)
0.2
0.6
-3
-1
0
1
2
3
x

9.21. X uniform on $[0, \theta]$

$H_0: \theta = 1$

$H_1: \theta = 2$

a) reject H_0 if $X > 1$

Then $\text{Prob}(X > 1 \mid H_0) = 0$ is the significant level,

and $\text{Prob}(X > 1 \mid H_1) = \frac{1}{2}$ is the power.

b) reject if $X \in [0, \alpha]$

The significance level is α.

$\text{Prob}(X \in [0, \alpha] \mid H_1) = \frac{\alpha}{2}$ is the power.

c) reject if $X \in [1-\alpha, 1]$

$\text{Prob}(X \in [1-\alpha, 1] \mid H_0) = \alpha$

$\text{Prob}(X \in [1-\alpha, 1] \mid H_1) = \alpha/2$

d) reject if (e.g.) $X \in \left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2}\right)$

significance level $= \alpha$

power $= \alpha/2$

in the same way.

9.21 cont.

e) The likelihood ratio test rejects for small values of.

$$\frac{P(x/H_0)}{P(x/H_1)} = \begin{cases} 2 & 0 \le x \le 1 \\ 0 & 1 < x \le 2 \end{cases}$$

if $0 < \alpha < 1$ is chosen as a significance level, we can't find $c > 0$ s.t.

$$Prob\left(\frac{P(x/H_0)}{P(x/H_1)} < c \,/\, H_0\right) = \alpha$$

since $Prob\left(\frac{P(x/H_0)}{P(x/H_1)} = 2 \,/\, H_0\right) = 1.$

For all $2 > c > 0$, we have

$$Prob\left(\frac{P(x/H_0)}{P(x/H_1)} < c \,/\, H_0\right) = 0 \le \alpha.$$

So there is no unique rejection region determined by the likelihood ratio.

9.21 cont.

f) if the null and alternative are interchanged

$H_0: \theta = 2$

$H_1: \theta = 1.$

$$\frac{P(x|H_0)}{P(x|H_1)} = \frac{1}{2} \qquad 0 \le x \le 1$$

$$= \infty \qquad 1 < x < 2$$

Again, no rejection region can be uniquely determined at level $0 < \alpha < 1$.

9.42

a) $$\hat{p} = \frac{157.0 + 69.1 + 35.2 + 3.17 + 4.1 + 5.1}{5(280)}$$

$$\approx .142$$

(Put $T = 280$)

b)

Breaks/Bar	Observed	Expected ($\hat{p}$)
0	157	$(1-\hat{p})^5 T \simeq 130.20$
1	69	$5\hat{p}(1-\hat{p})^4 T \simeq 107.74$
2	35	$10\hat{p}^2(1-\hat{p})^3 T \simeq 35.66$
3, 4, 5	19	$(1 - (\text{previous } 3))T \simeq 6.4$

$$\chi^2 = \frac{(157-130.2)^2}{130.2} + \frac{(69-107.74)^2}{107.74} + \frac{(35.66-35)^2}{35.66} + \frac{(19-6.4)^2}{6.4}$$

$$\simeq 44.26 \quad \text{and} \quad \chi^2 \sim \chi^2_2 \text{ under } H_0 \qquad P(\chi^2 \ge 44.26 | H_0) \cong 2.45 \times 10^{-10}$$

b) cont. The p-value for the Pearson chi-square test is negligible and we may conclude that the value of p varies from bar to bar.

c). Using the statistic from problem 41, as discussed in class,

$$T = \frac{157(0-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{69(1-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{35(2-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{17(3-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{(4-5\hat{p})^2}{5\hat{p}(1-\hat{p})} + \frac{(5-5\hat{p})^2}{5\hat{p}(1-\hat{p})}$$

$$\approx 429.$$

and T is $\approx \chi^2_{279}$ which, in turn,

is $\approx N(279, 558)$.

$$\text{Prob}\{T \geq 429\} = \text{Prob}\left\{\frac{T-279}{\sqrt{558}} \geq \frac{429-279}{\sqrt{558}}\right\}$$

$$\approx 1 - \Phi\left(\frac{429-279}{\sqrt{558}}\right)$$

$$\approx 1.077 \times 10^{-10}$$

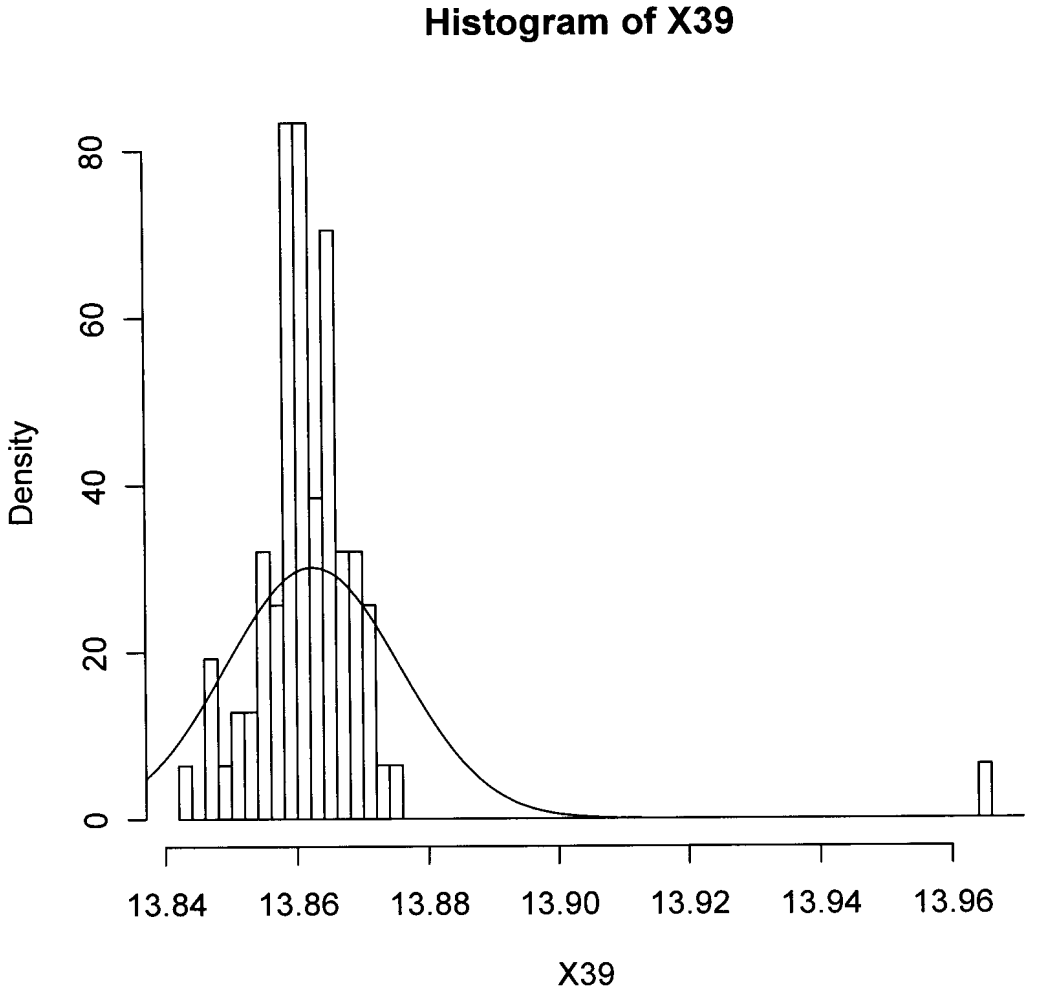
and we reach the same conclusion.

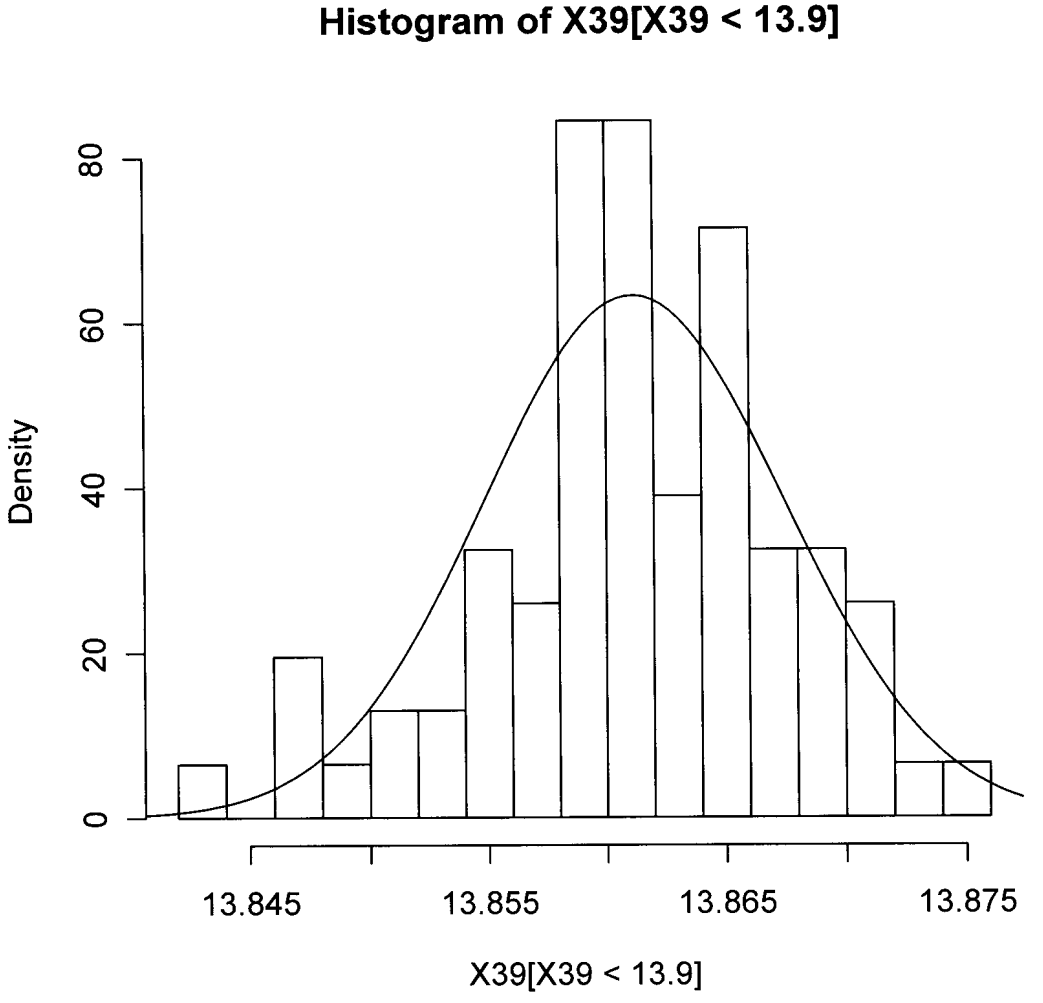
```
> #Problem 45
>
>
> #using the statistic developed in problem 41 (as demonstrated in class).
>
>
> fr<-c(7,45,181,478,829,1112,1343,1033,670,286,104,24,3)
>
> phat<- sum(fr*(0:12))/sum(12*fr)
>
> phat
[1] 0.480785
>
> sum(fr*((0:12)-12*phat)^2/(12*phat*(1-phat)))
[1] 7122.813
>
> 1-pchisq(7122.813,df=6114)
[1] 0
>
> (7122.813-6114)/sqrt(2*6114)
[1] 9.1229
> #[1] 9.1229
>
> 1-pnorm((7122.813-6114)/sqrt(2*6114))
[1] 0
> #[1] 0
>
>
> #The p-value is negligible.
>
> #Or using the Pearson Chi squared statistic.
>
> dbinom(0:12,12,phat)
 [1] 0.0003838550 0.0042653244 0.0217229277 0.0670502964 0.1396969168
 [6] 0.2069714320 0.2235943345 0.1774669993 0.1027072791 0.0422690325
[11] 0.0117421375 0.0019769154 0.0001525494
>
> #the expected values under the null hypothesis are:
>
> e<-6115*dbinom(0:12,12,phat)
>  e
 [1]    2.3472734   26.0824586  132.8357027  410.0125627  854.2466464
 [6] 1265.6303069 1367.2793552 1085.2107008  628.0550119  258.4751335
[11]   71.8031709   12.0888374    0.9328394
> #Pooling the appropriate cells to get at least 5
> #expected observations per cell:
>
> e<-c(e[1] + e[2],e[3:11],e[12] + e[13])
>
> #and pooling the observed frequencies ion the same way:
>
> fr<-c(fr[1] + fr[2],fr[3:11],fr[12] + fr[13])
>
>  fr
 [1]   52  181  478  829 1112 1343 1033  670  286  104   27
>  #[1]   52  181  478  829 1112 1343 1033  670  286  104   27
>
>
> #The chi square statistic is:
>
>  sum((fr-e)^2/e)
[1] 105.7913
> #[1] 105.7913
>
>  1-pchisq(105.79,df=9)
[1] 0
> #[1] 0
```

```
>
> #and the p-value is, again, negligible.
>
>
> #To understand why the model fails, it may be useful to see where the large
> #contributions to the chi square value are coming from:
>
> ((fr-e)^2/e)
 [1] 19.5414269 17.4636749 11.2735366  0.7461465 18.6486299  0.4311387
 [7]  2.5119152  2.8013183  2.9311070 14.4371869 15.0052503
>  #[1] 19.5414269 17.4636749 11.2735366  0.7461465 18.6486299  0.4311387
>  #[7]  2.5119152  2.8013183  2.9311070 14.4371869 15.0052503
>
> #In particular, the contributions from the "tails" (0,1,2) or (10,11,12)
> #boys are too large. The independence model in the hypothesis may be incorrect.
> #Perhaps these families kept growing until they reached a desired number of
> #either girls or boys.
>
```

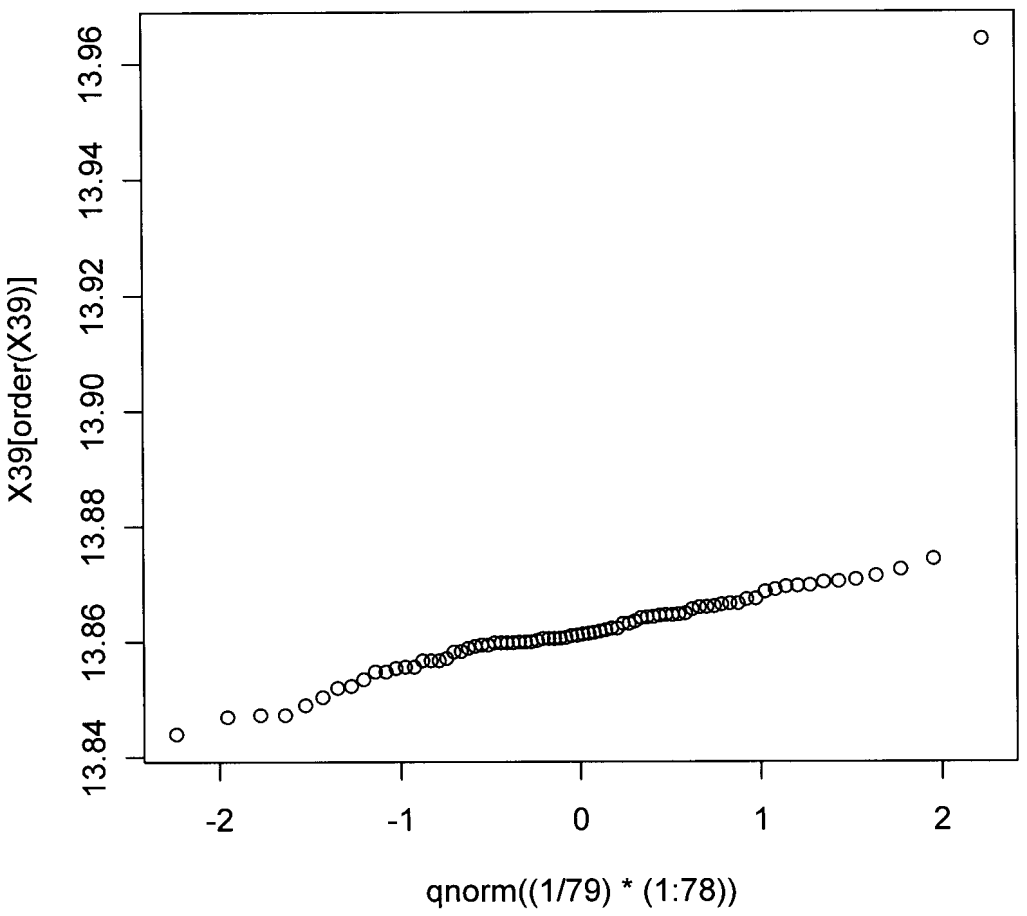
```
 #problem 9.52
> potassium<-read.table("http://math.cmc.edu/moneill/Math152/Handouts/POTASS.txt",header=T)
>
> attach(potassium)
> names(potassium)
[1] "X39" "X41"
>
> hist(X39,breaks=50,prob=T)
> curve(dnorm(x,mean=mean(X39),sd=sd(X39)),add=T)
>
> #Notice the presence of the outlying data point in the attached histogram.
>
> #With the outlier deleted, we have (also attached):
>
> hist(X39[X39<13.9],breaks=20,prob=T)
> curve(dnorm(x,mean=mean(X39[X39<13.9]),sd=sd(X39[X39<13.9])),add=T)
>
>
> #The outlier is evident in the probability plot as well (attached):
>
> length(X39)
[1] 78
> plot(qnorm((1/79)*(1:78)),X39[order(X39)])
>
> #Here is the probability plot without the outlier:
>
> plot(qnorm((1/78)*(1:77)),X39[X39<13.9][order(X39[X39<13.9])])
>
> #Now we apply our correlation test:
>
> cor(qnorm((1/78)*(1:77)),X39[X39<13.9][order(X39[X39<13.9])])
[1] 0.9876971
>
>
> f<-function(x){l<-rnorm(78)
+ cor(qnorm((1/79)*(1:78)),l[order(l)])
+ }
>
>
> d<-sapply(1:10000,f)
> d[order(d)[50]]
[1] 0.971669
> d[order(d)[100]]
[1] 0.974718
> d[order(d)[250]]
[1] 0.9793975
> d[order(d)[500]]
[1] 0.9826589
> d[order(d)[1000]]
[1] 0.9860163
> d[order(d)[2000]]
[1] 0.9891315
> d[order(d)[2500]]
[1] 0.9901766
> d[order(d)[5000]]
[1] 0.9931908
> d[order(d)[7500]]
[1] 0.995208
> d[order(d)[9000]]
[1] 0.9964919
> d[order(d)[9500]]
[1] 0.9970316
> d[order(d)[9750]]
[1] 0.9974406
> d[order(d)[9900]]
[1] 0.9978867
> d[order(d)[9950]]
```

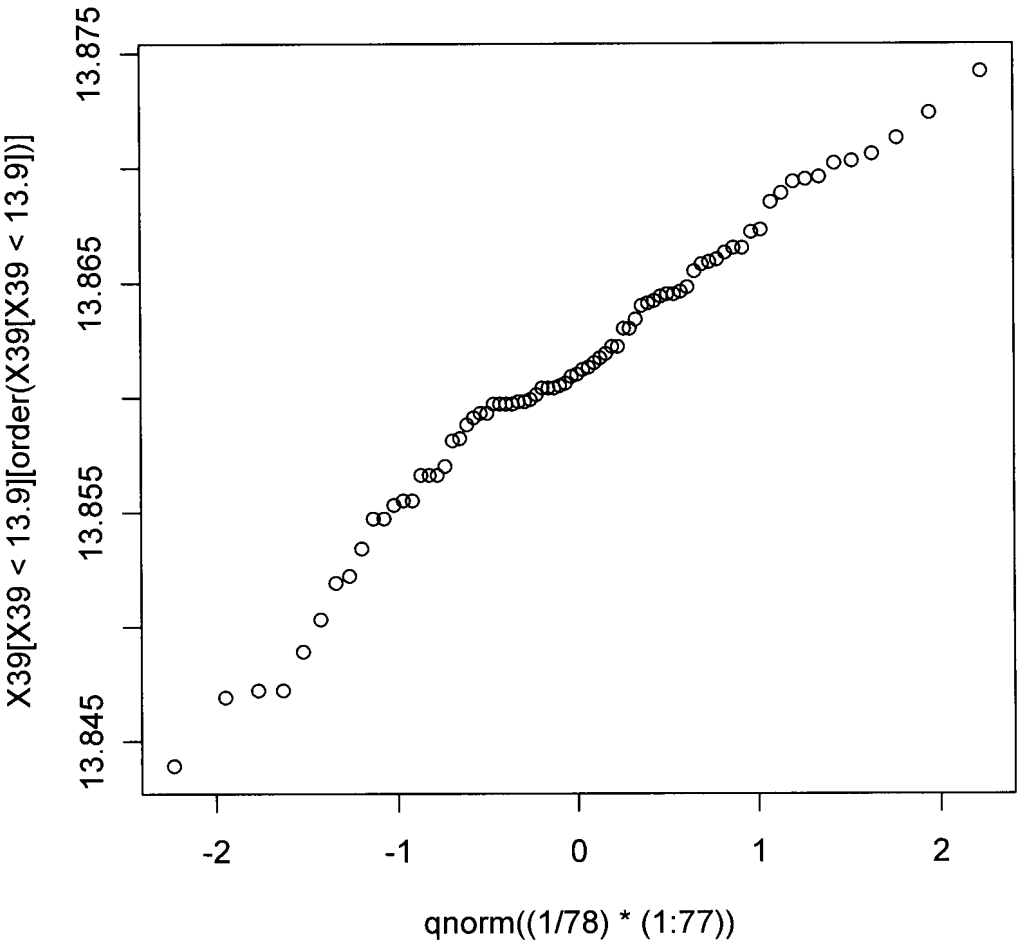
```
[1] 0.9981187
>
> #With the outlier deleted, our test has a p-value of about .09.
>
>
>  #With the outlier included:
> cor(qnorm((1/79)*(1:78)),X39[order(X39)])
[1] 0.686803
>
> #We would clearly reject the null hypothesis on the basis of
> #this numerical value alone.
>
>
> #We now do the same analysis or the other isotope:
>
>
> hist(X41,prob=T)
> curve(dnorm(x,mean=mean(X41),sd=sd(X41)),add=T)
>
> #The right hand tail may contain an outlier:
>
> hist(X41[X41<579],prob=T)
>  curve(dnorm(x,mean=mean(X41[X41<579]),sd=sd(X41[X41<579])),add=T)
>
>
> #Histograms and fitted densities with and without the possible
> # outlier are attached.
>
> #The probability plot is more revealing:
>
> plot(qnorm((1/79)*(1:78),mean=mean(X41),sd=sd(X41)),X41[order(X41)])
>
>  cor(qnorm((1/79)*(1:78),mean=mean(X41),sd=sd(X41)),X41[order(X41)])
[1] 0.9835837
> #[1] 0.9835837
> #5% level
>  cor(qnorm((1/78)*(1:77)),X41[X41<579][order(X41[X41<579])])
[1] 0.9961584
> #[1] 0.9961584
> #75% level
>
>
> #If we delete the single data point at 580,
> #the p-value  (for normality) changes from .05 to .75.
>
```


Histogram of X39
Density
0
20
40
60
80
13.84
13.86
13.88
13.90
13.92
13.94
13.96
X39

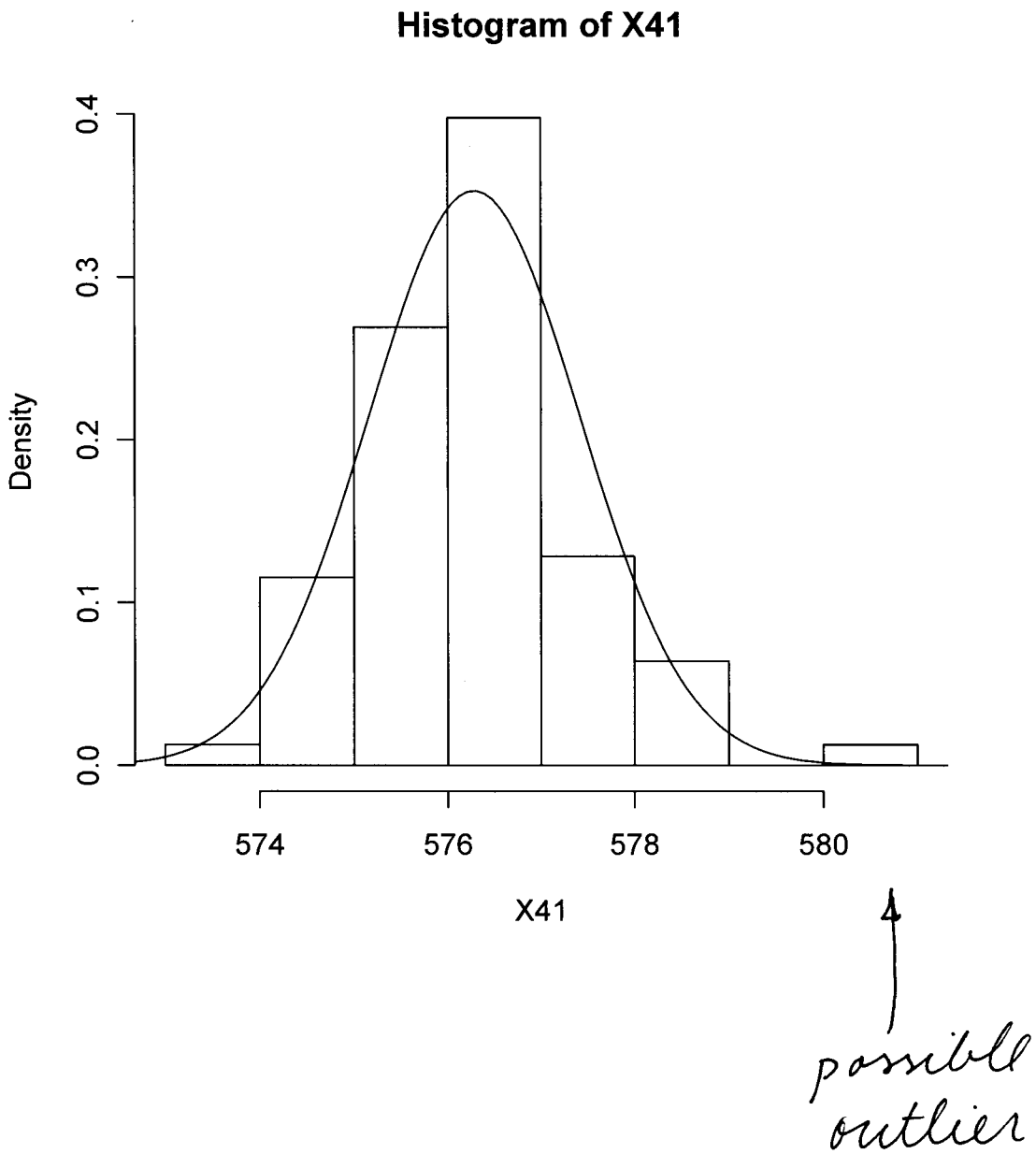


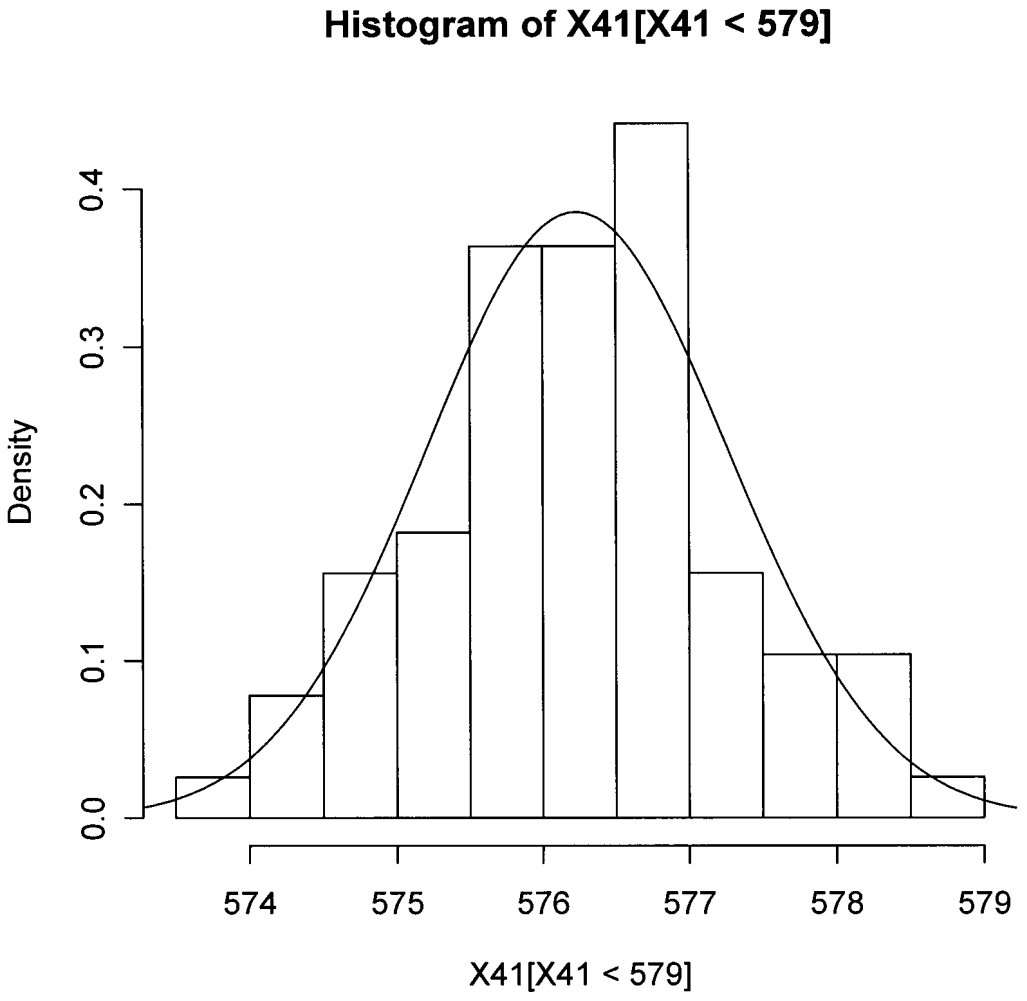
Outlier deleted.


X39[order(X39)]
13.84
13.86
13.88
13.90
13.92
13.94
13.96
-2
-1
0
1
2
qnorm((1/79) * (1:78))

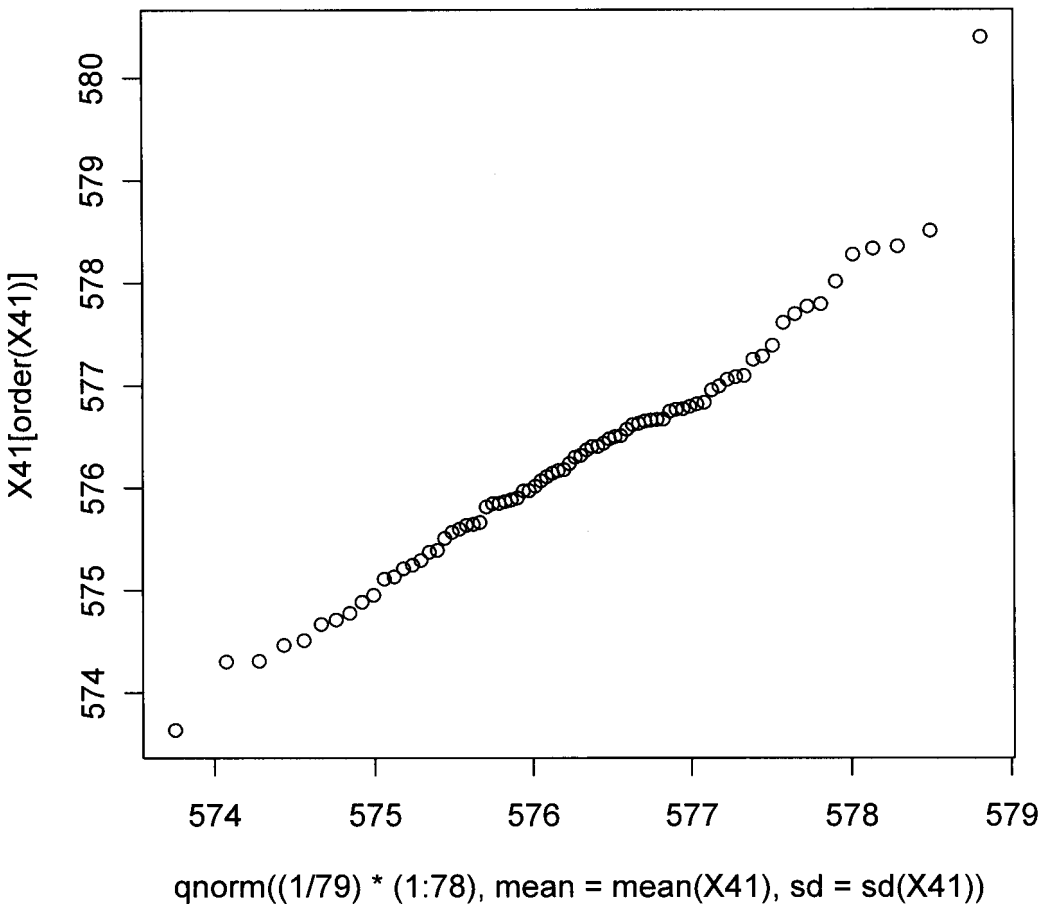


Outlier deleted.


Histogram of X41
Density
0.0
0.1
0.2
0.3
0.4
574
576
578
580
X41
possible
outlier



possible outlier deleted



The outlier is more evident in the probability plot.

9.54

a). If $Y = \log(X)$ is normally distributed with mean μ and variance σ^2 then

$X = e^Y$ has.

$$\text{Prob}\{X \le x\} = \begin{cases} 0 & x \le 0. \\ \text{Prob}\{Y \le \log x\} & x > 0. \end{cases}$$

and

$$\text{Prob}\{Y \le \log x\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\log x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

so that

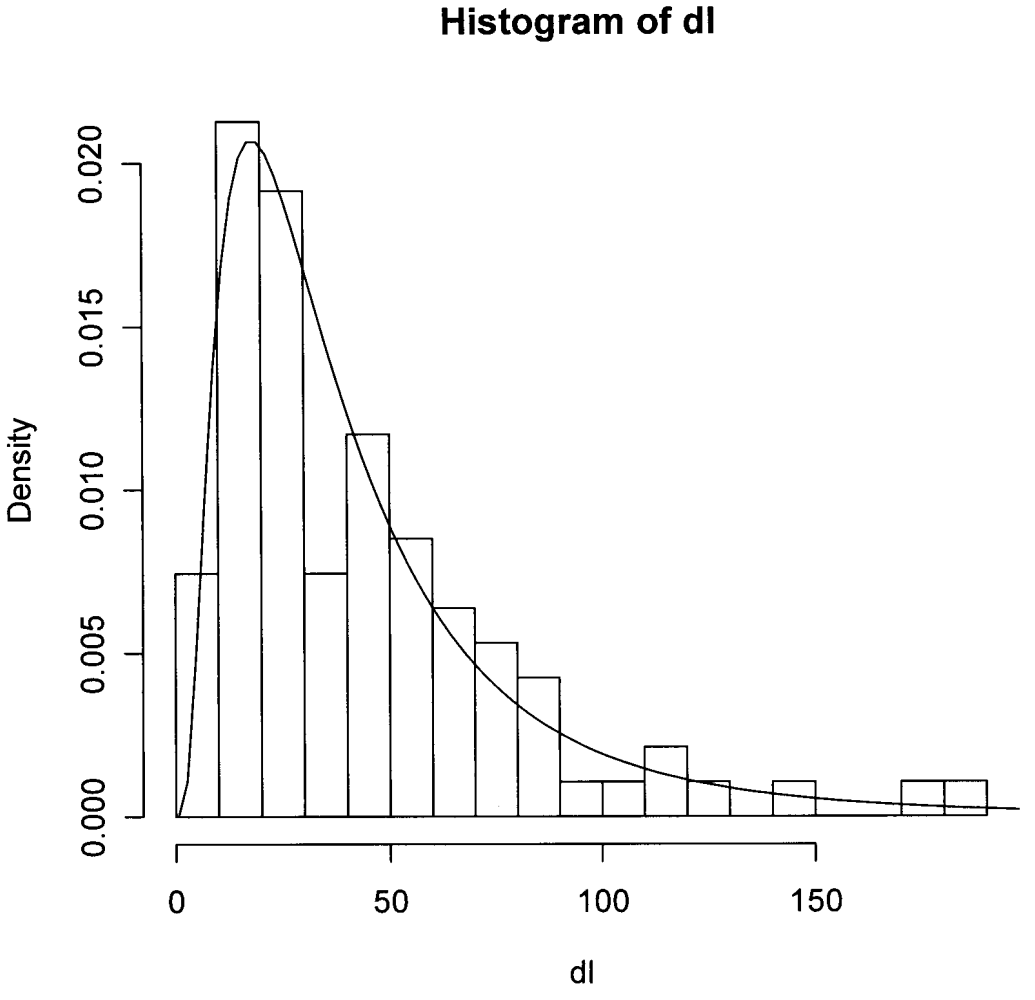
$$f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(\log x - \mu)^2}{2\sigma^2}\right)} \cdot \frac{1}{x}$$

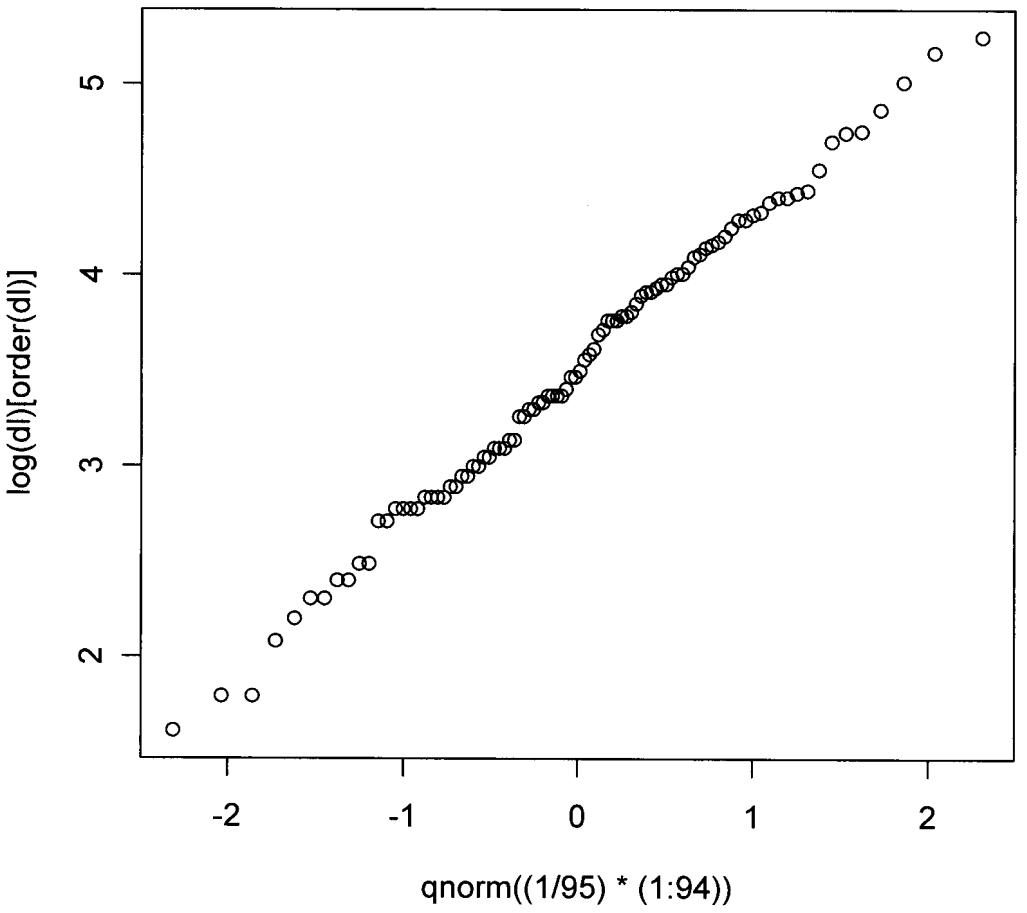
See the discussion on pg 187, example G of Rice for an heuristic motivation for the lognormal model.

The idea is that a taxonomic split changes species size by a random proportionality factor.

```
 #Problem 54
>
> octopods<-read.table("http://math.cmc.edu/moneill/Math152/Handouts/OCTO.txt",header=T)
>
>
> attach(octopods)
>
> mu<-mean(log(dl))
>
> sig<-sd(log(dl))
>
> #The histogram with the fitted lognormal density is attached:
>
> hist(dl,prob=T,breaks=20)
>  curve(dnorm(log(x),mean=mu,sd=sig)*(1/x),add=T)
Warning message:
NaNs produced in: log(x)
>
> length(dl)
[1] 94
> #[1] 94
>
>
> #The built in normality test which is similar to the one we developed in class
> #gives a p-value of 80% for the hypothesis that the log
> #of the dorsal lengths is normally distributed:
>
> shapiro.test(log(dl))

        Shapiro-Wilk normality test

data:  log(dl)
W = 0.9914, p-value = 0.8068


>
> #The probability plot is also attached:
>
> plot(qnorm((1/95)*(1:94)),log(dl)[order(dl)])
>
> #Now we apply the test for normality based on the correlation coefficient
> #of the probability plot:
>
> cor(qnorm((1/95)*(1:94)),log(dl)[order(dl)])
[1] 0.9973262
>
>
> f<-function(x){l<-rnorm(94)
+ cor(qnorm((1/95)*(1:94)),l[order(l)])
+ }
>
>
> d<-sapply(1:10000,f)
> d[order(d)[50]]
[1] 0.9753753
> d[order(d)[100]]
[1] 0.9786927
> d[order(d)[250]]
[1] 0.982547
> d[order(d)[500]]
[1] 0.9850412
> d[order(d)[1000]]
[1] 0.9880045
> d[order(d)[2000]]
[1] 0.990577
> d[order(d)[2500]]
[1] 0.9914204
> d[order(d)[5000]]
```

```
[1] 0.9941973
> d[order(d)[7500]]
[1] 0.995918
> d[order(d)[9000]]
[1] 0.9969936
> d[order(d)[9500]]
[1] 0.9974473
> d[order(d)[9750]]
[1] 0.9977948
> d[order(d)[9900]]
[1] 0.9981282
> d[order(d)[9950]]
[1] 0.998346
>
> #We see that the p-value is slightly less than .95. The lognormal model fits
> #quite well, and if anything, perhaps a little too well.
```


Histogram of dl
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
dl


log(dl)[order(dl)]
2
3
4
5
-2
-1
0
1
2
qnorm((1/95) * (1:94))