qbs124_hw1_gibran

Gibran Erlangga

4/2/2022

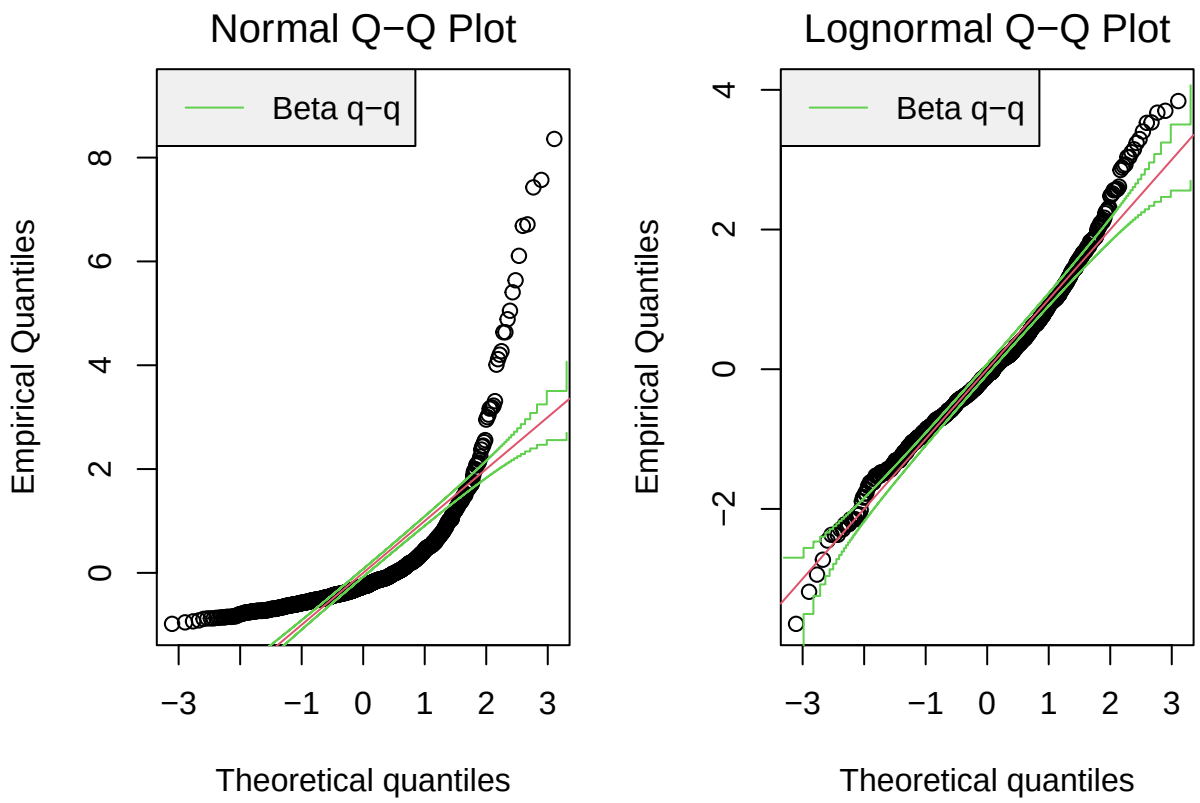
Question 1

(10 points). Modify toears code to display the q-q plot for testing the normal and lognormal distribution along with the 95% beta q-q confidence band. Use par(mfrow=c(1,2), mar=c(4.5,4.5,3,1), cex.lab=1, cex.main=1.5). Comment on the usefulness of the log transformation.

```
data <- scan('toears.txt')
n = length(data)
log_data = log(data)
ii=1:n
ii_reverse = n-ii+1
quantile = qnorm((1:n)/n)
thq=qnorm(((1:n)-.5)/n)
q=seq(from=-8,to=8,length=n)
data = data[order(data)]
data_plot = (data-mean(data))/sd(data)
log_data_plot = (log_data-mean(log_data))/sd(log_data)

par(mfrow=c(1,2), mar=c(4.5,4.5,3,1), cex.lab=1, cex.main=1.5)
plot(quantile, data_plot, xlab="Theoretical quantiles",ylab="Empirical Quantiles")
mtext(side=3,"Normal Q-Q Plot",cex=1.25,line=.5)
abline(coef = c(0,1), col=2)
upB=qnorm(qbeta(.5+.95/2,shape1=ii,shape2=ii_reverse))
lowB=qnorm(qbeta(.5-.95/2,shape1=ii,shape2=ii_reverse))
lines(thq,(upB-mean(data_plot))/sd(data_plot),type="s",col=3)
lines(thq,(lowB-mean(data_plot))/sd(data_plot),type="s",col=3)
legend("topleft",c("Beta q-q"),col=3,lty=1,bg="gray94",cex=1)

plot(quantile, log_data_plot, xlab="Theoretical quantiles",ylab="Empirical Quantiles")
mtext(side=3,"Lognormal Q-Q Plot",cex=1.25,line=.5)
abline(coef = c(0,1), col=2)
upB=qnorm(qbeta(.5+.95/2,shape1=ii,shape2=n-ii+1))
lowB=qnorm(qbeta(.5-.95/2,shape1=ii,shape2=n-ii+1))
lines(thq,(upB-mean(log_data_plot))/sd(log_data_plot),type="s",col=3)
lines(thq,(lowB-mean(log_data_plot))/sd(log_data_plot),type="s",col=3)
legend("topleft",c("Beta q-q"),col=3,lty=1,bg="gray94",cex=1)
```



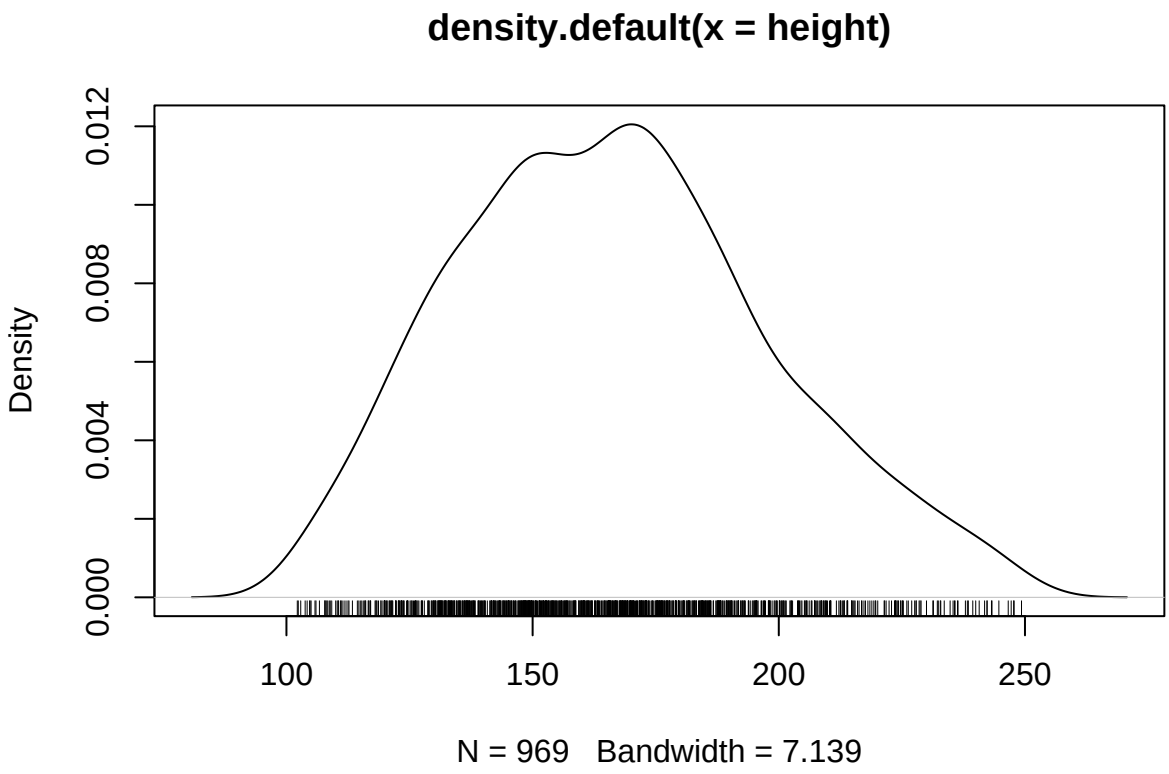
From both graphs above, we can see that after applying log transformation to the original, right-skewed data (shown by the left graph), we got a more normally distributed data (shown by the right graph). Therefore, log transformation is a great way to convert a skewed distribution into a normal distribution.

Question 2

(10 points). File height.csv contains height (cm) of random people. Use density to reconstruct and plot the pdf (use rug command to show the data). Explain the result. Estimate the number of people taller than 250 cm in town with population 100,000. Display the result on the graph. [Hint: See "Don't be late to the meeting" Example and jackM code.]

```
height_data <- read.csv('height1000.csv')
height <- height_data$height_cm

# plot density + rug
plot(density(height))
rug(height)
```




```
pop = 100000
num_ppl_more_than_250 = pnorm(250, mean=mean(height), sd=sd(height), lower.tail=F)*100000
num_ppl_more_than_250
```

```
## [1] 385.8122
```

Based on the dataset, the number of people taller than 250 cm in a town with 100,000 population is approximately 386.

Question 3.

(10 points). Compute three central tendency measures for original and log-transformed toears (don't forget to exponentiate the log-transformed centers). Explain the results: why some are different and some are close. Print out as the data frame with two columns and three rows. [Hint: Consult Section 2.11. Use arithmetic and geometric mean and their inequality, explain why the medians are the same.]

```
pdf_data = density(data)
log_data = log(data)
pdf_log_data = density(log_data)

# compute central tendency measures for original data
og_mean = mean(data)
og_median = median(data)
og_mode = pdf_data$x[pdf_data$y==max(pdf_data$y)]
```

```r
# compute central tendency measures for original data
log_mean = exp(mean(log_data))
log_median = exp(median(log_data))
log_mode = exp(pdf_log_data$x[pdf_log_data$y==max(pdf_log_data$y)])

labels <- c("Mean", "Median", "Mode")
origin_measures <- c(og_mean, og_median, og_mode)
log_measures <- c(log_mean, log_median, log_mode)

df <- data.frame(origin_measures, log_measures)
rownames(df) = labels
colnames(df) = c("Original toears", "Log-transformed toears")
df
```

```
##        Original toears Log-transformed toears
## Mean        0.11111306             0.08837587
## Median      0.08400000             0.08400000
## Mode        0.06600328             0.07397543
```

The table above shows the comparison of three central tendency measures (mean, median and mode) for toears data, both the original and the log-transformed one. We can observe that the median value is exactly the same while mean and mode values are slightly different. The median value of both data is exactly the same because it refers to the same data point (the middle point) in the data set. The mean value of both data (arithmetic mean for original data and geometric/multiplicative mean for back-transformed data) is different due to compounding effect, so the arithmetic mean will always be greater than geometric mean. The mode value of both data is different due to distribution changes (right-skewed to normal) after we applied the log-transformation to the data, therefore we took the data point which has the highest density on log-scale but not in the original scale.