Verzani Problem Set

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

Problem 9.1

Create 15 random numbers that are normally distributed with mean 10 and standard deviation 5.

```
> x <- rnorm(15, mean = 10, sd = 5)
```

Find a 1-sample z-test at the 95% level. Did it get it right?

```
> mean(x)
[1] 10.44665
> 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
> simple.z.test(x, sigma = 5, conf.level = 0.95)
[1]  7.916348 12.976953
```

Yes

What would happen if we reduce the confidence level to 0.90

```
> simple.z.test(x, sigma = 5, conf.level = 0.90)
```

```
[1]  8.323154 12.570148
```

What would happen if we enlarge the confidence level to 0.99

```
> simple.z.test(x, sigma = 5, conf.level = 0.99)
[1]  7.12127 13.77203
```

Problem 9.2

Do the above 100 times. Compute what percentage of the means of the samples is in a 95% confidence interval?

```
> f <- function () {
+   mean(rnorm(15, mean = 10, sd = 5))
+ }
> library(UsingR)
> xbar <- simple.sim(100, f)
> SE <- 5/sqrt(15)
> alpha <- 0.05
> zstar <- qnorm(1 - alpha/2)
> sum(abs(xbar - 10) < zstar*SE)
[1] 97
> persentage <- sum(abs(xbar - 10) < zstar*SE) / 100;
persentage
[1] 0.97
```

Problem 9.3

The t-test is just as easy to do. Do a t-test on the same data. Is it correct now? Comment on the relationship between the confidence intervals.

```
> mean(x)
[1] 10.44665
> t.test(x, conf.level = 0.95)

    One Sample t-test

data:  x
t = 13.805, df = 14, p-value = 1.518e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
  8.823651 12.069650
sample estimates:
mean of x
 10.44665
> left <- 0; right <- 0
> alpha <- 0.05
> zstar <- qnorm(1 - alpha/2)
> for(i in 1:100) {
+   x <- rnorm(15, mean = 10, sd = 5)
+   SE <- sd(x) / sqrt(15)
+   left[i] <- mean(x) - zstar * SE
+   right[i] <- mean(x) + zstar * SE
+
+ }
> sum(left < 10 & right > 10)
[1] 89
> sum(left < 10 & right > 10) / 100
[1] 0.89
```

Problem 9.4

Find an $80\,\%$ and $95\,\%$ confidence interval for the median for the `exec.pay` dataset.

```
> median(exec.pay)
[1] 27
```

$80\,\%$ confidence interval for the median

```
> wilcox.test(exec.pay, conf.level = 0.80, conf.int=TRUE)

    Wilcoxon signed rank test with continuity correction

data:  exec.pay
V = 19306, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
80 percent confidence interval:
 27.00005 31.49996
sample estimates:
(pseudo)median
      29.00002
```

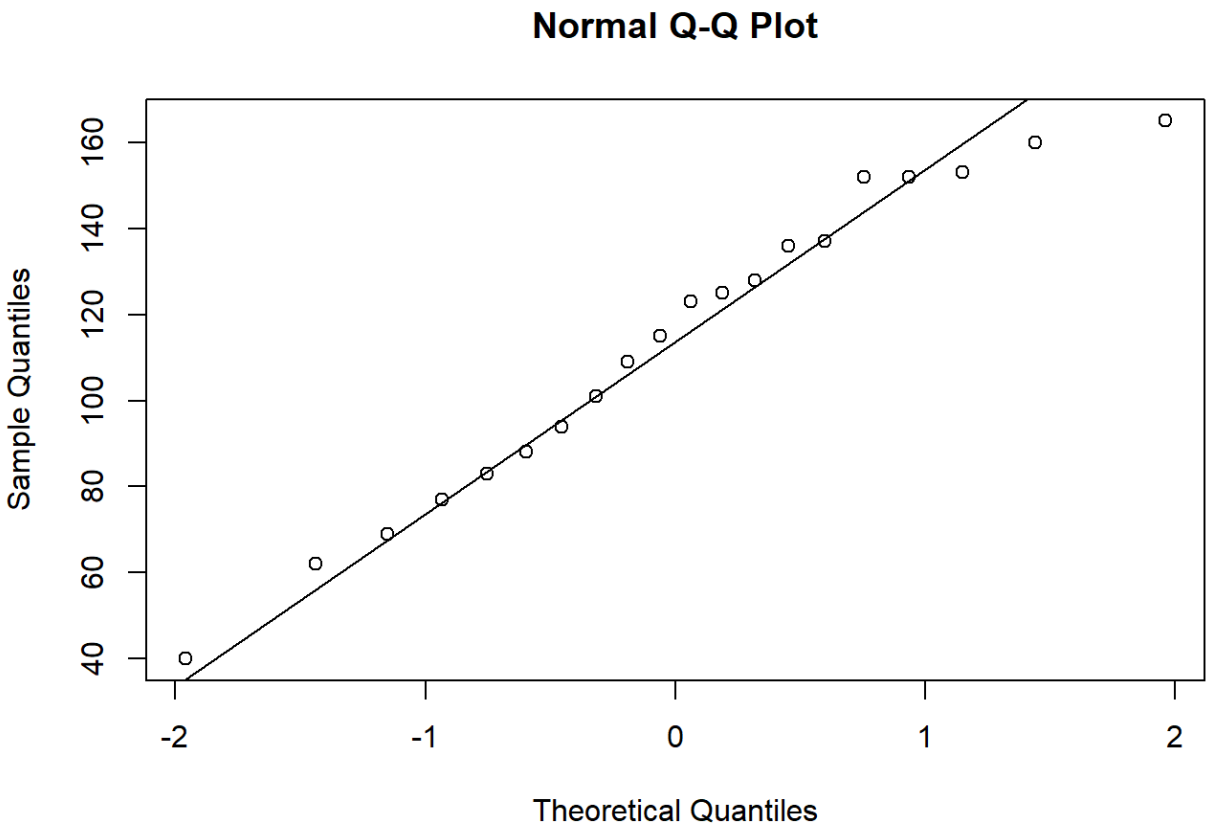
80 % confidence interval for the median

```
> wilcox.test(exec.pay, conf.level = 0.95, conf.int=TRUE)

    Wilcoxon signed rank test with continuity correction

data:  exec.pay
V = 19306, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 25.99998 32.99994
sample estimates:
(pseudo)median
      29.00002
```

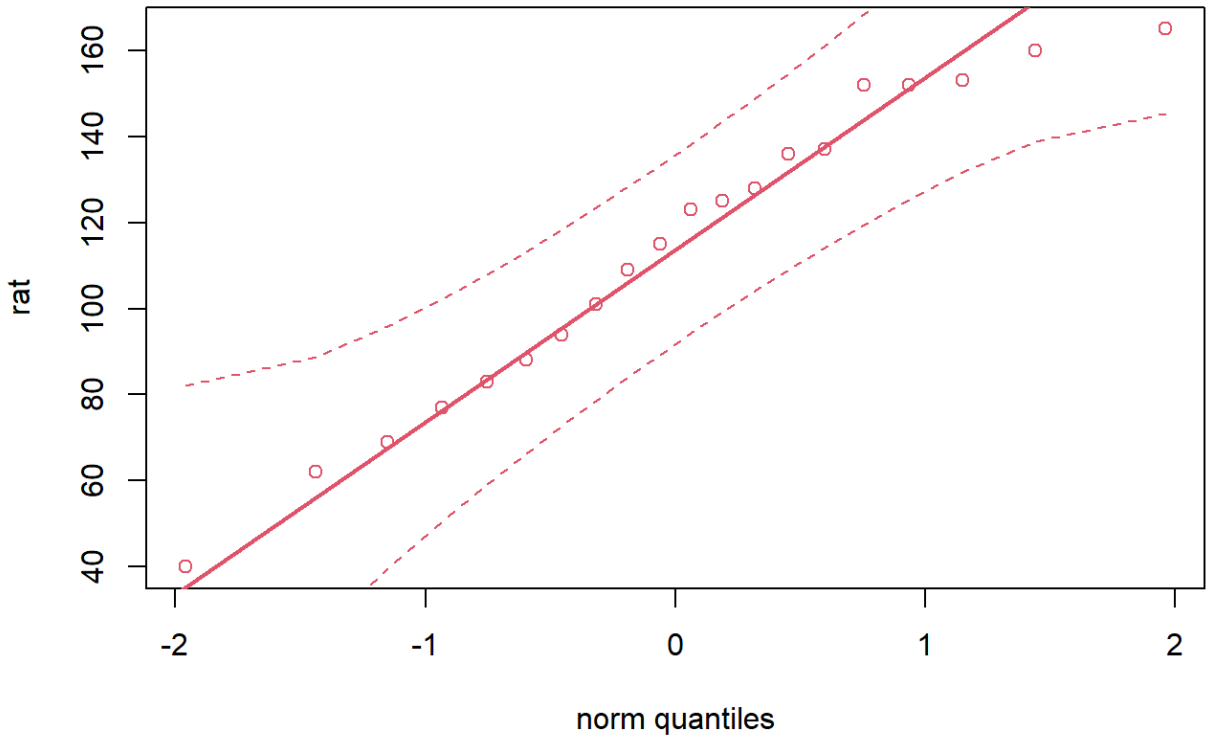
Problem 9.5

The data set `rat` records survival times for rats. Do a t-test for mean if the data suggests it is appropriate. If not, say why not.

```
> qqnorm(rat)
> qqline(rat)
```




```
> 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(rat, "norm")
```




```
> shapiro.test(rat)

    Shapiro-Wilk normality test

data:  rat
W = 0.96134, p-value = 0.571
```

As we see from the graphics and $p-value = 0.571 > 0.05 = \alpha$ we can assume that the data is normally distributed.

The data sample is very small.

```
> length(rat)
[1] 20
```

It has only 20 observations and σ is unknown, therefore, we use the `t.test`

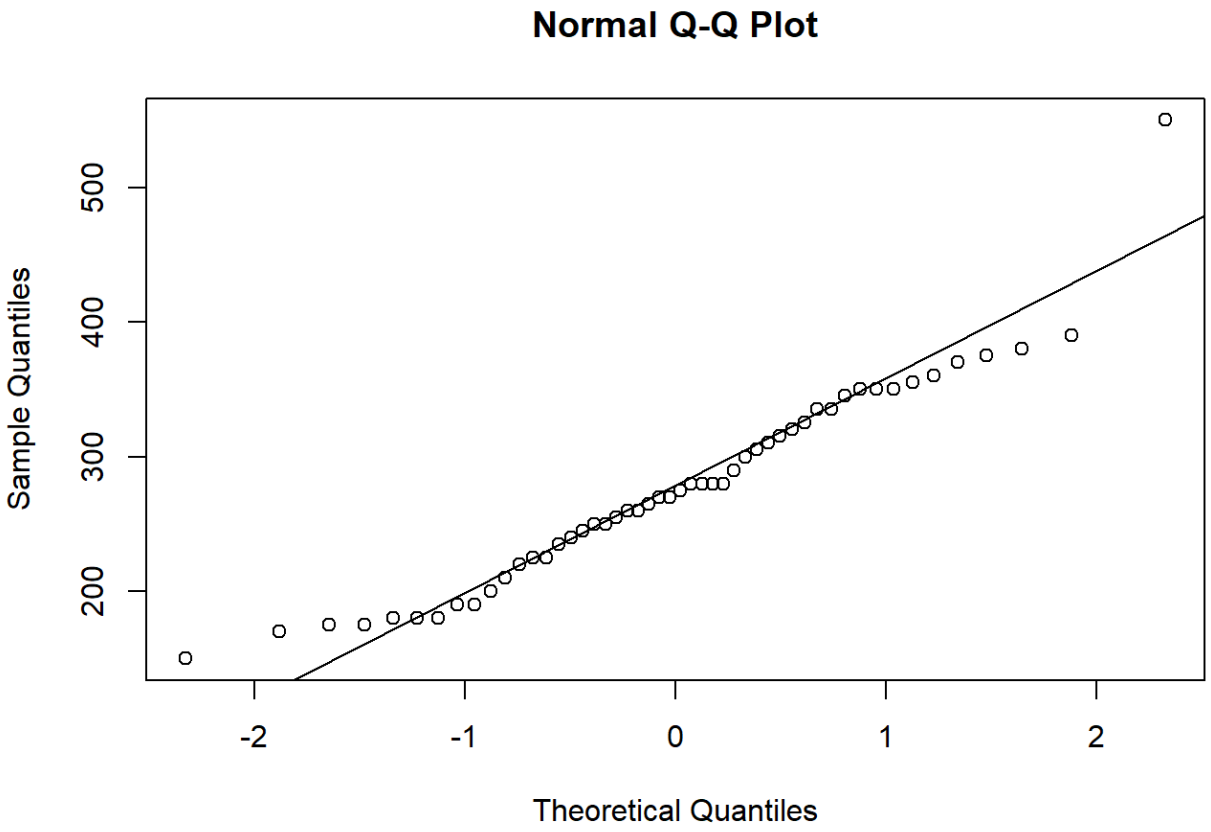
```
> mean(rat)
[1] 113.45
> t.test(rat)
```

```
    One Sample t-test

data:  rat
t = 14.176, df = 19, p-value = 1.48e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  96.69997 130.20003
sample estimates:
mean of x
   113.45
```

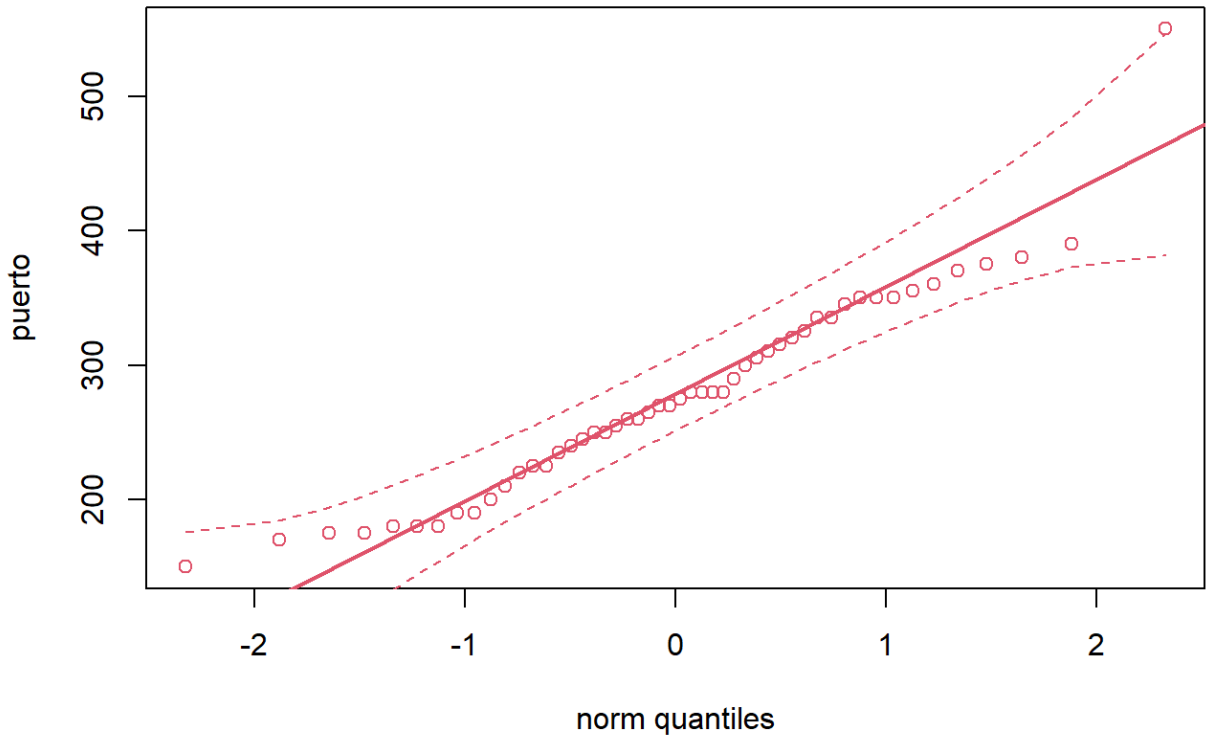
Problem 9.6

Repeat the previous for the data set `puerto` containing weekly incomes of Puerto Ricans in Miami.

```
> qqnorm(puerto)
> qqline(puerto)
```




```
> qqplot.das(puerto, "norm")
```




```
> shapiro.test(puerto)

    Shapiro-Wilk normality test

data:  puerto
W = 0.94538, p-value = 0.02212
```

As we see from the graphics and $p-value = 0.021 < 0.05 = \alpha$, so we can't assume that the data is normally distributed.

The data sample is not very small.

```
> length(puerto)
[1] 50
```

It has 50 observations and $\sigma$is unknown

If the $\sigma < \infty$ we can use the CLT and obtain

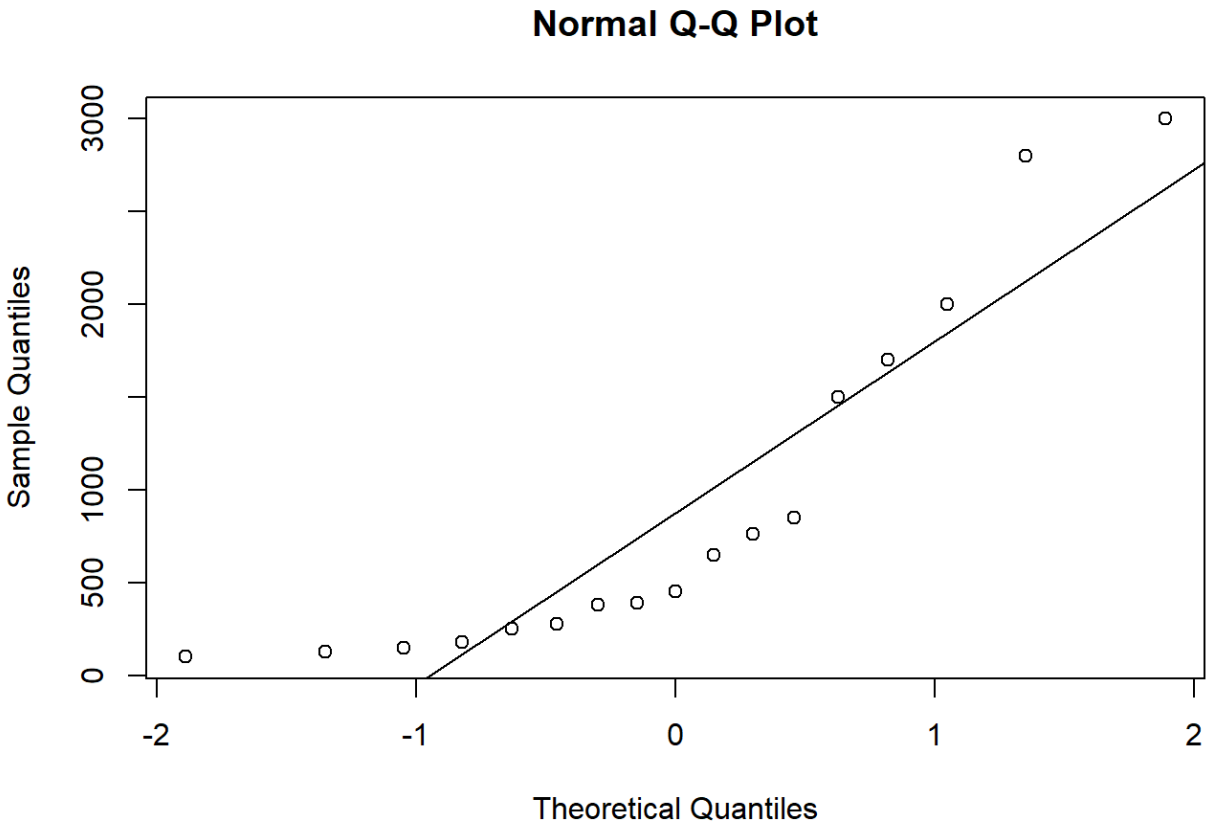
```
> mean(puerto)
[1] 277.5
> t.test(puerto)
```

```
    One Sample t-test

data:  puerto
t = 25.847, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 255.9244 299.0756
sample estimates:
mean of x
    277.5
```

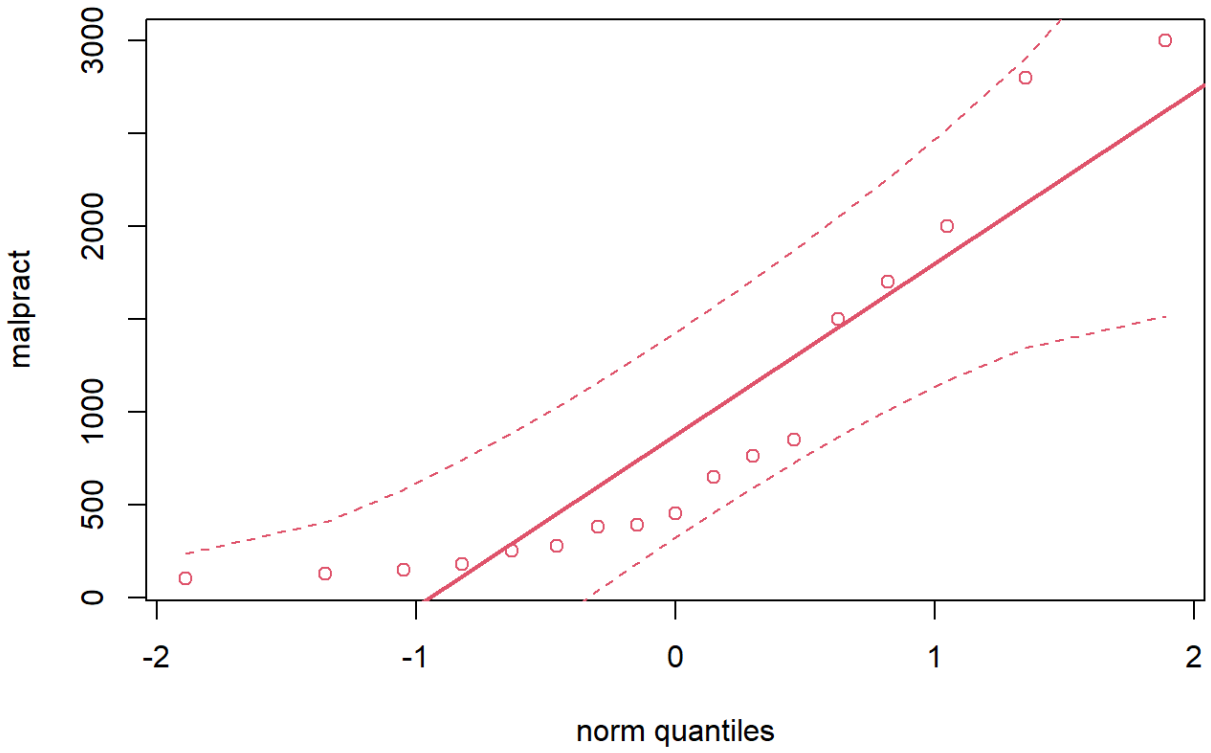
Problem 9.7

The median may be the appropriate measure of center. If so, you might want to have a confidence interval for it too. Find a $90\,\%$ confidence interval for the median for the data set `malpract`, containing the size of malpractice awards. Comment why this distribution doesn't lend itself to the z-test or t-test.

```
> qqnorm(malpract)
> qqline(malpract)
```



```
> qqplot.das(malpract, "norm")
```




```
> shapiro.test(malpract)

    Shapiro-Wilk normality test

data:  malpract
W = 0.80547, p-value = 0.002414
```

As we see from the $p-value = 0.002 < 0.05 = \alpha$, so we can't assume that the data is normally distributed.

The data sample is small.

```
> length(malpract)
[1] 17
```

It has 17 observations and we performe Wilcoxon confidence interval for the median.

```
> median(malpract)
[1] 450
> wilcox.test(malpract, conf.level = 0.90, conf.int =
TRUE)
```

```
    Wilcoxon signed rank exact test

data:  malpract
V = 153, p-value = 1.526e-05
alternative hypothesis: true location is not equal to 0
90 percent confidence interval:
  385 1325
sample estimates:
(pseudo)median
           800
```

If the sample size was large we could use the normal approximation, but here we have only 17 observations.