MAT 2379, Introduction to biostatistics

Assignment 4

Due date: Friday, November 14, 2014 at 3 pm

(based on Chapters 9 and 10)

Total = 100 marks

Question 1: [15 points]

(a) The first quartile is $q_1 = 5.1$mmol/L, the third quartile is $q_3 = 5.7$mmol/L and the interquartile range is $\text{IQR} = q_3 - q_1 = 0.6$mmol/L.

(b) The lower fence is $q_1 - 1.5\text{IQR} = 5.1 - 1.5(0.6) = 4.2$ and the upper fence is $q_3 + 1.5\text{IQR} = 5.7 + 1.5(0.6) = 6.6$. The value 6.9 is the only outlier, since it is the only value that is outside of the fences.

(c) The sample standard deviation is

$$s = \sqrt{\frac{(\sum x_i^2) - (\sum x_i)^2/n}{n-1}} = \sqrt{\frac{(394.59) - (71.1)^2/13}{13-1}} = 0.69087.$$

(d) A point estimate for μ is $\overline{x} = 5.469$mmol/L and the (estimated) standard error of the estimate is $s/\sqrt{n} = 0.69087/\sqrt{13} = 0.1916$mmol/L.

Question 2: [15 points]

(a) Note that pOH is a linear transformation of pH. Thus, the mean pOH is $\overline{\text{pOH}} = 14 - \overline{\text{pH}} = 14 - 7.75 = 6.25$. The sample variance of the pOH is $s^2_{\text{pOH}} = (-1)^2 s^2_{\text{pH}} = (-1)^2 (1.57)^2 = 1.57^2$, and thus the sample standard deviation of the pOH is $s_{\text{pOH}} = \sqrt{s^2_{\text{pOH}}} = \sqrt{1.57^2} = 1.57$.

(b) Let $y = -\ln(10)\,\text{pH}$. So $\overline{y} = -\ln(10)\,\overline{\text{pH}} = -\ln(10)\,(7.75)$, since y is a linear transformation of pH. Futhermore, $y = -\ln(10)\,\text{pH} = \ln(x)$. Thus, exponentiating the mean of y will give the geometric mean of x. Thus, the geometric mean of the hydrogen ion activity is

$$g = e^{\overline{y}} = e^{-\ln(10)\,(7.75)} = 1.78 \times 10^{-8}.$$

Question 3: [15 points]

(i) (a) Since the estimated standard error of the mean is $s/\sqrt{n} = 1.378$, then the sample standard deviation is $s = 1.378\sqrt{n} = 1.378\sqrt{15} = 5.3370$.

(b) A 95% confidence interval for the mean water hardness is

$$\overline{x} \pm t\,\frac{s}{\sqrt{n}} = [99.1, 105.0],$$

where $t = t_{0.025,14} = 2.145$, $\overline{x} = 102.03$ and $s/\sqrt{n} = 1.378$.

(c) If the mean hardness is between 100 mg/l and 180 mg/l, then it can be classified as medium hard. Since not all values in the confidence interval are between 100 and 180, then we cannot conclude the mean hardness of the water is medium hard at a level of confidence of 95%.

(ii) A 80% confidence interval for the mean water hardness is

$$\overline{x} \pm t \frac{s}{\sqrt{n}} = [100.2, 103.9],$$

where $t = t_{0.10,14} = 1.345$, $\overline{x} = 102.03$ and $s/\sqrt{n} = 1.378$. We are 80% confident, that the mean hardness of the water is medium hard.

Question 4: [15 points]

(a) The calcium concentration is highly skewed to the right. The median calcium concentration is 39 ppm and the central half of the calcium concentrations are between 14 ppm and 75 ppm, which is an interquartile range of 61 ppm. The distribution of the mortality rates is approximately symmetric. The mean and medium mortality rates are 1524 deaths per 100 000 residents and 1555 deaths per 100 000 residents. the central half of the mortality rates are between 1379 deaths per 100 000 residents and 1668 deaths per 100 000 residents, which is an interquartile range of 289 deaths per 100 000 residents.

(b) For **calcium**: The lower fence is $q_1 - 1.5\text{IQR} = 14 - 1.5(61) = -77.5$ and the upper fence is $q_3 + 1.5\text{IQR} = 75 + 1.5(61) = 166.5$. Since the minimum and the maximum values are within the fences, then all the values are within the fences. There are no outlying calcium concentrations.

For **mortality**: The lower fence is $q_1 - 1.5\text{IQR} = 1379 - 1.5(289) = 945.5$ and the upper fence is $q_3 + 1.5\text{IQR} = 1668 + 1.5(289) = 2101.5$. Since the minimum and the maximum values are within the fences, then all the values are within the fences. There are no outlying mortality rates.

(c) (i) The town with the largest mortality is in the north.

(ii) The town with the smallest mortality is in the south.

(iii) The median mortality for the towns in the north is larger than the median mortality for the towns in the south.

(iv) By comparing interquartile ranges, the mortality is slightly more dispersed in the south.

Question 5: [15 points]

(a) By the Central Limit Theorem,

$$Z = \frac{\overline{X}_{\text{boys}} - 9.7}{6/\sqrt{40}}$$

has a standard normal distribution approximately. Thus,

$$\begin{aligned} P(9 < \overline{X}_{\text{boys}} < 10) &= P\left(\frac{9-9.7}{6/\sqrt{40}} < Z < \frac{10-9.7}{6/\sqrt{40}}\right) \\ &\approx \Phi(0.32) - \Phi(-0.74) = 0.6255 - 0.2296 = 0.3959. \end{aligned}$$

(b) Let $\overline{X}_{\text{girls}}$ be the sample mean for the $n = 50$ girls. By the Central Limit Theorem,

$$Z = \frac{\overline{X}_{\text{girls}} - 15.6}{9.5/\sqrt{50}}$$

has a standard normal distribution approximately. Thus,

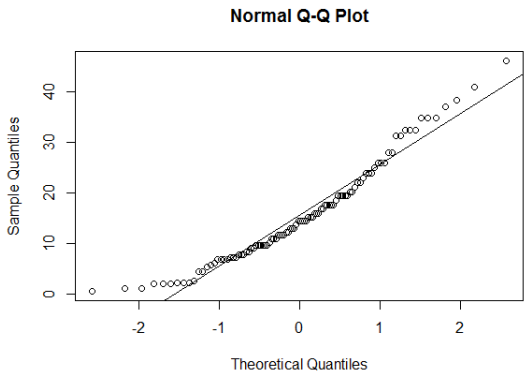
$$\begin{aligned} P(\overline{X}_{\text{girls}} > 18) &= 1 - P\left(Z < \frac{18-15.6}{9.5/\sqrt{50}}\right) \\ &\approx 1 - \Phi(1.79) = 1 - 0.9633 = 0.0367. \end{aligned}$$

Question 6: [25 points] In `R`, we assign the data to the dataframe `data` and display the names of the columns. We assign the lifetime to the variable x.

```
> data=read.table(file.choose(),header=TRUE,sep="\t")
> names(data)
[1] "lifetime..in.seconds."
> x=data$lifetime..in.seconds.
```

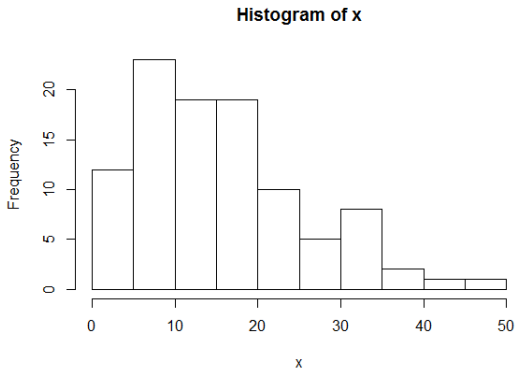
(a) Here are the commands to produce the quantile-quantile plot of the lifetime that is displayed below. There is a non-linear tendency in the plot. This is evidence against normality.

```
> qqnorm(x)
> abline(mean(x),sd(x))
```



Here is the command to produce the histogram of the lifetime that is displayed below. The lifetime is highly skewed to the right. This is evidence against normality.

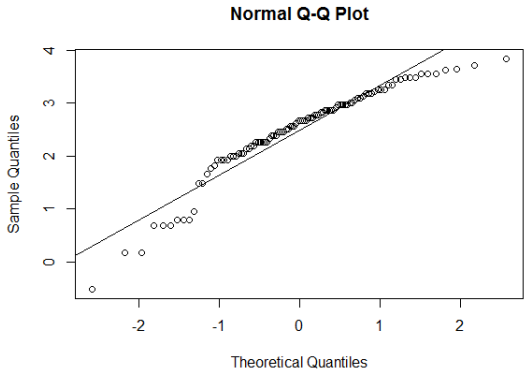
```
> hist(x)
```



We conclude that it is not reasonable to model the lifetime with a normal distribution.

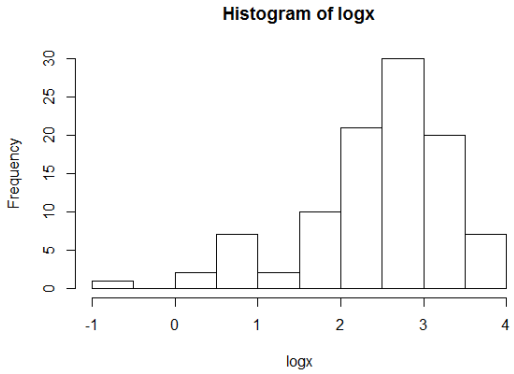
(b) Here are the commands to produce the quantile-quantile plot of the logarithmic lifetime that is displayed below. There is a non-linear tendency in the plot. This is evidence against normality.

```
> logx=log(x)
> qqnorm(logx)
> abline(mean(logx),sd(logx))
```



Here is the command to produce the histogram of the logarithmic lifetime that is displayed below. The logarithmic lifetime is highly skewed to the left. This is evidence against normality.

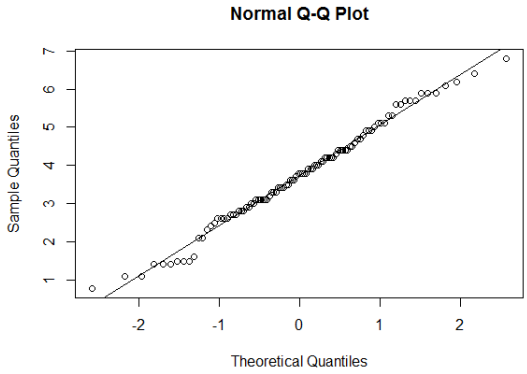
```
> hist(logx)
```



We conclude that it is not reasonable to model the logarithmic lifetime with a normal distribution.

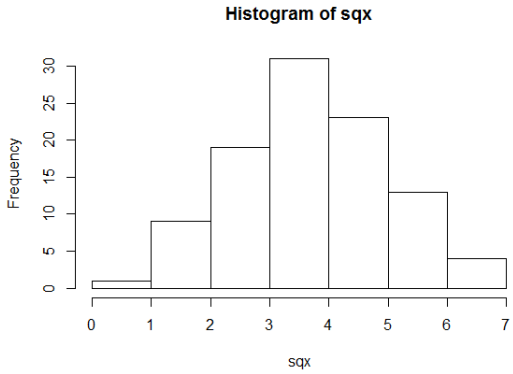
(c) Here are the commands to produce the quantile-quantile plot of the square root of the lifetime that is displayed below. There is a linear tendency in the plot. Based on this plot, we conclude that the square root of the lifetime is approximately normal.

```
> sqx=sqrt(x)
> qqnorm(sqx)
> abline(mean(sqx),sd(sqx))
```



Here is the command to produce the histogram of the square root of the lifetime that is displayed below. The histogram is approximately symmetric. This is further support that the square root of the lifetime is approximately normal.

```
> hist(logx)
```



We conclude that it is reasonable to model the square root of the lifetime with a normal distribution.