Verzani Problem Set

Next are considered the problems from Verzani's book on page 68.

Problem 10.1

Load the data set vacation. This gives the number of paid holidays and vacation taken by workers in the textile industry.

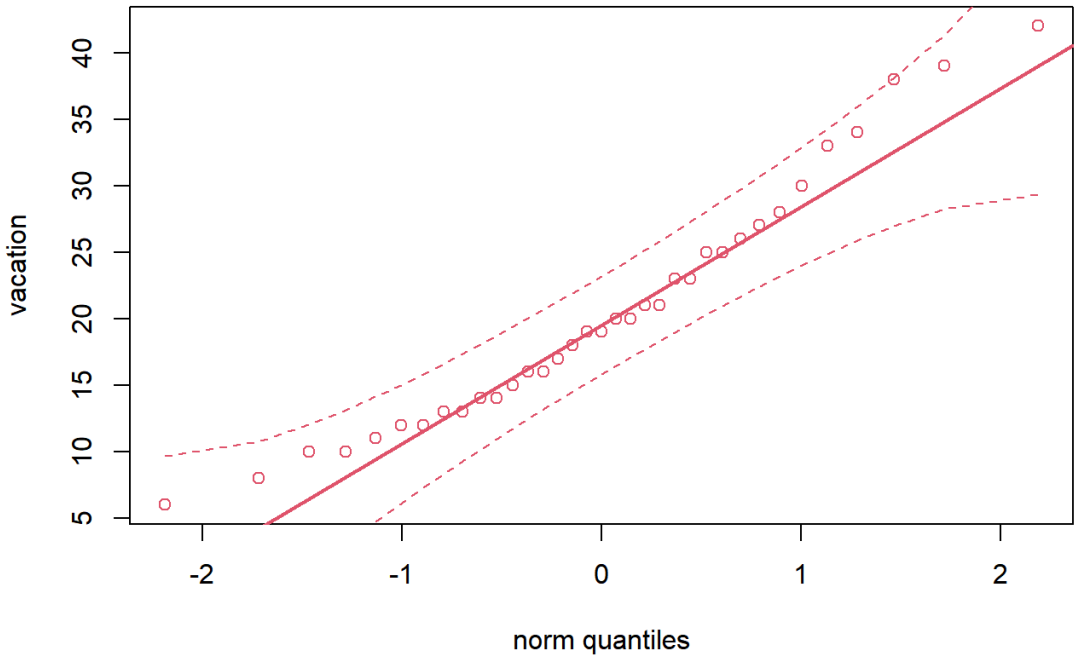
```
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':

    cancer
> head(vacation)
[1] 23 12 10 34 25 16
```

1. Is a test for $\overline{y}$ appropriate for this data?

H_0 : Y is normally distributed

H_A : Y isn't normally distributed

```
> library(StatDA)
Warning: package 'StatDA' was built under R version 4.0.3
Loading required package: sgeostat
Warning: package 'sgeostat' was built under R version 4.0.3
Registered S3 method overwritten by 'geoR':
  method         from
  plot.variogram sgeostat
> qqplot.das(vacation)
```




```
> shapiro.test(vacation)

    Shapiro-Wilk normality test

data:  vacation
W = 0.95272, p-value = 0.1374
```

The $p-value = 0.1374 > 0.05 = \alpha$, so we can assume that the data is normally distributed.

2. Does a t-test seem appropriate? Yes, because the data is normally distributed.
3. If so, test the null hypothesis that $\mu = 24$. (What is the alternative?)

$$H_0 : \mu = 24$$
$$H_A : \mu \neq 24$$

```
> t.test(vacation,
+        alternative = "two.sided",
+        mu = 24,
+        conf.level = 0.95)

    One Sample t-test

data:  vacation
t = -2.2584, df = 34, p-value = 0.03045
alternative hypothesis: true mean is not equal to 24
95 percent confidence interval:
 17.37768 23.65089
sample estimates:
```

```
mean of x
 20.51429
```

The $p-value = 0.03045 < 0.05 = \alpha$, so we reject H_0 and $\mu \neq 24$.

Test if the mean is greater than 24.

$$H_0 : \mu = 24$$
$$H_A : \mu > 24$$

```
> t.test(vacation,
+        alternative = "greater",
+        mu = 24,
+        conf.level = 0.95)

    One Sample t-test

data:  vacation
t = -2.2584, df = 34, p-value = 0.9848
alternative hypothesis: true mean is greater than 24
95 percent confidence interval:
 17.90448      Inf
sample estimates:
mean of x
 20.51429
```

The $p-value = 0.9848 > 0.05 = \alpha$, so we don't have a reason to reject H_0.

Test if the mean is smaller than 24.

$$H_0 : \mu = 24$$
$$H_A : \mu < 24$$

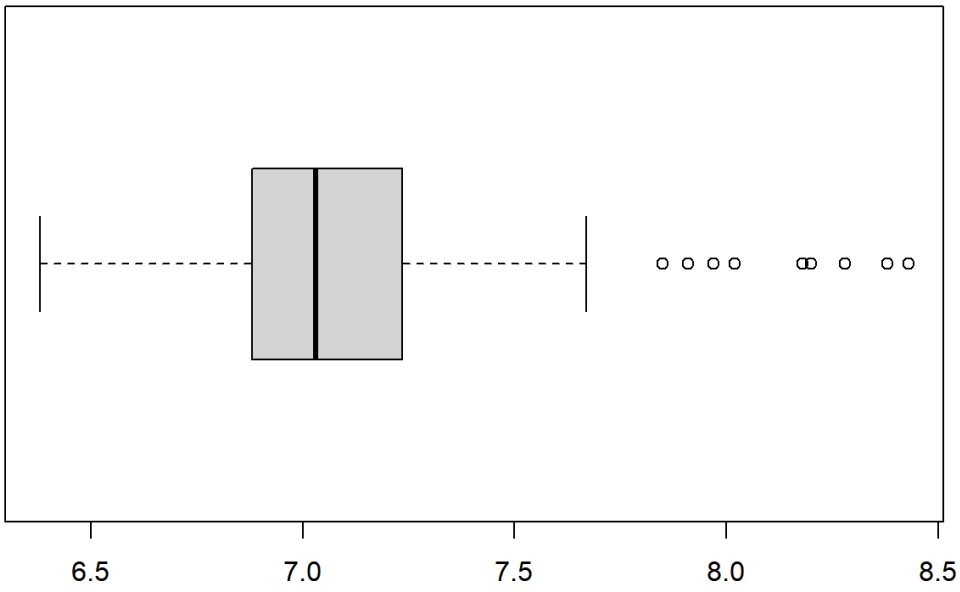
```
> t.test(vacation,
+        alternative = "less",
+        mu = 24,
+        conf.level = 0.95)

    One Sample t-test

data:  vacation
t = -2.2584, df = 34, p-value = 0.01522
alternative hypothesis: true mean is less than 24
95 percent confidence interval:
     -Inf 23.12409
sample estimates:
mean of x
 20.51429
```

The $p-value = 0.01522 < 0.05 = \alpha$, so we reject H_0.

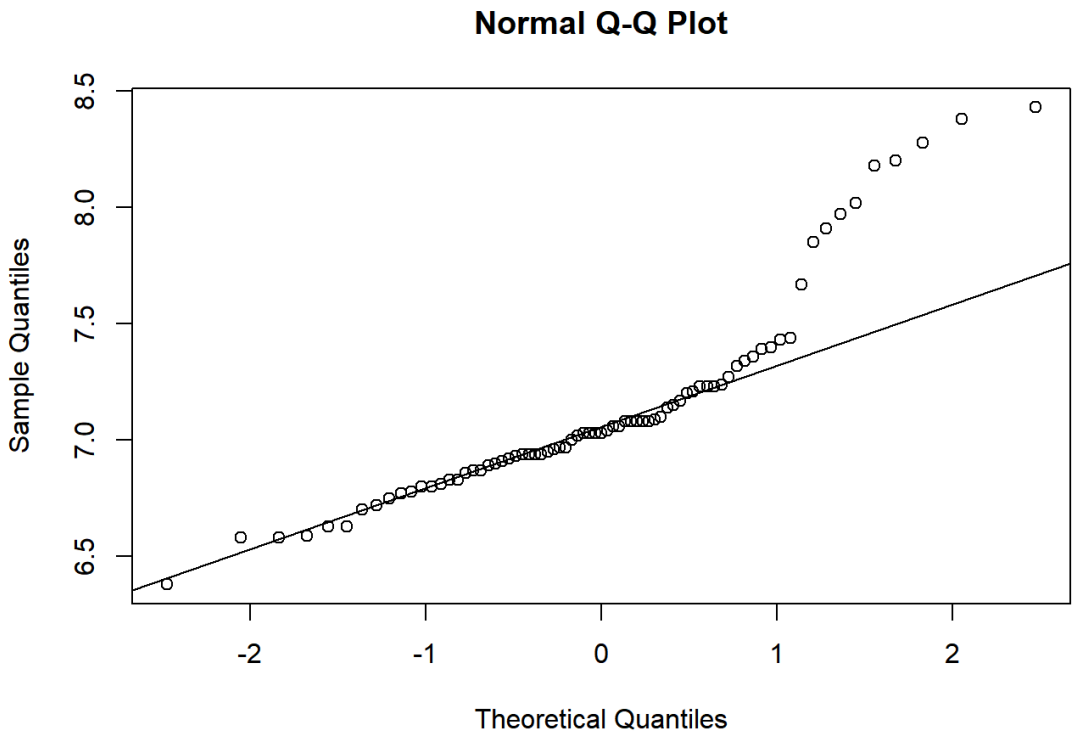
Problem 10.2
Repeat the above for the data set smokyph. This data set measures pH levels for water samples in the Great Smoky Mountains. Use the waterph column to test the null hypothesis that $\mu = 7$. What is a reasonable alternative?

```
> head(smokyph)
  waterph  elev code
1    7.91 0.244    0
2    7.14 0.375    0
3    6.81 0.567    0
4    6.97 0.512    0
5    7.21 0.408    0
6    6.94 0.512    0
```

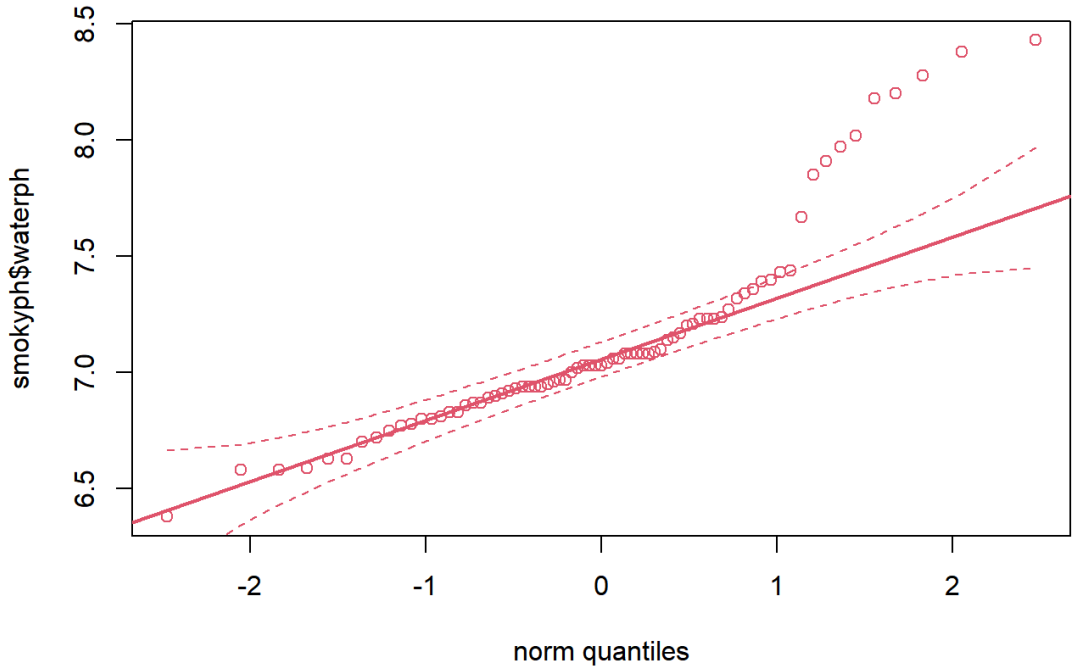
```
> boxplot(smokyph$waterph, horizontal = TRUE)
```



```
> qqnorm(smokyph$waterph)
> qqline(smokyph$waterph)
```




```
> qqplot.das(smokyph$waterph)
```



H_0 : X is normally distributed
H_A : X isn't normally distributed

```
> shapiro.test(smokyph$waterph)

    Shapiro-Wilk normality test

data:  smokyph$waterph
W = 0.86654, p-value = 1.178e-06
```

The $p-value = 1.178e-06 < 0.05 = \alpha$, so we reject H_0 the data is not normally distributed.

We can make hypothesis test only for the median. Let's see what is the median of the sample.

```
> median(smokyph$waterph)
[1] 7.03
```

So it is reasonable to test

$$H_0 : Me = 7$$
$$H_A : Me \neq 7$$

```
> wilcox.test(smokyph$waterph,
+       alternative = "two.sided",
+       mu = 7,
+       conf.level = 0.95)

    Wilcoxon signed rank test with continuity correction

data:  smokyph$waterph
V = 1711, p-value = 0.08177
alternative hypothesis: true location is not equal to 7
```

The $p-value = 0.08177 > 0.05 = \alpha$, so we have no reason to reject H_0
or

$$H_0 : Me = 7$$
$$H_A : Me > 7$$

```
> wilcox.test(smokyph$waterph,
+       alternative = "greater",
+       mu = 7,
+       conf.level = 0.95)

    Wilcoxon signed rank test with continuity correction

data:  smokyph$waterph
```

```
V = 1711, p-value = 0.04089
alternative hypothesis: true location is greater than 7
```

The $p-value = 0.04089 < 0.05 = \alpha$, so we reject H_0.

Problem 10.3

An exit poll by a news station of 900 people in the state of Florida found 440 voting for Bush and 460 voting for Gore. Does the data support the hypothesis that Bush received $p = 50\,\%$ of the state's vote?

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

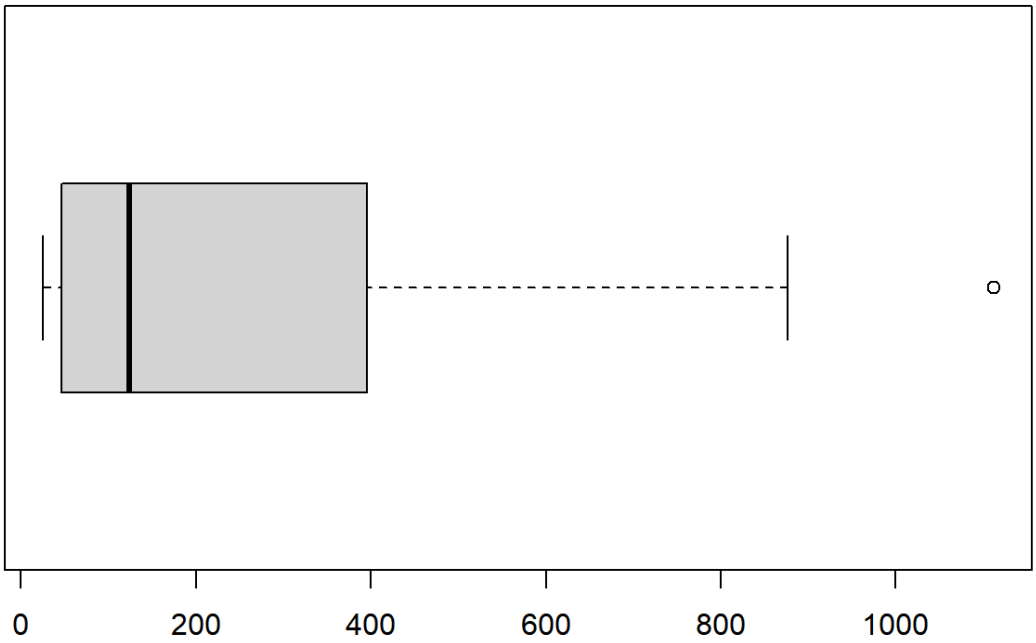
```
> prop.test(440, 900, p = 0.5, conf.level = 0.95)

    1-sample proportions test with continuity correction

data:  440 out of 900, null probability 0.5
X-squared = 0.40111, df = 1, p-value = 0.5265
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4557952 0.5220786
sample estimates:
       p
0.4888889
```

The $p-value = 0.5265 > 0.05 = \alpha$, so we have no evidence to reject H_0. We expect Bush to receive 0.5 from the voting.

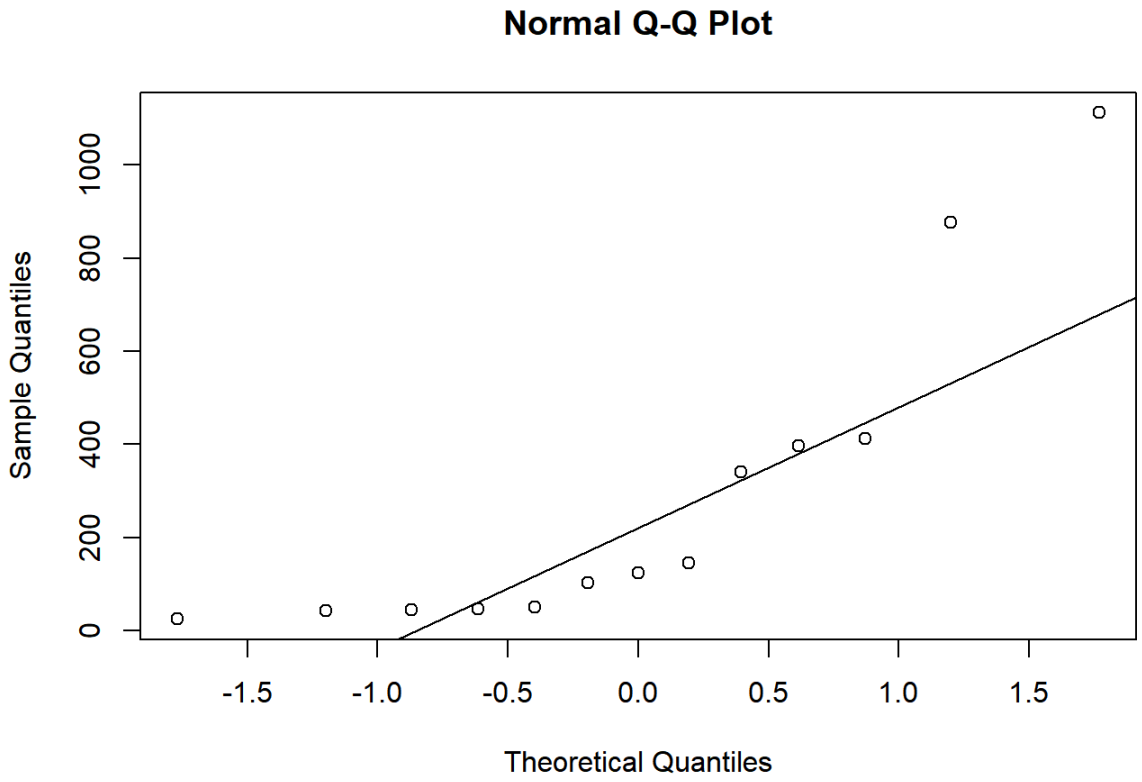
Problem 10.4

Load the data set cancer. Look only at cancer[['stomach']]. These are survival times for stomach cancer patients taking a large dosage of Vitamin C. Test the null hypothesis that the Median is 100 days. Should you also use a t-test? Why or why not? (A boxplot of the cancer data is interesting.)

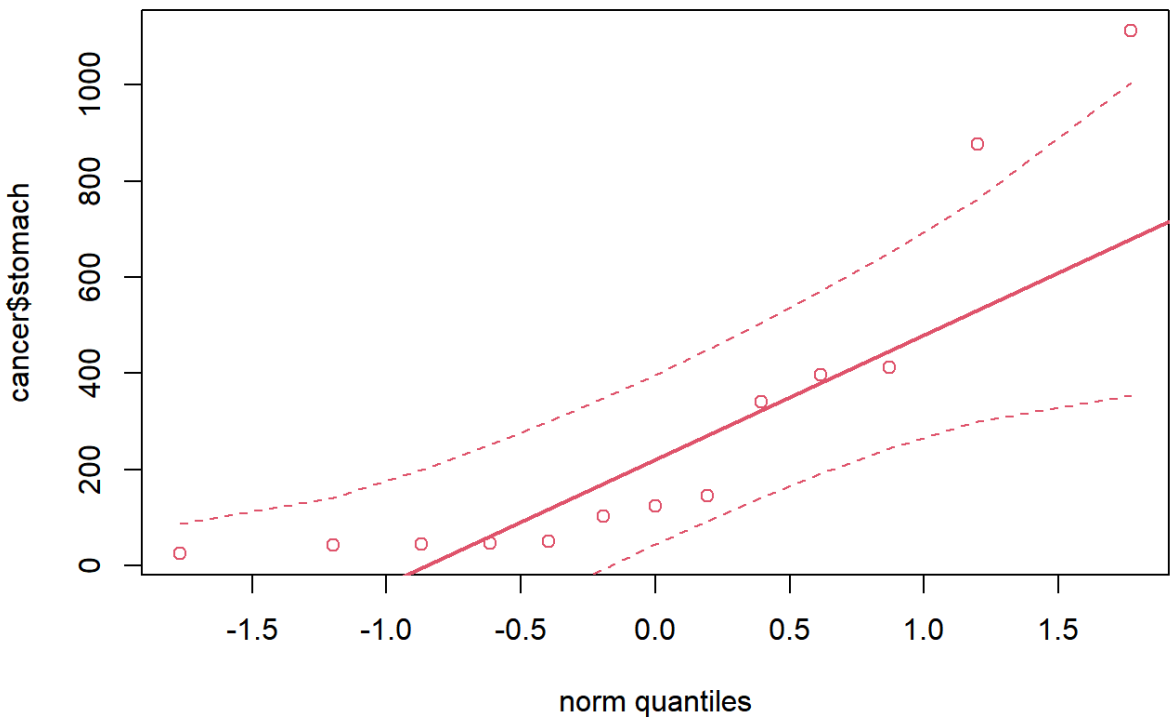
```
> str(cancer)
List of 5
 $ stomach : num [1:13] 124 42 25 45 412 ...
 $ bronchus: num [1:17] 81 461 20 450 246 166 63 64 155 859 ...
 $ colon   : num [1:17] 248 377 189 1843 180 ...
 $ ovary   : num [1:6] 1234 89 201 356 2970 ...
 $ breast  : num [1:11] 1235 24 1581 1166 40 ...
> boxplot(cancer$stomach, horizontal = TRUE)
```




```
> qqnorm(cancer$stomach)
> qqline(cancer$stomach)
```



```
> qqplot.das(cancer$stomach)
```




```
> shapiro.test(cancer$stomach)

    Shapiro-Wilk normality test

data:  cancer$stomach
W = 0.75473, p-value = 0.002075
```

The $p-value = 0.002075 < 0.05 = \alpha$, so we reject H_0. The data is not normally distributed. We can't use a t-test. We can only make a hypothsis test for the median.

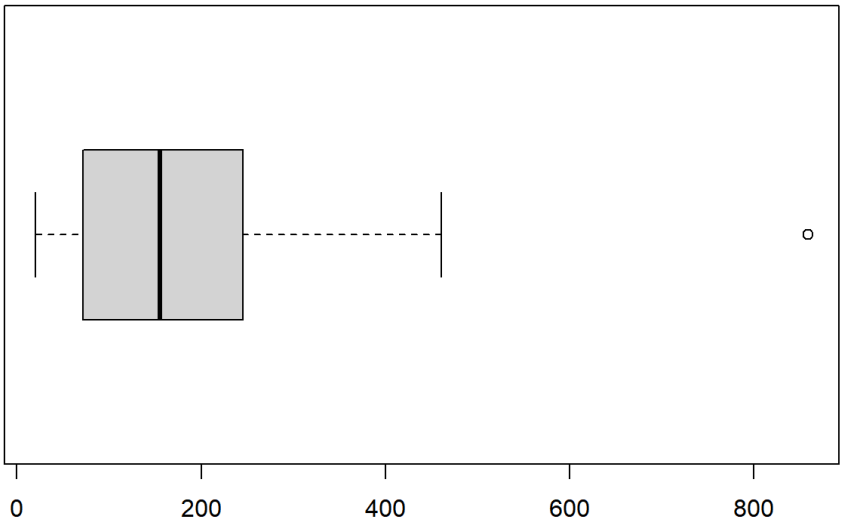
$$H_0 : Me = 100\ days$$
$$H_A : Me \neq 100\ days$$

```
> wilcox.test(cancer$stomach, mu = 100, alternative = "two.sided")

    Wilcoxon signed rank exact test

data:  cancer$stomach
V = 61, p-value = 0.3054
alternative hypothesis: true location is not equal to 100
```

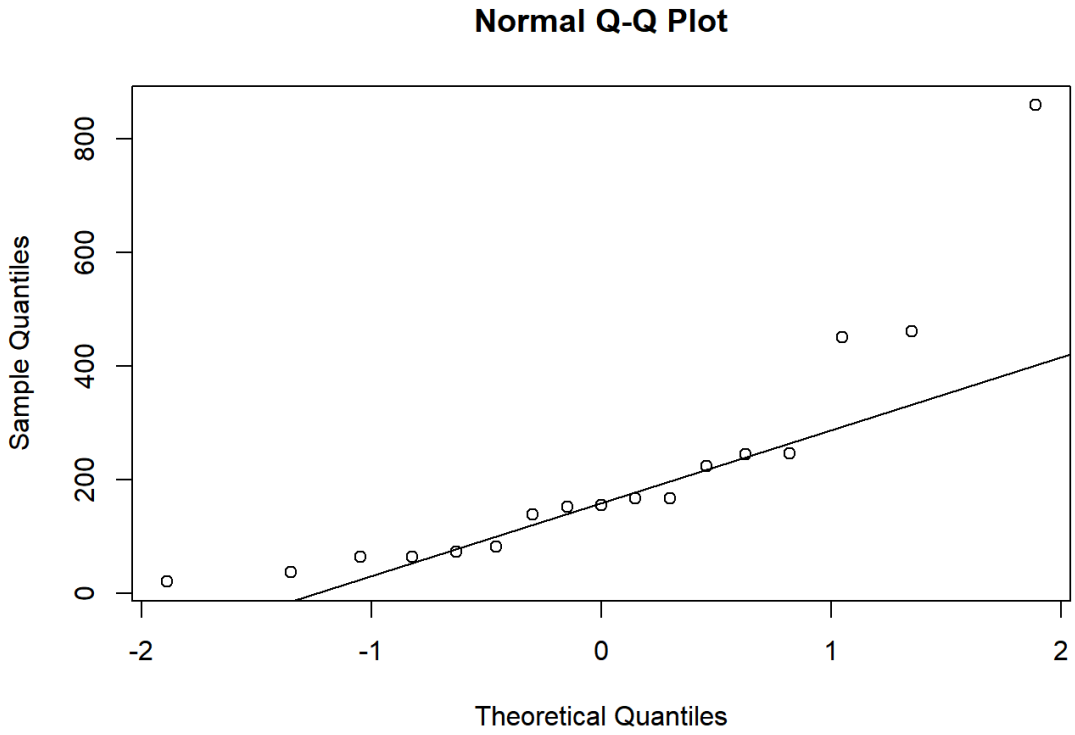
The $p-value = 0.3054 > 0.05 = \alpha$, so we have no evidence to reject H_0. We can assume that people having stomach cancer in average survive 100 days.

Let's review the data for the brochus cancer. Test the hypothesis that the Median is 100 days. Is the t-test appropriate here?

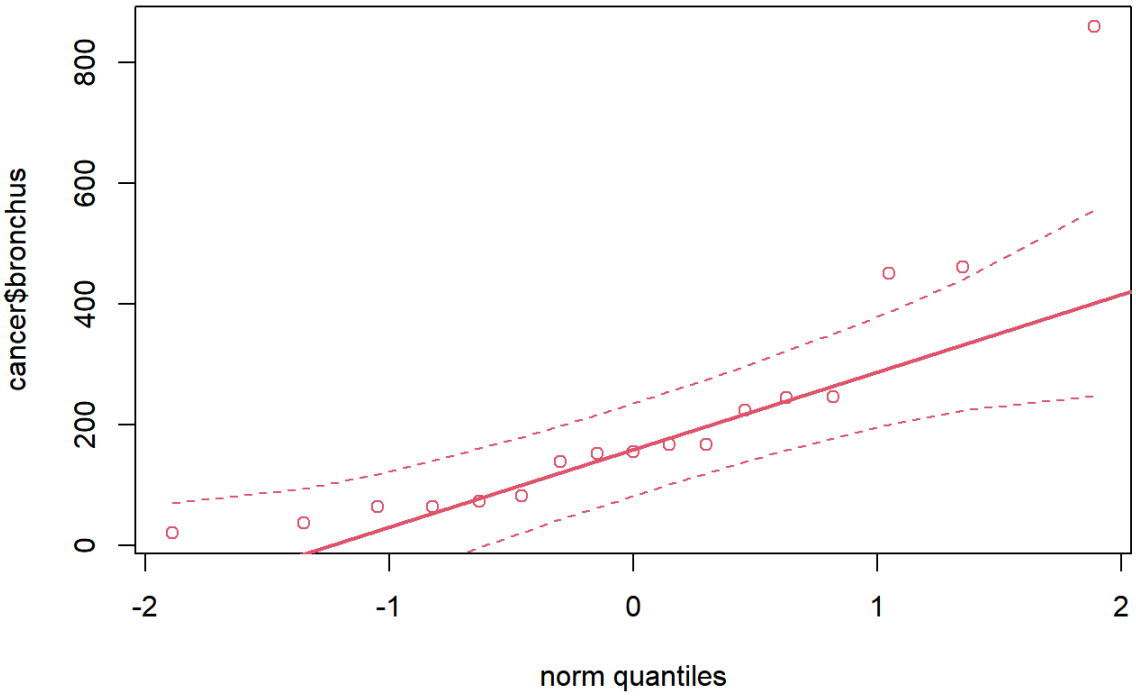
```
> boxplot(cancer$bronchus, horizontal = TRUE)
```




```
> qqnorm(cancer$bronchus)
> qqline(cancer$bronchus)
```



```
> qqplot.das(cancer$bronchus)
```




```
> shapiro.test(cancer$bronchus)

    Shapiro-Wilk normality test

data:  cancer$bronchus
W = 0.76596, p-value = 0.0007186
```

The $p-value = 0.0007186 < 0.05 = \alpha$, so we reject H_0. The data is not normally distributed. We can't use a t-test. We can only make a hypothesis test for the median.

$$H_0 : Me = 100\ days$$
$$H_A : Me \neq 100\ days$$

```
> wilcox.test(cancer$bronchus, mu = 100, alternative = "two.sided", conf.level = 0.95)
Warning in wilcox.test.default(cancer$bronchus, mu = 100, alternative =
"two.sided", : cannot compute exact p-value with ties

    Wilcoxon signed rank test with continuity correction

data:  cancer$bronchus
V = 124, p-value = 0.02607
alternative hypothesis: true location is not equal to 100
```

The $p-value = 0.02607 < 0.05 = \alpha$, so we reject H_0. The average time of survival from bronchus cancer is different than 100 days.