Statistics 2 Unit 5

Team 8

May 1, 2018

Contents

1 Task 85

1.1 a)

We are testig simple against composite hypothesis, thus the generalized likelihood ratio is in the following form:

$$\Lambda = \frac{\binom{n}{x}(0.5)^n}{sup_{p\in(0,1)}\binom{n}{x}p^x(1-p)^{n-x}}$$

Maximizing this the denominator first, we can find

$$\begin{aligned} log(\binom{n}{x}p^x(1-p)^{n-x}) &= const + xlog(p) + (n-x)log(1-p) \\ \frac{\partial l(f)}{\partial p} &= \frac{x}{p} - \frac{n-x}{1-p} = 0 \\ x - x - np + xp &= 0 \\ p &= \frac{x}{n} \end{aligned}$$

Thus, the GLR is the following expression:

$$\begin{aligned} \Lambda &= \frac{\binom{n}{x}(0.5)^n}{\binom{n}{x}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \\ &= \frac{(0.5)^n}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \end{aligned}$$

1.2 b)

The test rejects the null for the small values of Λ, which can be the case if, after we make a substitution for $y = x - \frac{n}{2} \Leftrightarrow x = y + \frac{n}{2}$:

$$\begin{aligned} \Lambda &= \frac{(0.5)^n}{(\frac{y+\frac{n}{2}}{n})^{y+\frac{n}{2}}(1-\frac{y+\frac{n}{2}}{n})^{n-y+\frac{n}{2}}} \\ &= \frac{(\frac{1}{2})^n}{(\frac{1}{2}+\frac{y}{n})^{y+\frac{n}{2}}(\frac{1}{2}-\frac{y}{n})^{\frac{n}{2}-y}} \end{aligned}$$

As $\Lambda(y)$ is symmetric and achieves its maximum at 0, we can conclude, that the null hypothesis is rejected for large values of $|x - \frac{n}{2}|$.

1.3 c)

Under the null the cdf of X is $f(x) = \binom{n}{x}(0.5)^n$ and the significance level is defined as:

$$\begin{aligned}
\alpha = P(X > c|H_0) &= 1 - F(c) \\
P(|X - \frac{n}{2}| > c) &= P(c + \frac{n}{2} < X) + P(X < -c + \frac{n}{2}) \\
&= 1 - F(c + \frac{n}{2}) + F(-c + \frac{n}{2}) \\
&= \sum_{c+\frac{n}{2}+1}^{n} \binom{n}{x}(0.5)^n + \sum_{0}^{-c+\frac{n}{2}-1} \binom{n}{x}(0.5)^n
\end{aligned}$$

1.4 d)

Having $n = 10$ and $c = 2$, we can compute the significance level:

$$\begin{aligned}
\alpha = P(X > c|H_0) &= 1 - F(c) \\
P(|X - 5| > 2) &= (0.5)^{10}(\sum_{8}^{10} \binom{10}{x} + \sum_{0}^{2} \binom{10}{x}) \\
&= 0.109375
\end{aligned}$$

2 Task 86

a) TRUE
Because $L_0 > L_1$.

b) FALSE
Because we reject when p-value $<$ significance level. In our case is the other way around.

c) TRUE
p-value$<$significance level. so p-value < 0.06.

d) FALSE
P-value is the probability of rejecting the null hypothesis when the null is true.

e) FALSE
the p-value differs from the likelihood ratio.

f) FALSE

```
pchisq(8.5,4,lower.tail=FALSE)<0.05
```

```
## [1] FALSE
```

3 Task 87

a)

The task says that test rejects for large values of $\mid T \mid$. We can write the p - value (probability of rejection H_0, when H_0 was true):

$$\text{p-value} = P(\mid T \mid > 1.50 | H_0) = P(T > 1.50) + P(T < -1.50)$$

Next, according to the task T statistics has standard normal diustrubution under null hypothesis.

$$\text{p-value} = 1 - \Phi(1.50) + \Phi(-1.50) = 1 - \Phi(1.50) + 1 - \Phi(1.50) = 2 - 2\Phi(1.50) = 0.1336144$$

```
p_value = 2 - 2*pnorm(1.50)
```

```
p_value
```

```
## [1] 0.1336144
```

b)

Now, we do the same but for the test:

$$\text{p-value} = P(T > 1.50 | H_0) = 1 - \Phi(1.50) = 0.0668072$$

```
p_value2 = 1 - pnorm(1.50)

p_value2

## [1] 0.0668072
```

4 Task 88

In this exercise a significance level α is given for which a test statistics rejects if it is greater than a certain thresholds.

$$\alpha = \mathbb{P}\left(T > t_0\right).$$

Let's consider a monotone-increasing transformation of the test statistics and of the threshold:

$$\gamma = \mathbb{P}\left(g(T) > g(t_0)\right).$$

The exercise asks, if this monotone-increasing transformation still is a α-test, i.e. does it hold that $\alpha = \gamma$?

Yes, it holds, the monotone-increasing transformation is a level α test. Since the transformation of the test statistics and the threshold are monotone-increasing and it holds that

$$T > t_0 \overset{mon.incr}{\Longleftrightarrow} g(T) > g(t_0)$$

what is roughly the definition of a strictly increasing function.
Note that we can take inverses of monotone increasing functions. Thus, by taking the probability:

$$\mathbb{P}\left(g(T) > g(t_0)\right) = \mathbb{P}\left(T > t_0\right) = \alpha.$$

We see that a monotone-increasing transformation of the test-statistic and the threshold yields again a level α test.

5 Task 89

5.1 a)

Let the value of the observed T statistic be $T = t$. Under H_0,

$$p - \text{value} = P(T > t) = 1 - P(T \leq t) = 1 - F(t) \Rightarrow V = 1 - F(T)$$

5.2 b)

To show that V is uniformly distributed it is enough to show that F(T) is. We know that F is continuous, thus we can find:

$$\begin{aligned} P(F(T) \leq t) &= P(T \leq F^{-1}(t)) \\ &= F(F^{-1}(t)) = t \end{aligned}$$

Note that the probability lies in $[0; 1]$ and this shows that the domain of F is $[0; 1]$ which leads to the conclusion, that F is uniformly distributed on $[0; 1]$ and so is V.

5.3 c) and d)

$$\begin{aligned} P(V > 0.1|H_0) &= 1 - F(0.1) \\ &= 1 - 0.1 = 0.9 \\ P(V < \alpha|H_0) = 1 - P(V > \alpha|H_0) & \\ &= 1 - (1 - \alpha) \\ &= \alpha. \end{aligned}$$

6 Task 90

We retrieve the log-likelihood function from the probability density function of the Poisson distribution:

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

With the $\theta := (\lambda_1, \ldots, \lambda_n)$
Log-likelihood:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda_i} \tag{1}$$

$$l(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log\left(\frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}\right) \tag{2}$$

$$= \sum_{i=1}^{n} (x_i \log(\lambda_i) - \log(x_i!) - \lambda_i) \tag{3}$$

$$\frac{\partial l(\theta|x_1, \ldots, x_n)}{\partial \lambda_k} = -1 + \frac{x_k}{\lambda_k} \tag{4}$$

$$S(\theta) = \left(-1 + \frac{x_1}{\lambda_1}, \ldots, -1 + \frac{x_n}{\lambda_n}\right)' \tag{5}$$

By setting all first partial derivative w.r.t. λ to zero, we obtain $\forall k \in \{1, \ldots, n\}$ as the MLE: $\lambda_k^{MLE} = x_k$ but with the null: $\lambda_1 = \cdots = \lambda_n$, we obtain the usual MLE of a Poisson: $\lambda^{MLE} = \bar{X}$.
Fisher information matrix:

$$I_{ij}(\theta) = -\frac{1}{n} E\left(\frac{\partial^2 l}{\partial \lambda_i \partial \lambda_j}\right) = \begin{cases} 0, & \text{if } i \neq j \\ \frac{\bar{X}}{\lambda_i^2}, & \text{if } i = j \text{Because the } X_i \text{ are Poisson} \end{cases} \tag{6}$$

$$\Rightarrow \qquad I(\theta) = \operatorname{diag}\left(\frac{\bar{X}}{\lambda_1^2}, \ldots, \frac{\bar{X}}{\lambda_n^2}\right) \tag{7}$$

$$I(\theta)^{-1} = \operatorname{diag}\left(\frac{\lambda_1^2}{\bar{X}}, \ldots, \frac{\lambda_n^2}{\bar{X}}\right) \tag{8}$$

Test statistic:

$$T_S = S(\theta)' I(\theta)^{-1} S(\theta) = \sum_{i=1}^{n} \left(\frac{\lambda_i^2}{\bar{X}} - \frac{\lambda_i x_i}{\bar{X}} - \frac{\lambda_i x_i}{\bar{X}} + \frac{x_i^2}{\bar{X}}\right) \tag{9}$$

$$= \sum_{i=1}^{n} \frac{(x_i - \lambda_i)^2}{\bar{X}} \tag{10}$$

By using $\lambda_i^{MLE} = \bar{X}$ we get:

$$T_S = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\bar{X}} \tag{11}$$

7 Task 91

We will use chi-squared test in order to answer the question in the task. First of all, in order to check whether the temporal trend takes place, we should state that our data (observations of number of bites) is uniformly distributed, i.e. every day we observe the same number of bites, does not matter which lunar period we have. Let's formulate the hypotheses:

H_0 : Number of bites has uniform distribution within lunar cycle
H_A : There is a temporal trend in the number of bites.

Now, let's recall the formula for Pearson's chi-squared statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where
O_i is the number of observations of period i.
E_i is expected frequency of period i.
n is the number of periods in the lunar cycle.

The expected frequency will be:

$$E_i = Np_i$$

For the dicrete uniform distribution we have that probability of every observation is equal. So, we have 29 days, hence probability for the period $i = 3/29$, where $i \in [1, 9]$, and for the last period we have only 2 days, so the probability will be $2/29$.

Now, we will calculate everything what we described above:

```
period <- seq(1:10)

observations <- c(137, 150, 163, 201, 269, 155, 142,
   146, 148, 110)

prob_uniform<-c(rep(3/29,9), 2/29)

ev<-prob_uniform*sum(observations)

f<-function(x) ((observations[x] - ev[x])^2)/ev[x]

test_Chi<-sum(sapply(period, f))

test_Chi

## [1] 85.47975

p_value <- pchisq(test_Chi, df = length(observations)
   - 1, ncp = 0, lower.tail = FALSE, log.p = FALSE)

p_value

## [1] 1.308432e-14
```

Since p-value is less than 0.05 (we test with the confidence level 95% and consequently $\alpha = 0.05$), we reject H_0 that number of bites is uniformly distributed.

8 Task 92

We have given a multinomial distribution with only two cells. First, there are two important relationships to note. We have $n = X_1 + X_2$ and $1 = p_1 + p_2$. Thus we can substitute in our formula as follows and derive the expression from the exercise:

$$X^2 = \sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{((n - X_1) - n(1 - p_1))^2}{n(1 - p_1)}$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(np_1 - X_1)^2}{n(1-p_1)}$$
$$= \frac{(1-p_1)(X_1 - np_1)^2 + p_1(np_1 - X_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1(1-p_1)}.$$

And well, this is just the expression we should derive!

Let us now show that

$$\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

is approximately standard normal.
First note that the X_1 is the number of observations, i.e. the sum of observed Bernoulli-variables (e.g. 1 if observed and 0 if not). The mean of a Bernoulli distribution is just $\mu = p$ and the standard deviation is just $\sigma = \sqrt{p(1-p)}$. By taking that into account, our desired result follows straight from the Central Limit Theorem.
Recall:
Let $R_1, ..., R_n$ be iid random variables with mean μ, variance σ^2 and a finite third moment. Denote the sum of the random variables as X_1. Then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 - n\mu}{\sqrt{n}\sigma} \leq z\right) = \Phi(z), \quad z \in (-\infty, \infty).$$

Thus, by plugging in the mean and the standard deviation of the Bernoulli distribution, the expression

$$\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

is approximately standard normal by the CLT (if the observations are independent (that is namely not mentioned in the exercise)).

The square of a standard normal variable is just $\mathcal{X}^2$ distributed with 1 degree of freedom. Since we have shown that the expression - which is just the square root of the X^2 - is approximately normal, X^2 has to be $\mathcal{X}^2$ distributed.

9 Task 93

The null is "p's are all the same", thus one can write the GLR as follows:

$$\Lambda = \frac{sup_{\Theta_0} \prod_{i=1}^{m} \binom{n_i}{x_i} (p)^{x_i}(1-p)^{n_i - x_i}}{sup_{p \in \Theta_0 \cup \Theta_A} \prod_{i=1}^{m} \binom{n_i}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}}$$

Maximizing this the nominator first, we can find

$$log(\prod_{1}^{m}\binom{n_i}{x_i}p^{x_i}(1-p)^{n_i-x_i}) = const + \sum_{i=1}^{m}x_i log(p) + \sum_{i=1}^{m}(n_i - x_i)log(1-p)$$

$$\frac{\partial l(f)}{\partial p} = \frac{\sum_{i=1}^{m}x_i}{p} - \frac{\sum_{i=1}^{m}(n_i - x_i)}{1-p} = 0$$

$$p\sum_{i=1}^{m}(n_i - x_i) = \sum_{i=1}^{m}x_i - p\sum_{i=1}^{m}x_i$$

$$p = \frac{\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}n_i}$$

while when we treat p's to be unequal and maximize wrt to each, we get $p_i = \frac{x_i}{n_i}$ Thus, we have:

$$T = \frac{\prod_{i=1}^{m}\binom{n_i}{x_i}(\frac{\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}n_i})^{x_i}(1-\frac{\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}n_i})^{n_i-x_i}}{\prod_{i=1}^{m}\binom{n_i}{x_i}(\frac{x_i}{n_i})^{x_i}(1-\frac{x_i}{n_i})^{n_i-x_i}}$$

$$= \frac{\prod_{i=1}^{m}(\frac{\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}n_i})^{x_i}(1-\frac{\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}n_i})^{n_i-x}}{\prod_{i=1}^{m}(\frac{x_i}{n_i})^{x_i}(1-\frac{x_i}{n_i})^{n_i-x_i}} \sim \chi^2_{m-1}$$

*According to Wilk's theorem, which states that as n goes to infinity the $-2log(\Lambda)$ is assymptotically Chi-squared distributed with number of parameters of alternative H minus number of parameters of null H, in our case m-1.

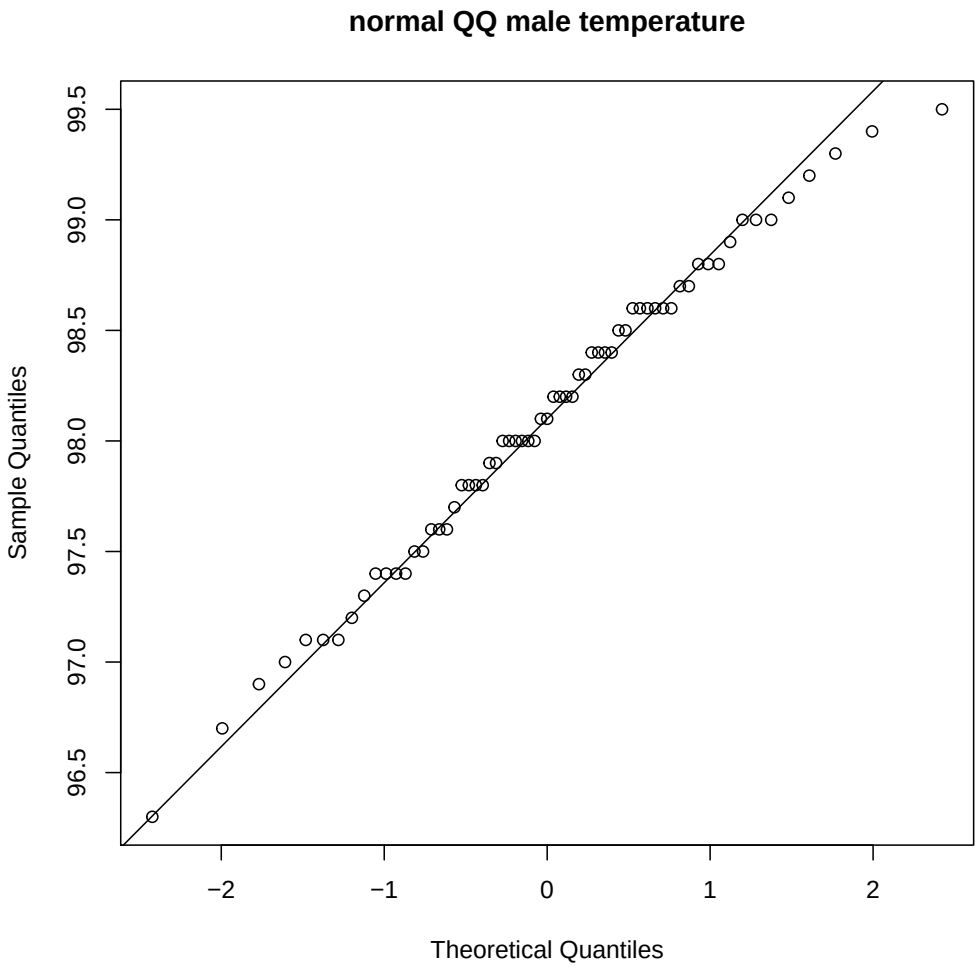
10 Task 98

```
bodytemp <- read.table(file.choose(), header = TRUE,
   sep = ",",  dec = ".")

head(bodytemp)

##   temperature gender rate
## 1        96.3      1   70
## 2        96.7      1   71
## 3        96.9      1   74
## 4        97.0      1   80
## 5        97.1      1   73
## 6        97.1      1   75

temperature <- as.vector(bodytemp[, 1])

gender <- as.vector(bodytemp[, 2])

rate <- as.vector(bodytemp[, 3])
```

a)

MALE:

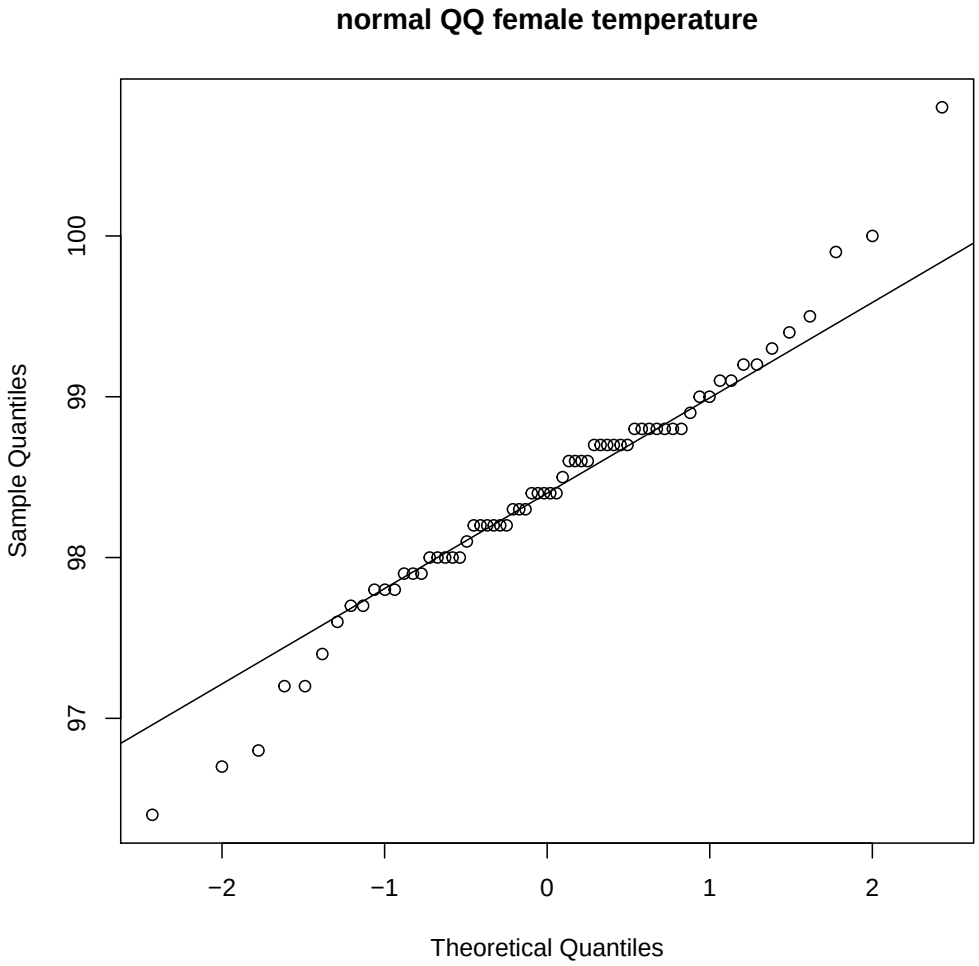
```
qqnorm(bodytemp[1:65, 1], main= "normal QQ male
   temperature"  )

qqline(bodytemp[1:65, 1])
```



So it seems they are normally distributed, even if the tails are slightly heavier.

FEMALE:

```
qqnorm(bodytemp[65:130, 1], main= "normal QQ female
   temperature")

qqline(bodytemp[65:130, 1])
```



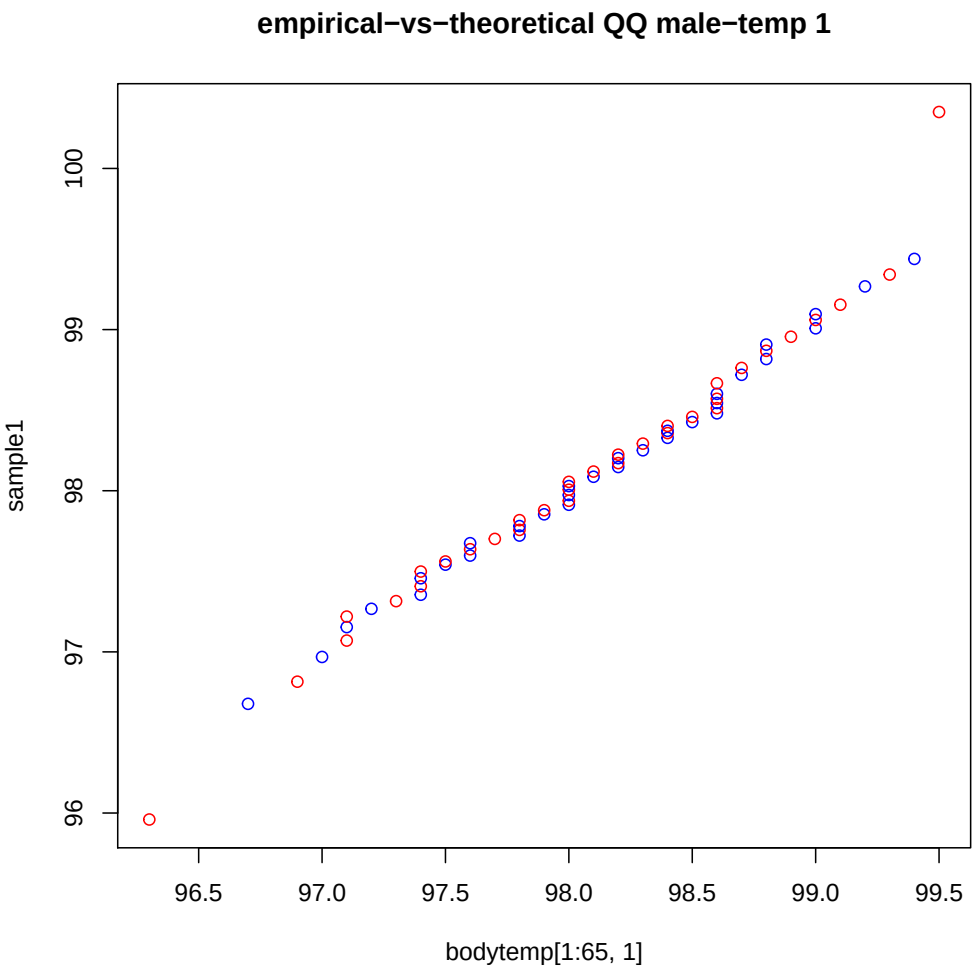
In the female case tails are even heavier and there seem to be some outliers.

We can now use the bootsstrapping method to get normal samples and compare them with the initial data.

MALE

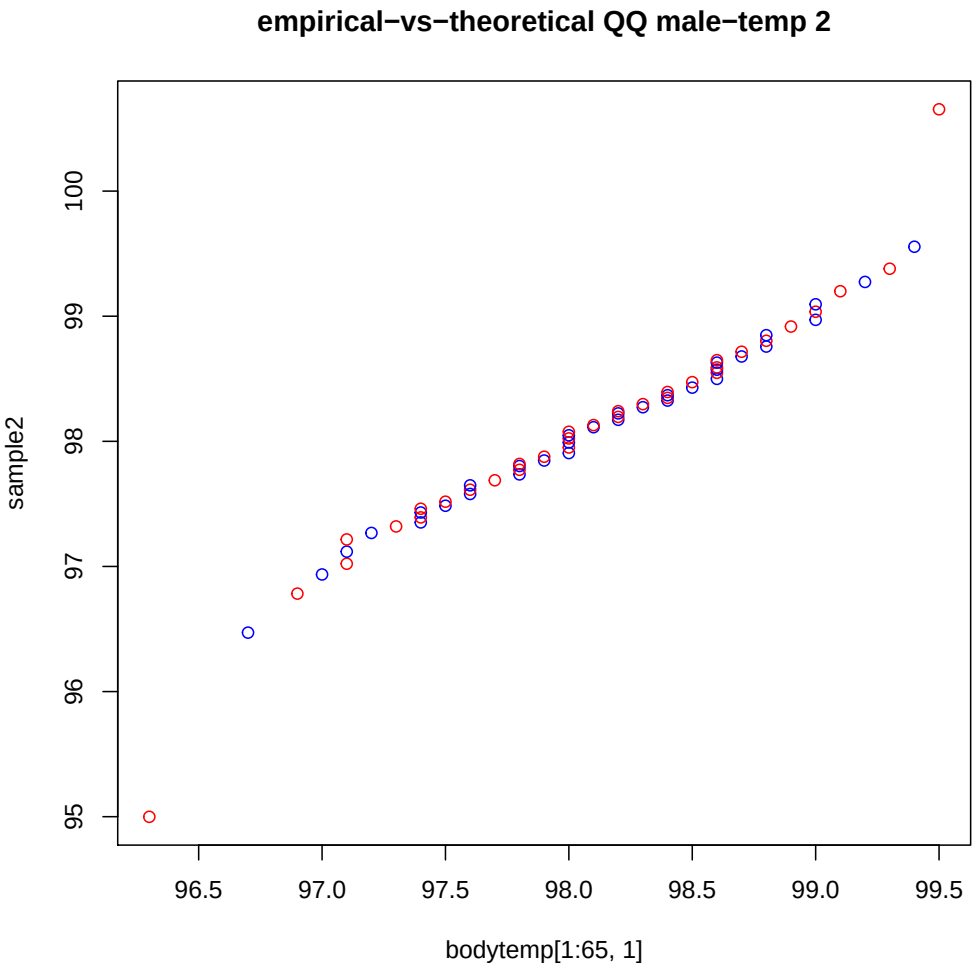
```
mean_m<-mean(bodytemp[1:65,1])

sd_m<-sd(bodytemp[1:65,1])

sample1<-rnorm(1000, mean_m, sd_m)
```

```
qqplot(bodytemp[1:65,1], sample1, col= c("red", "blue
   "), main= "empirical-vs-theoretical QQ male-temp 1"
    )
```



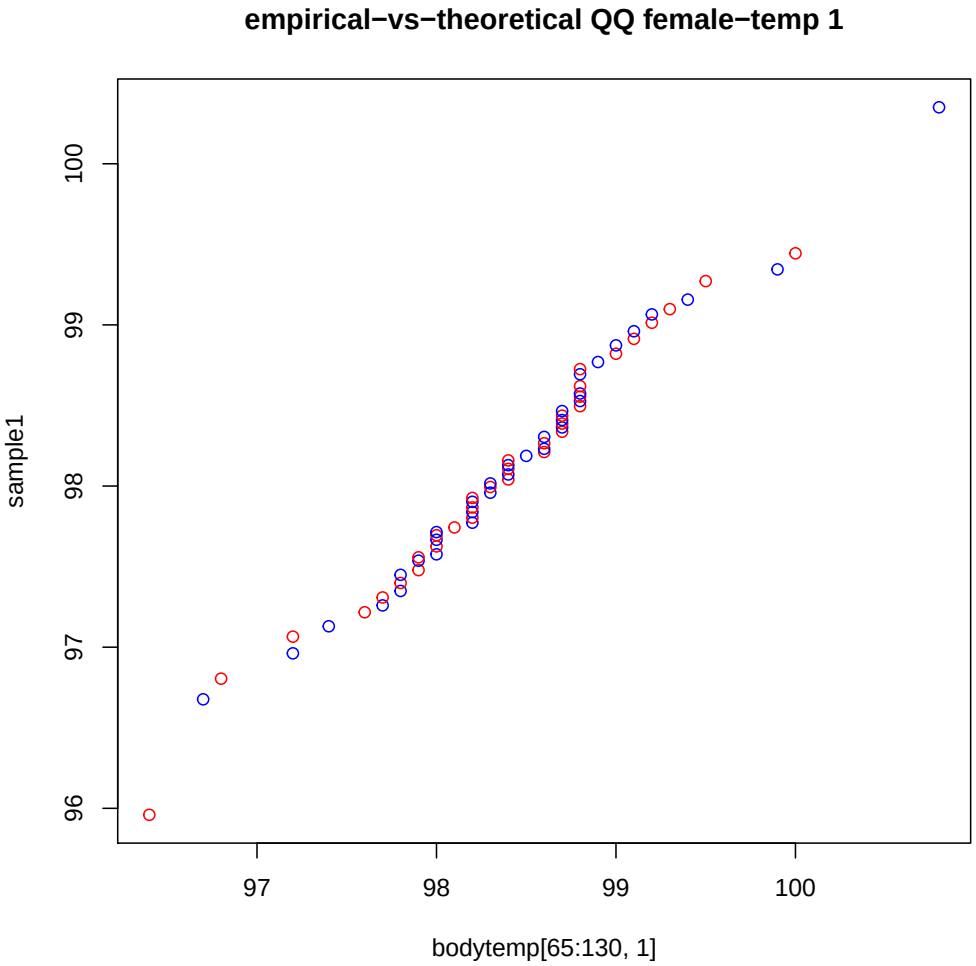
With red: the empirical data, blue: normalized data.

```
sample2<-rnorm(1000, mean_m, sd_m)

qqplot(bodytemp[1:65,1], sample2, col= c("red", "blue
   "), main= "empirical-vs-theoretical QQ male-temp 2"
     )
```



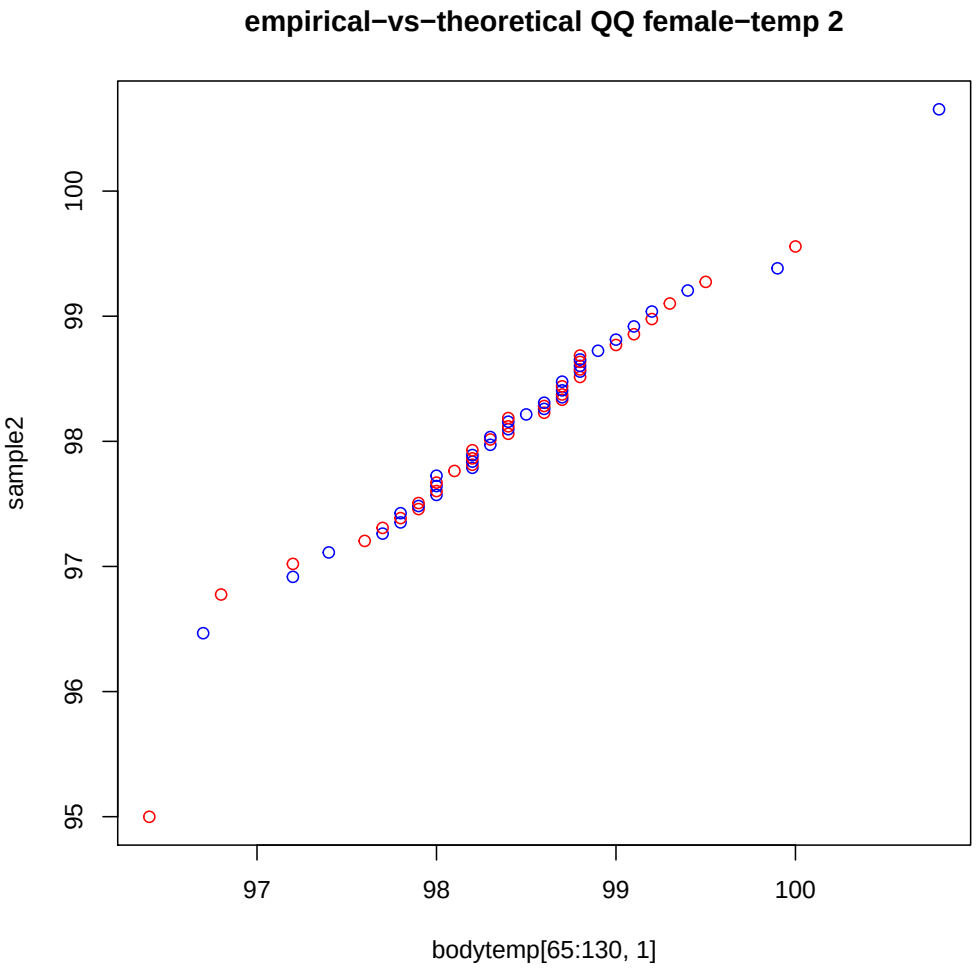
FEMALE

```
mean_f<-mean(bodytemp[65:130,1])

sd_f<-sd(bodytemp[65:130,1])

sample3<-rnorm(1000, mean_f, sd_f)

qqplot(bodytemp[65:130,1], sample1, col= c("red", "
   blue") , main= "empirical-vs-theoretical QQ female-
   temp 1" )
```



With red: the empirical data, blue: normalized data.

```
sample4<-rnorm(1000, mean_f, sd_f)

qqplot(bodytemp[65:130,1], sample2, col= c("red", "
   blue"), main= "empirical-vs-theoretical QQ female-
   temp 2" )
```

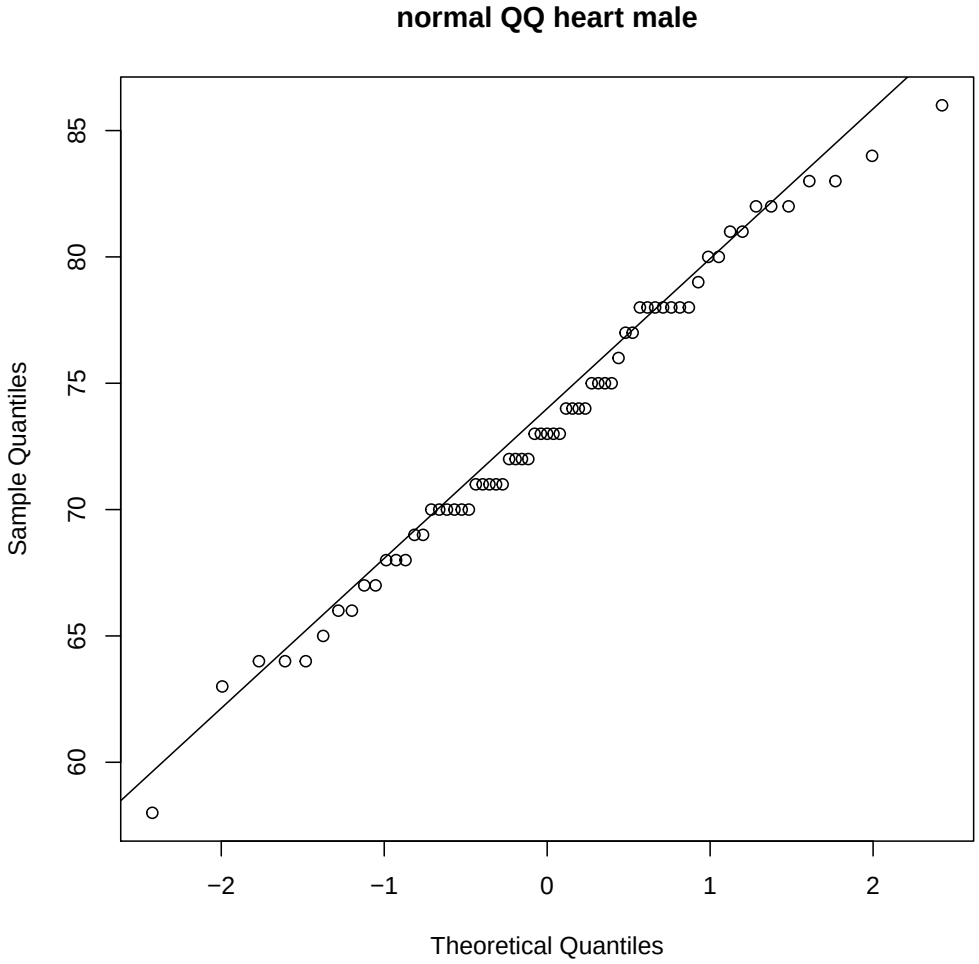


It seems that we have obtained not a perfect but a reasonable fit despite heavier tails.

b)

MALE:

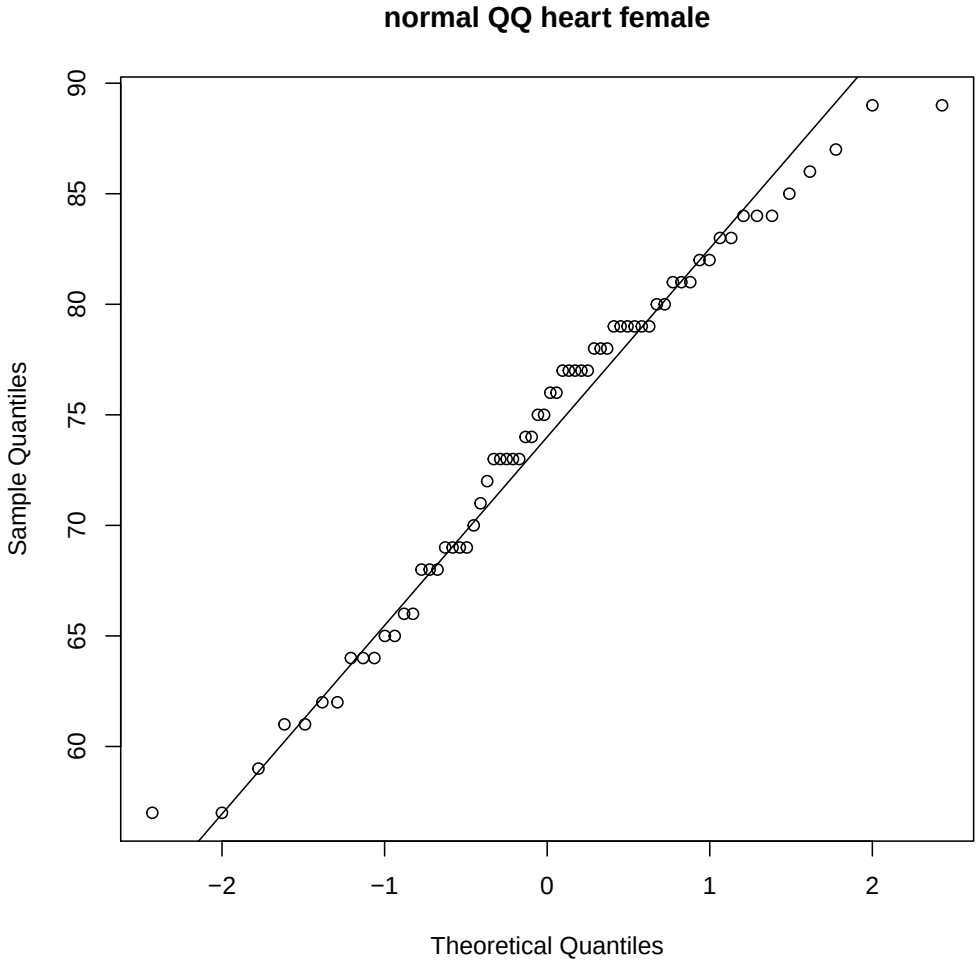
```
qqnorm(bodytemp[1:65, 3], main= "normal QQ heart male
   ")

qqline(bodytemp[1:65, 3])
```



So it seems they are normally distributed.

FEMALE:

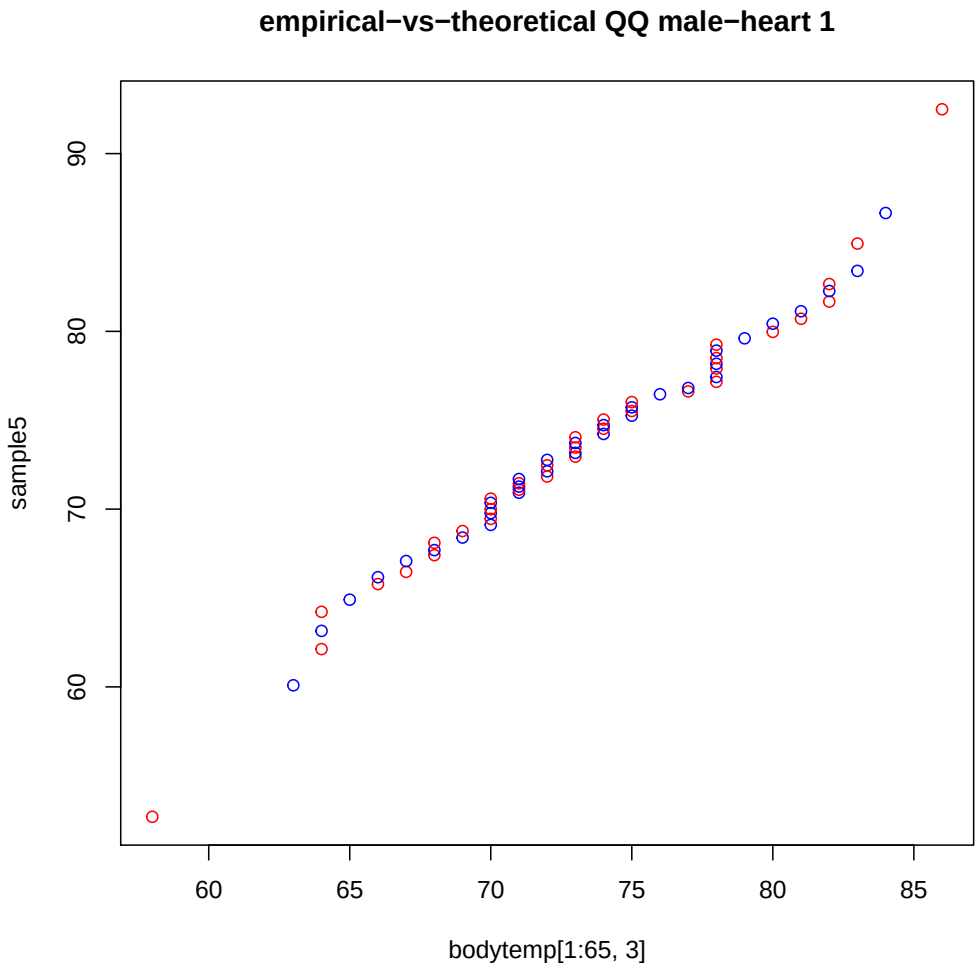
```
qqnorm(bodytemp[65:130, 3], main= "normal QQ heart
   female")

qqline(bodytemp[65:130, 3])
```



In the female case tails are heavy!

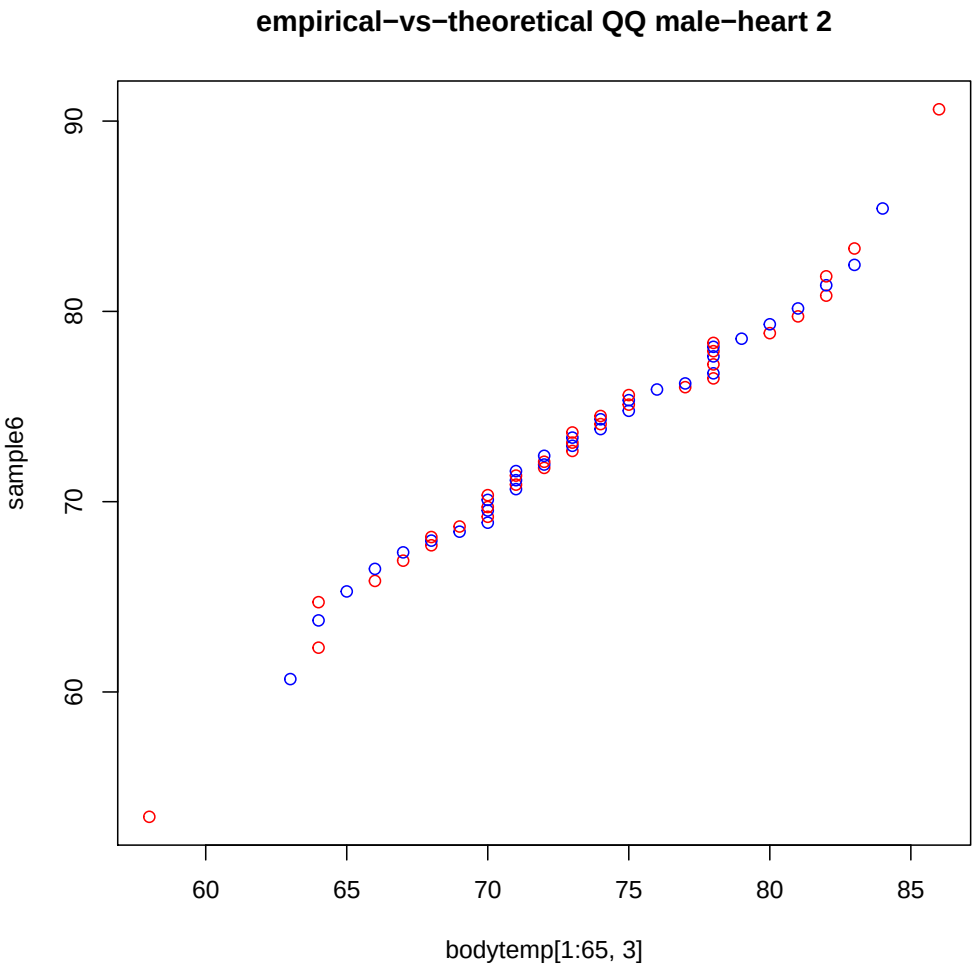
MALE

```
mean2_m<-mean(bodytemp[1:65,3])

sd2_m<-sd(bodytemp[1:65,3])

sample5<-rnorm(1000, mean2_m, sd2_m)

qqplot(bodytemp[1:65,3], sample5, col= c("red", "blue
   ") , main= "empirical-vs-theoretical QQ male-heart
   1" )
```



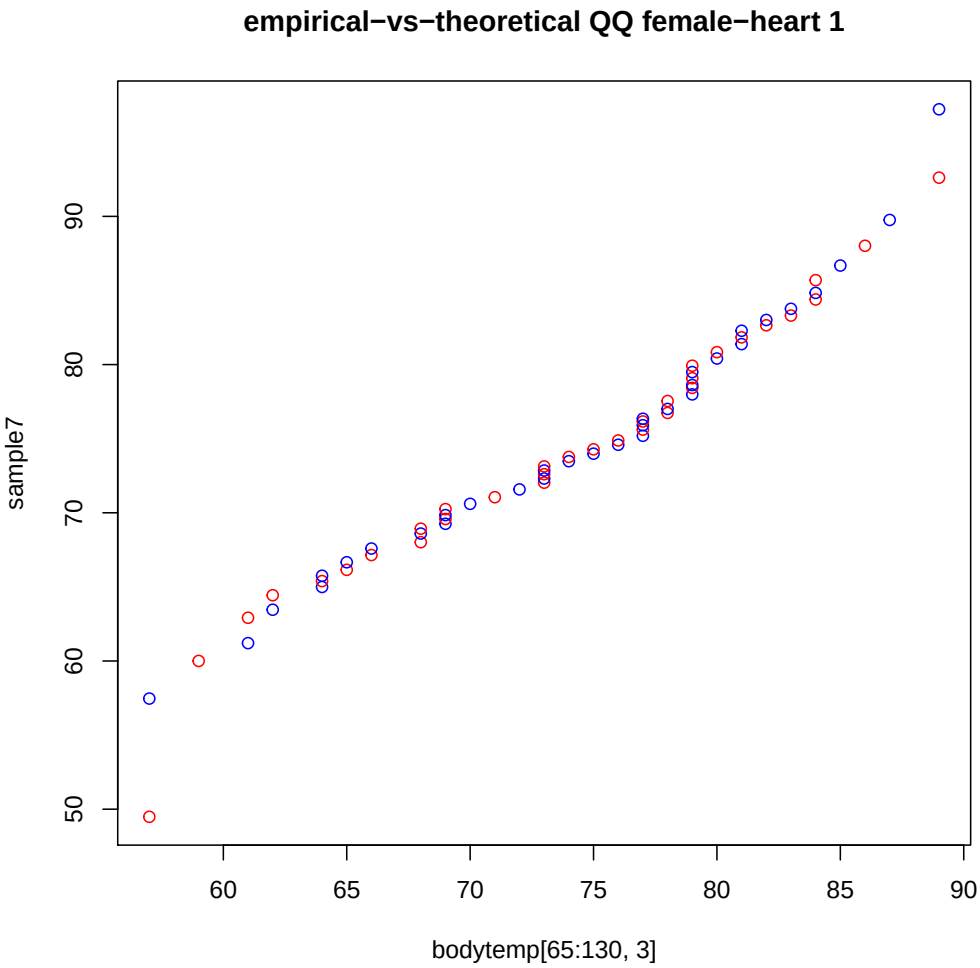
With red: the empirical data, blue: normalized data.

```
sample6<-rnorm(1000, mean2_m, sd2_m)

qqplot(bodytemp[1:65,3], sample6, col= c("red", "blue
   "), main= "empirical-vs-theoretical QQ male-heart
   2" )
```



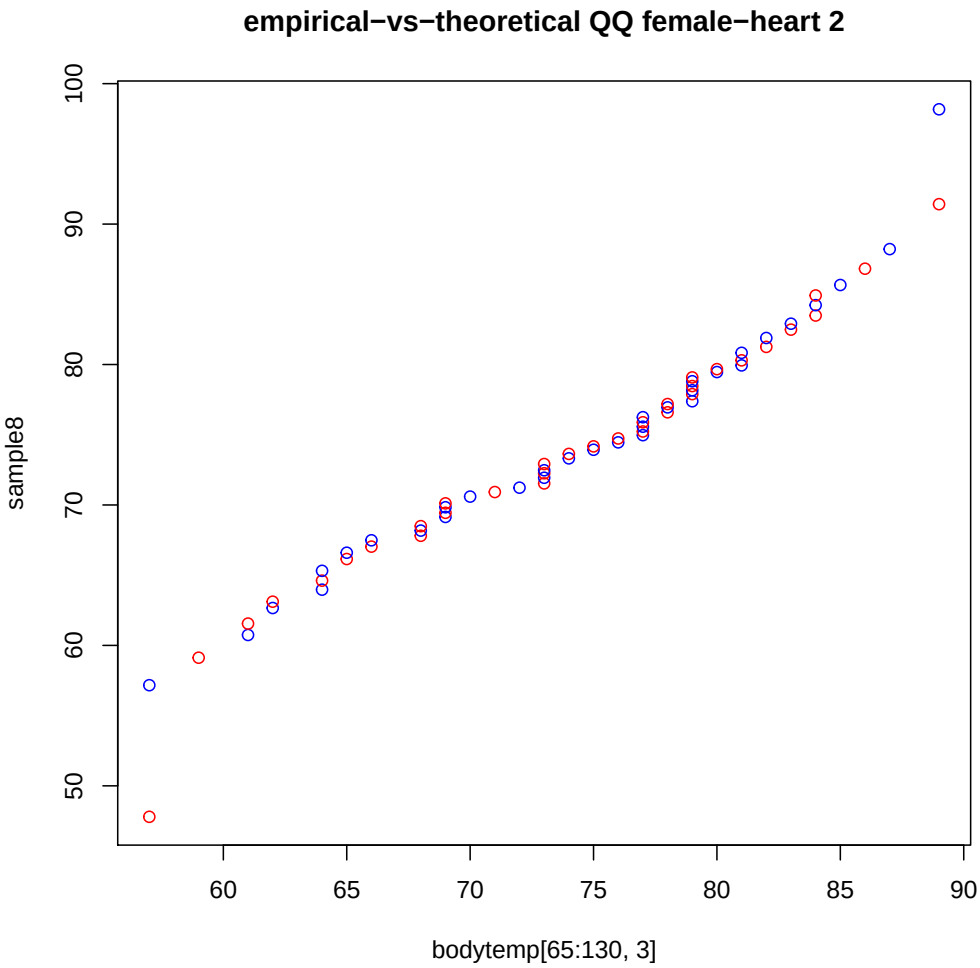
FEMALE

```
mean2_f<-mean(bodytemp[65:130,3])

sd2_f<-sd(bodytemp[65:130,3])

sample7<-rnorm(1000, mean2_f, sd2_f)

qqplot(bodytemp[65:130,3], sample7, col= c("red", "
   blue"), main= "empirical-vs-theoretical QQ female-
   heart 1" )
```



With red: the empirical data, blue: normalized data.

```
sample8<-rnorm(1000, mean2_f, sd2_f)

qqplot(bodytemp[65:130,3], sample8, col= c("red", "
   blue"), main= "empirical-vs-theoretical QQ female-
   heart 2" )
```



In both cases we can say the distributions fit the normal one quite well.

c)

Since the data seems to follow normal distribution, we can use a t-test.

MALE:

```
mu0=98.6

t.test(bodytemp[1:65, 1], mu=mu0)

##
```

```
##      One Sample t-test
##
## data:  bodytemp[1:65, 1]
## t = -5.7158, df = 64, p-value = 3.084e-07
## alternative hypothesis: true mean is not equal to
   98.6
## 95 percent confidence interval:
##  97.93147 98.27776
## sample estimates:
## mean of x
##  98.10462
```

Therefore we reject the null hypothesis.

FEMALE

```
t.test(bodytemp[65:130, 1], mu=mu0)

##
##      One Sample t-test
##
## data:  bodytemp[65:130, 1]
## t = -2.051, df = 65, p-value = 0.04431
## alternative hypothesis: true mean is not equal to
   98.6
## 95 percent confidence interval:
##  98.22618 98.59503
## sample estimates:
## mean of x
##  98.41061
```

We, again, reject the null. however, if we would have chosen a 99% confidence interval, we would have accepted the null.