Confidence Interval Estimation

2020

A bootstrap sample

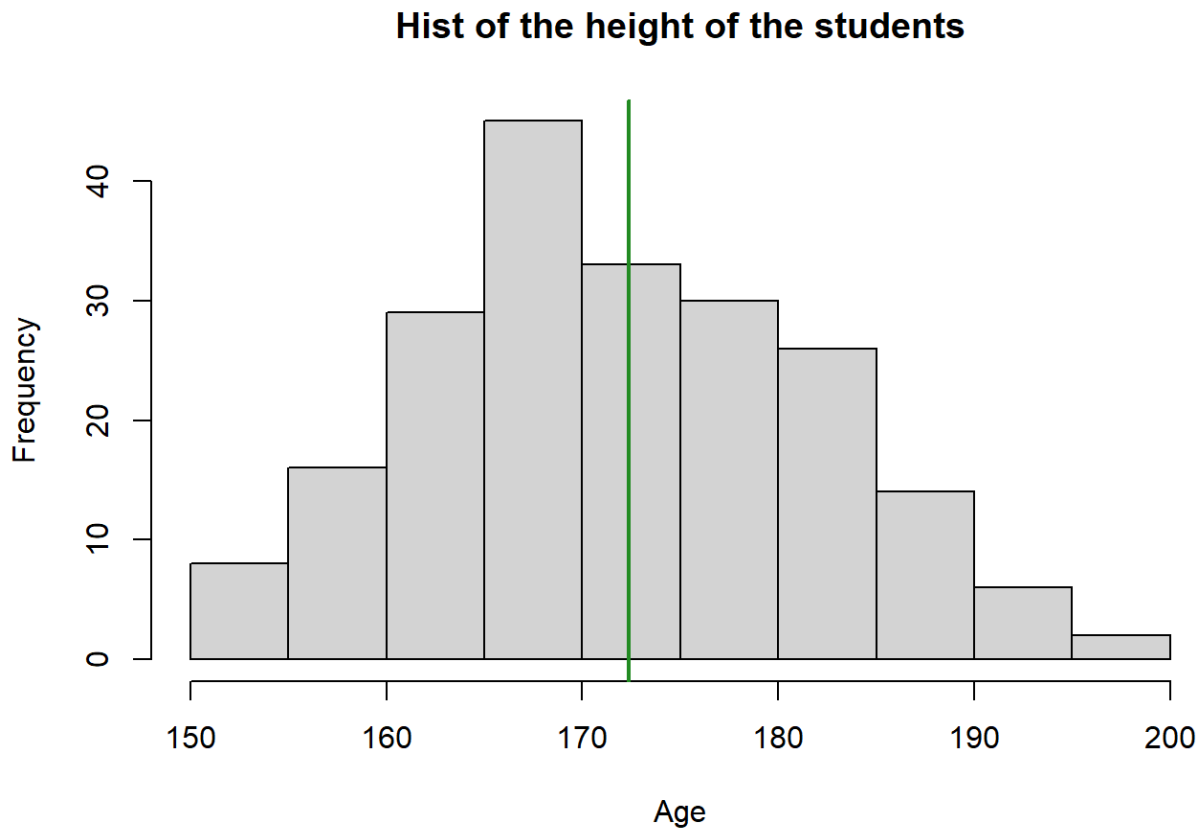
Bootstrapping is a method of sampling **with replacement** from a data set to make statistical inference.

Example:

Let's take for example the height of the students from the `survey` data set.

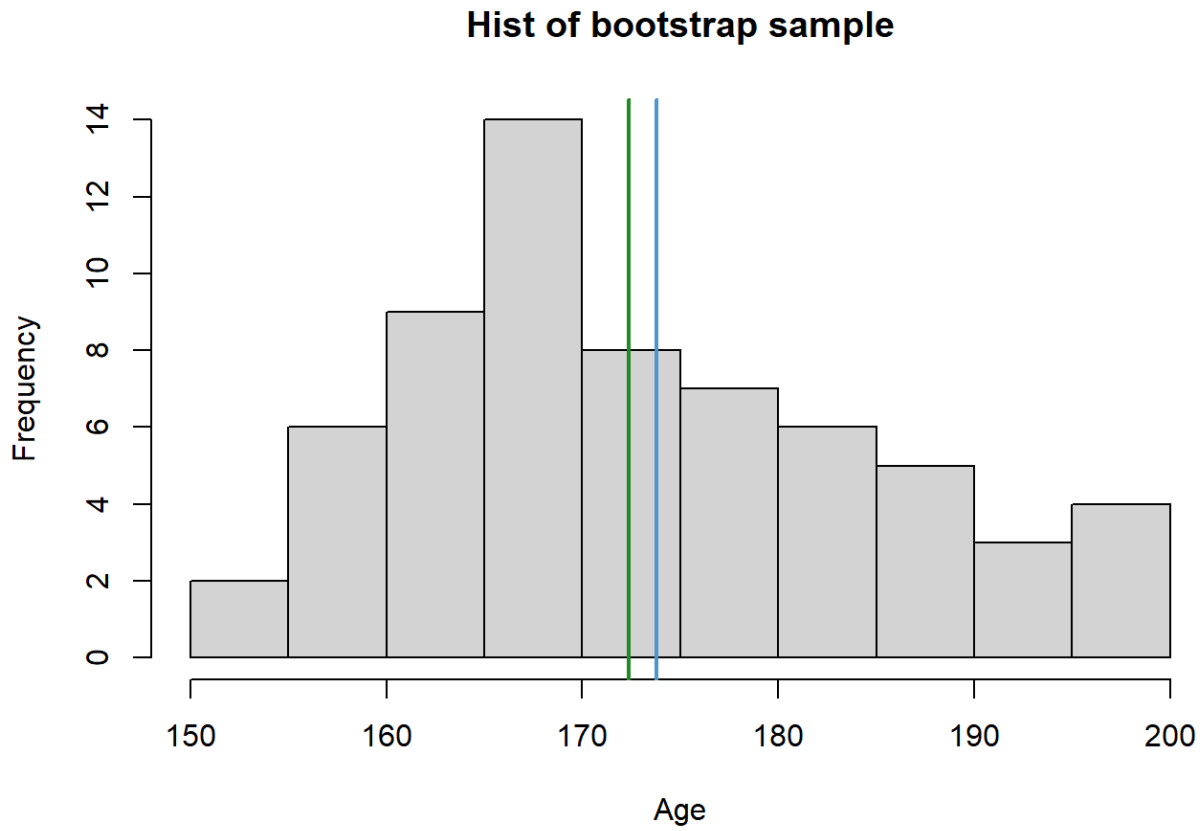
```
> library(MASS)
> str(survey)
'data.frame':   237 obs. of  12 variables:
 $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2
1 2 1 2 2 ...
 $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20
18.5 ...
 $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5
18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2
2 2 2 2 ...
 $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3
1 3 2 1 1 3 3 3 ...
 $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2
2 3 3 3 3 3 3 ...
 $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3
3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2
2 2 2 2 2 2 ...
 $ Height: num  173 178 NA 160 165 ...
 $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA
2 2 1 1 2 2 2 ...
 $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...
```

```
> summary(survey$Height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  150.0   165.0   171.0   172.4   180.0   200.0      28
> hist(survey$Height,
+      main = "Hist of the height of the students",
+      xlab = "Age",
+      breaks = 10)
> abline(v = mean(survey$Height, na.rm = TRUE),
+        lwd = 2,
+        col = "#228B22")
```

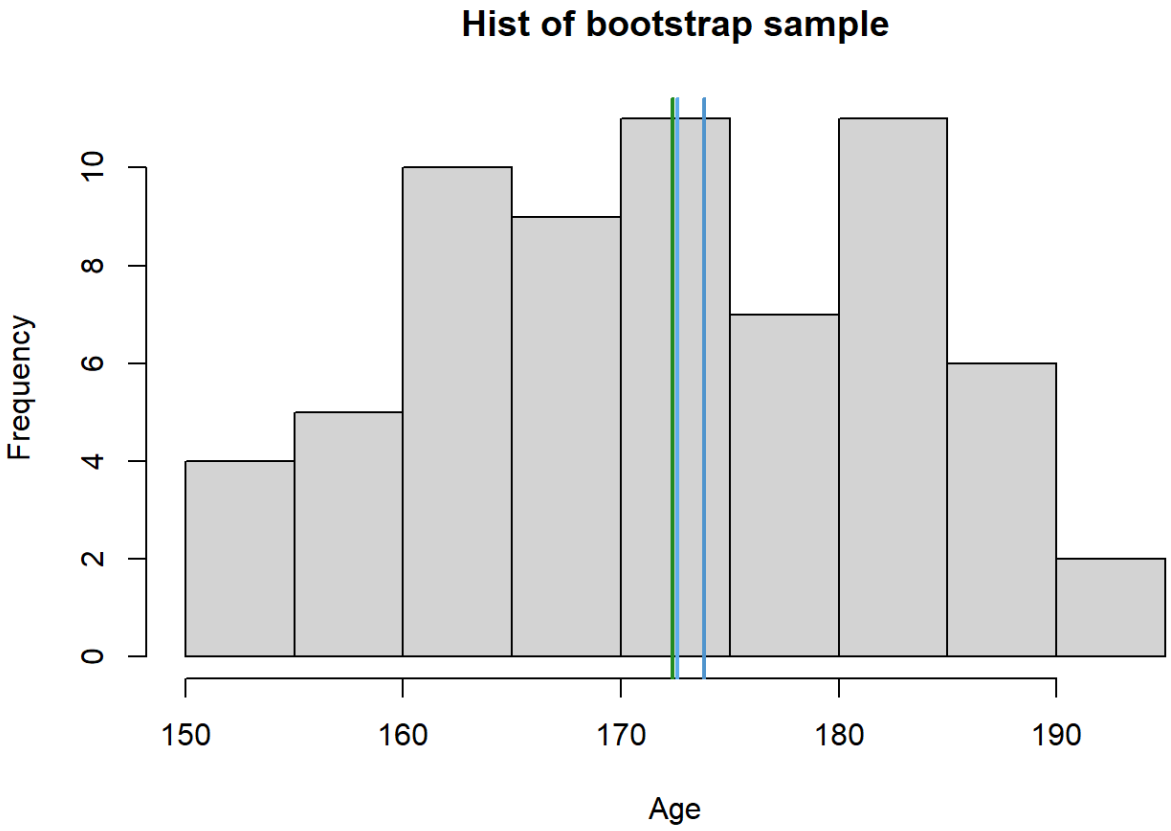



```
> samp1 <- sample(survey$Height, 70, replace = TRUE)
> summary(samp1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  152.4   165.0   170.2   173.8   180.9   200.0       6
> hist(samp1,
+      main = "Hist of bootstrap sample",
+      xlab = "Age",
+      breaks = 10)
> abline(v = mean(survey$Height, na.rm = TRUE),
+        lwd = 2,
+        col = "#228B22")
```

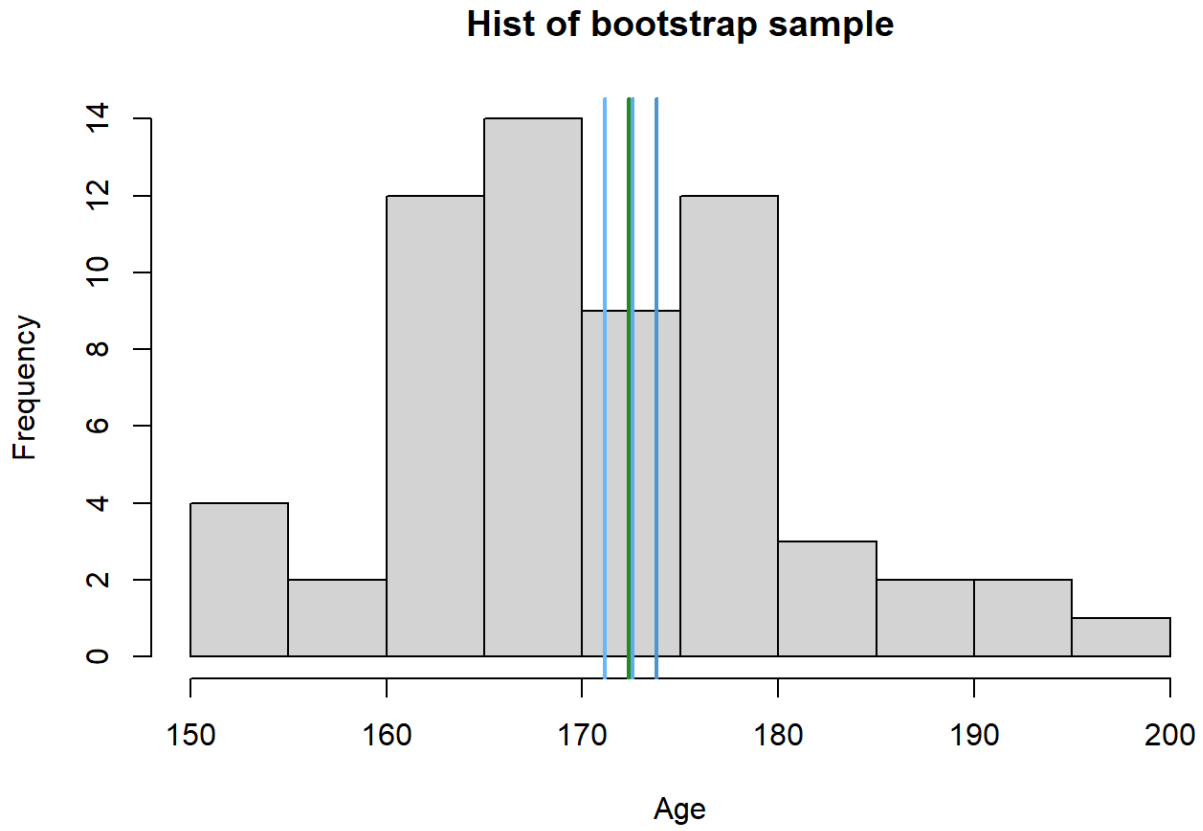
```
> abline(v = mean(samp1, na.rm = TRUE),
+        lwd = 2,
+        col = "#4F94CD")
```




```
> samp2 <- sample(survey$Height, 70, replace = TRUE)
> summary(samp2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  152.4   165.0   172.0   172.6   180.3   195.0       5
> hist(samp2,
+      main = "Hist of bootstrap sample",
+      xlab = "Age",
+      breaks = 10)
> abline(v = mean(survey$Height, na.rm = TRUE),
+        lwd = 2,
+        col = "#228B22")
> abline(v = mean(samp1, na.rm = TRUE),
+        lwd = 2,
+        col = "#4F94CD")
> abline(v = mean(samp2, na.rm = TRUE),
+        lwd = 2,
+        col = "#5CACEE")
```

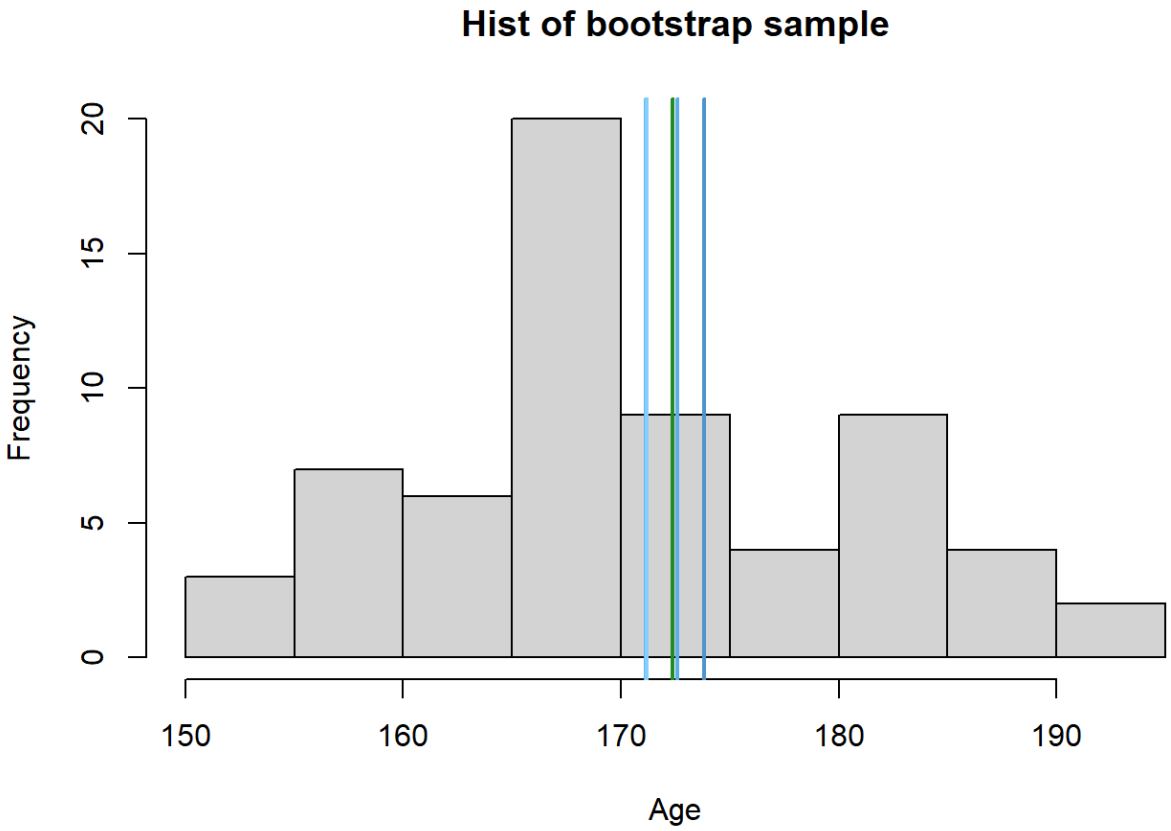



```
> samp3 <- sample(survey$Height, 70, replace = TRUE)
> summary(samp3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  152.4   165.0   170.0   171.2   177.0   196.0       9
> hist(samp3,
+      main = "Hist of bootstrap sample",
+      xlab = "Age",
+      breaks = 10)
> abline(v = mean(survey$Height, na.rm = TRUE),
+        lwd = 2,
+        col = "#228B22")
> abline(v = mean(samp1, na.rm = TRUE),
+        lwd = 2,
+        col = "#4F94CD")
> abline(v = mean(samp2, na.rm = TRUE),
+        lwd = 2,
+        col = "#5CACEE")
> abline(v = mean(samp3, na.rm = TRUE),
+        lwd = 2,
+        col = "#63B8FF")
```




```
> samp4 <- sample(survey$Height, 70, replace = TRUE)
> summary(samp4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  153.5   165.1   170.0   171.2   179.3   190.5       6
> hist(samp4,
+      main = "Hist of bootstrap sample",
+      xlab = "Age",
+      breaks = 10)
> abline(v = mean(survey$Height, na.rm = TRUE),
+        lwd = 2,
+        col = "#228B22")
> abline(v = mean(samp1, na.rm = TRUE),
+        lwd = 2,
+        col = "#4F94CD")
> abline(v = mean(samp2, na.rm = TRUE),
+        lwd = 2,
+        col = "#5CACEE")
> abline(v = mean(samp3, na.rm = TRUE),
+        lwd = 2,
+        col = "#63B8FF")
> abline(v = mean(samp4, na.rm = TRUE),
```

```
+         lwd = 2,
+         col = "#87CEFA")
```



Confidence Interval Estimation /Оценка на доверителни интервали/

Is used for estimating unknown parameters for a known or unknown distribution.

Confidence interval for the mean

If we have a sample of independent observations of a random variable X. We don't know $\mu = \mathbb{E}X$ and we have estimated it by the average of the observations $\overline{X}$. Let's choose $\alpha \in (0,1)$. What is the smallest interval that we can construct such that will cover the unknown

parameter μ with probability at least $(1-\alpha)$? $(1-\alpha)$ is called **confidence level /ниво на доверие/**.

z-test /known standard deviation/

- If $X \in N(\mu, \sigma^2)$ and σ is **known**, then

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1), \text{ and}$$

$$\mathbb{P}\left(z_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}\left(-z^* < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z^* \right) = 1-\alpha$$

$$\mathbb{P}\left(-z^*\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z^*\frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

$$\mathbb{P}\left(\overline{X}-z^*\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z^*\frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

$1-\alpha$ confidence interval for μ is

$$\left(\overline{X}-z^*\frac{\sigma}{\sqrt{n}}, \overline{X}+z^*\frac{\sigma}{\sqrt{n}} \right)$$

When we build up confidence intervals we have **stochastic errors**. These are the errors which are due to the random nature of the sample.

$SE := \frac{\sigma}{\sqrt{n}}$ is called **standard error**.

z^*SE is called **maximum stochastic error** of the mean.

$$(\overline{X} - z^*SE, \overline{X} + z^*SE)$$

- $\sigma < \infty$ is **known**, and n is **large** enough to apply the **CLT**, then

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} Z, Z \in N(0,1)$$

We apply the same computations for large enough n and obtain the same confidence interval.

$$\left(\overline{X} - z^*\frac{\sigma}{\sqrt{n}}, \overline{X} + z^*\frac{\sigma}{\sqrt{n}}\right)$$

Example:

A person weighs himself on a regular basis and finds his weight to be

```
> weight <- c(75, 76, 73, 75, 74, 73, 73, 76, 73, 79)
> sigma <- 1.5
```

Error in weighing is normally distributed

$$X_i = \mu + \varepsilon_i, \varepsilon \in N(0, 1.5^2)$$

We estimate the confidence interval as

```
> z.test <- function(weight, sigma, conf.level = 0.95) {
+   n <- length(weight)
+   xbar <- mean(weight)
+   alpha <- 1 - conf.level
+   zstar <- qnorm(1 - alpha / 2)
+   SE <- sigma / sqrt(n)
+   xbar + c(-zstar*SE, zstar*SE)
+ }
> z.test(weight, 1.5)
[1] 73.77031 75.62969
```

95% confidence interval that the mean of the wight is between $(73.7703, 75.6297)$

```
> mean(weight)
[1] 74.7
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
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(weight, 1.5)
[1] 73.77031 75.62969
```

Example:

Let's continue the example for the height of the students. If the standard deviation of the population height of the students is 10.

```
> height <- survey$Height[!is.na(survey$Height)]
> simple.z.test(height, 10)
[1] 171.0251 173.7366
```

t-test /unknown standard deviation/

Usually we do not know the population standard deviation σ of the observed random variable and we estimate it from the sample. The sample standard deviation is denoted by s.

- If $X \in N(\mu, \sigma^2)$and σ is **unknown**, then

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in t(n-1)$$

Analogously to the previous computations we obtain that the $1 - \alpha$ confidence interval for μ is

$$\left(\overline{X} - t^* \frac{s}{\sqrt{n}}, \overline{X} + t^* \frac{s}{\sqrt{n}}\right)$$

where t^* is $1 - \frac{\alpha}{2}$ quantile of $t(n-1)$ distribution.

In this case the **standard error** is $SE := \frac{s}{\sqrt{n}}$

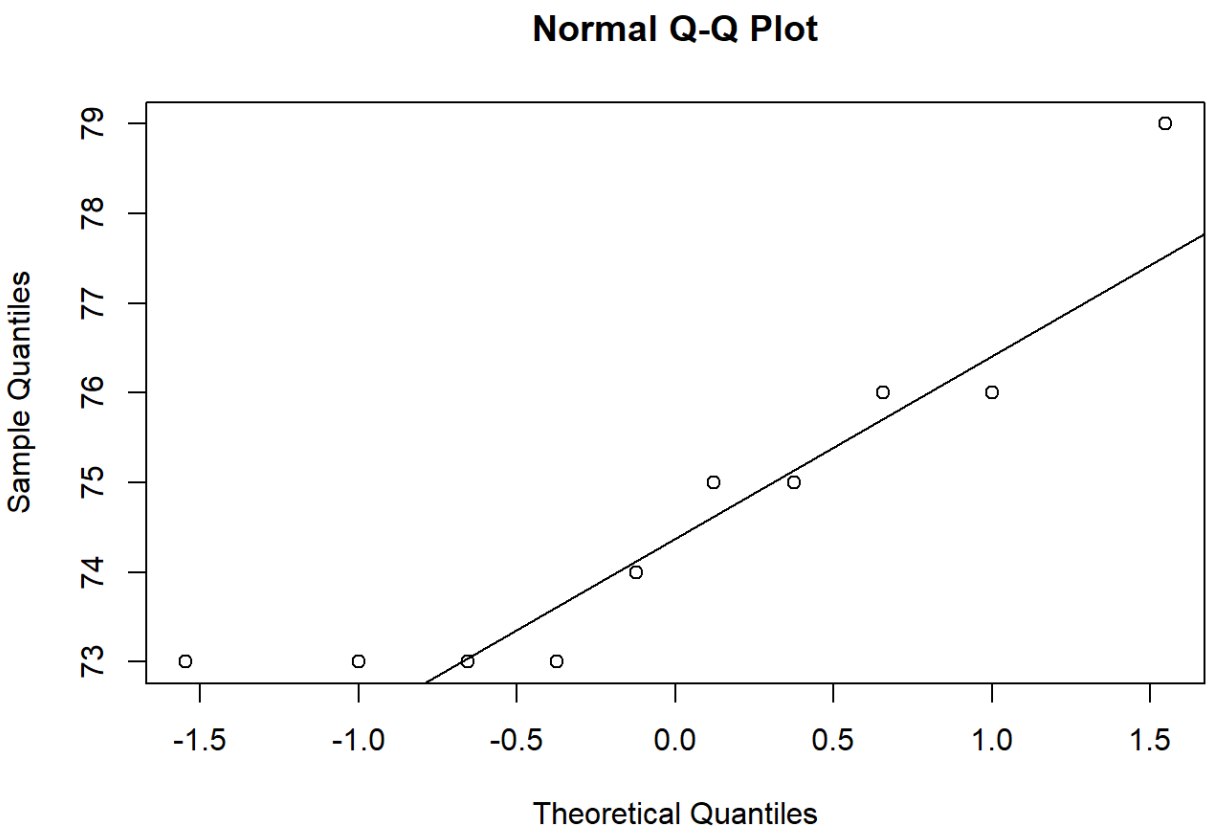
$$(\overline{X} - t^* SE, \overline{X} + t^* SE)$$

- If n is **large enough** and the **variance** of the observed random variable is **finite**, then the CLT allows us to use the same confidence interval.

Example:

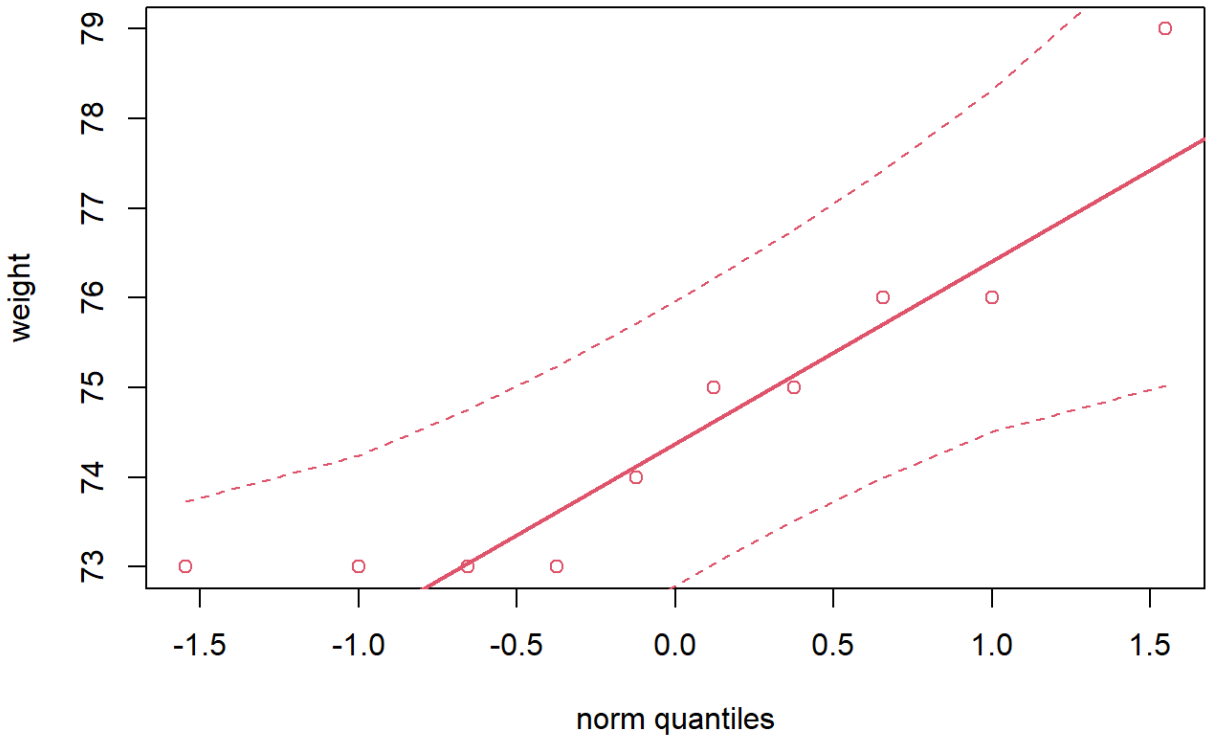
First we make a quick investigation if the data is normally distributed.

```
> qqnorm(weight)
> qqline(weight)
```




```
> 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(weight, "norm")
```



We have a small sample. But as we can see from the graphics we can assume that the data is normally distributed.

Now if we don't assume the standard deviation is 1.5 we can use t-test to estimate 95% confidence interval for the mean.

```
> t.test(weight)

    One Sample t-test

data:  weight
t = 121.36, df = 9, p-value = 8.896e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 73.30755 76.09245
sample estimates:
mean of x
     74.7
```

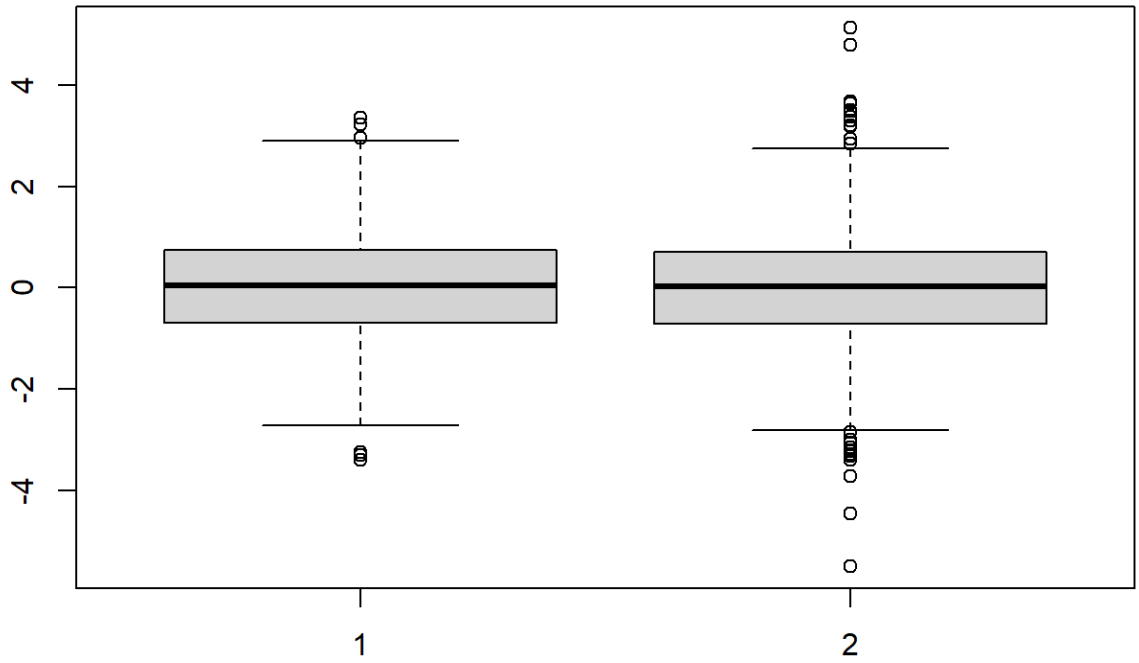
95% confidence interval for the mean of the wight
is (73.3076, 76.0924)

What have you observed comparing the confidence intervals from the z-test and t-test?

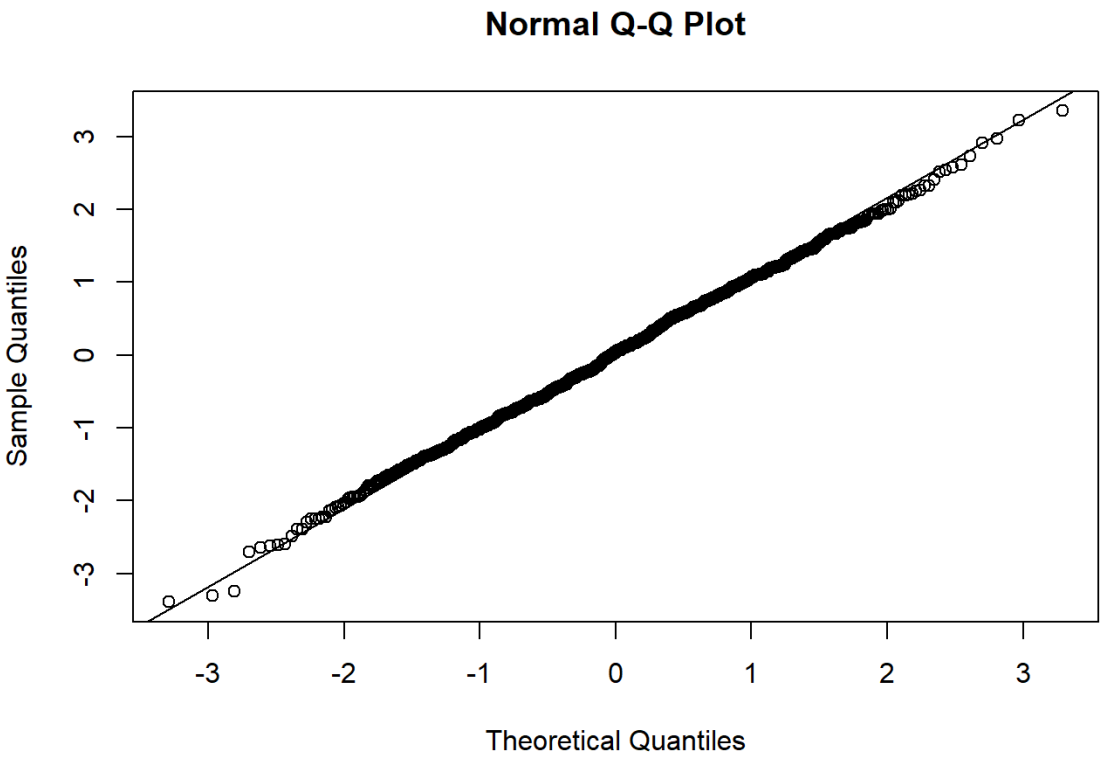
We see that the confidence interval for the t-test $(73.3076, 76.0924)$ is wider than the confidence interval from the z-test $(73.7703, 75.6297)$.

The standard error SE for the t also depends on s which is variable. You can also see the large variance of the t distribution.

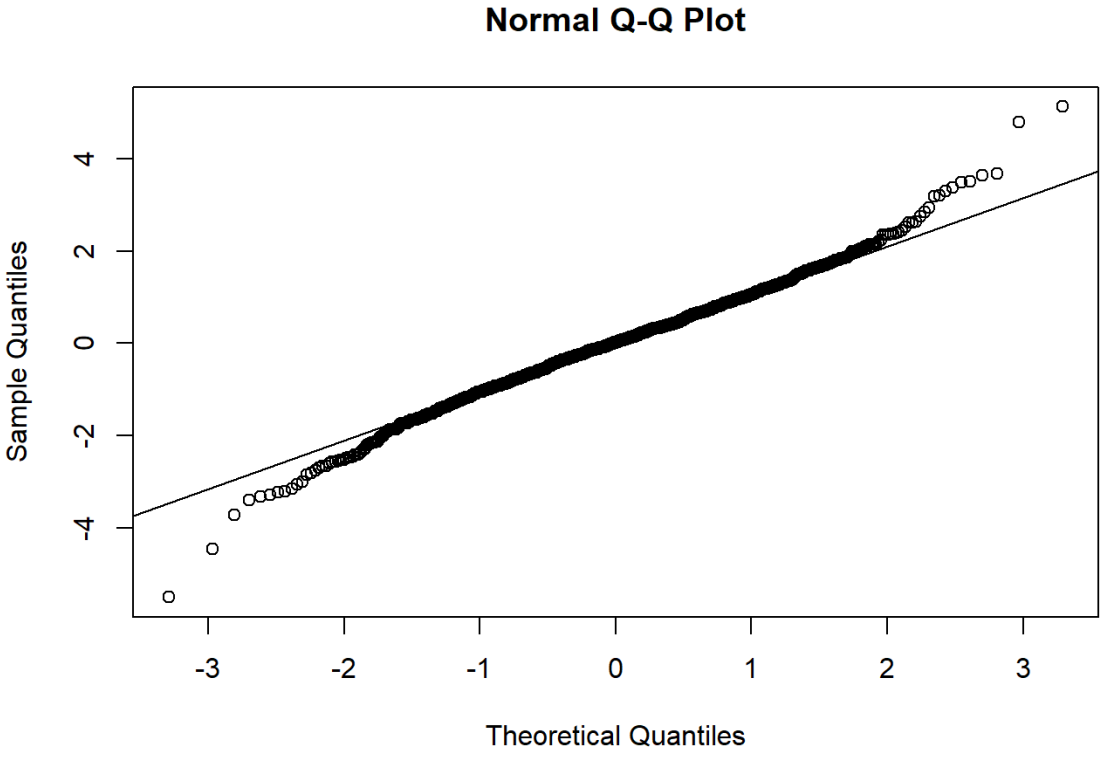
```
> x.norm <- rnorm(1000)
> x.t <- rt(1000, 9)
> boxplot(x.norm, x.t)
```



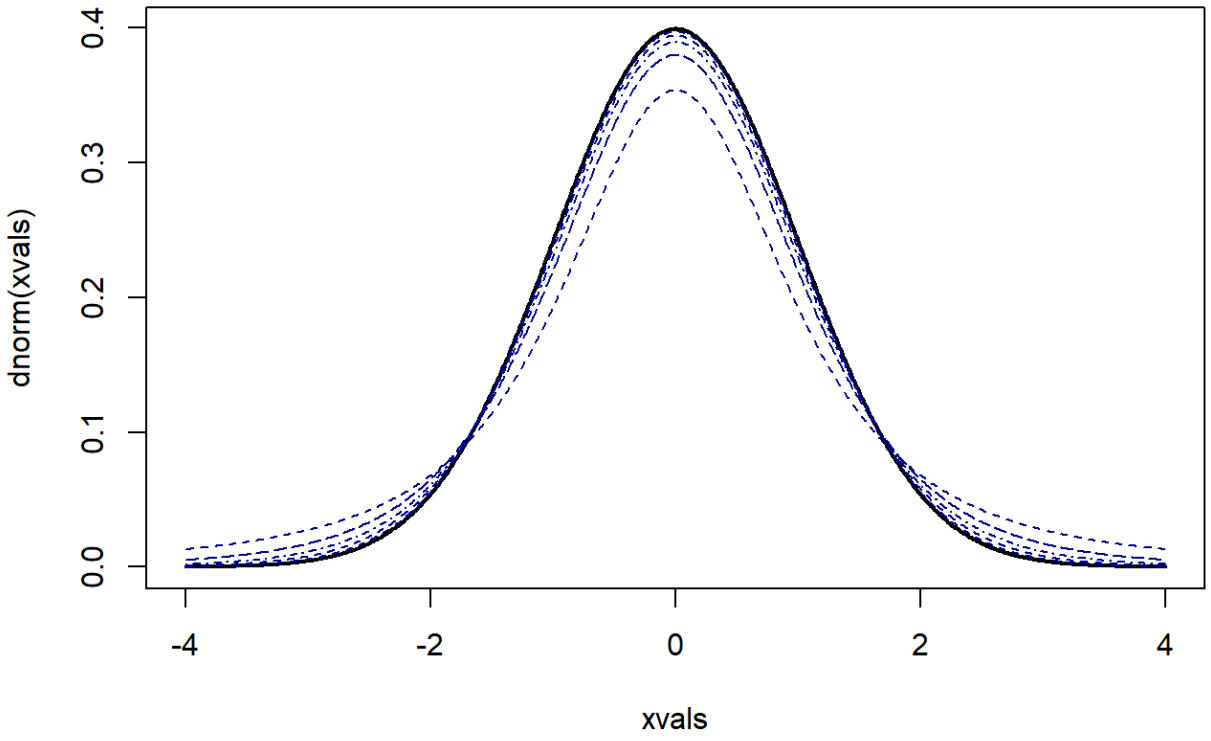
```
> qqnorm(x.norm)
> qqline(x.norm)
```




```
> qqnorm(x.t)
> qqline(x.t)
```



```
> xvals<- seq(-4, 4, 0.01)
> plot(xvals, dnorm(xvals), type = "l", lwd = 2)
> for(i in c(2, 5, 10, 20, 50))
+   points(xvals, dt(xvals, df = i), type = "l", lty = i,
col = "darkblue")
```



Example:

Let's continue the example for the height of the students. If the standard deviation of the population height of the students is unknown.

```
> t.test(survey$Height)

    One Sample t-test

data:  survey$Height
t = 253.07, df = 208, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 171.0380 173.7237
sample estimates:
```

```
mean of x
 172.3809
```

Confidence interval for the population proportion

Let's denote with p the population proportion of successes (or which is the same as probability for success) and with $\hat{p}$ the corresponding sample proportion.

$$p = \frac{\#\ of\ successes\ in\ the\ population}{sie\ of\ the\ population}$$

and

$$p = \frac{\#\ of\ successes\ in\ the\ sample}{sie\ of\ the\ sample}$$

If a **random sample** is made, then we know that $n\hat{p} \in Bi(n, p)$.

If we define random variables $I_1, I_2, \ldots, I_n$ such that $I_i = 1$ when we have "success" and $I_i = 0$ otherwise, then

$$\hat{p} = \frac{I_1 + I_2 + \ldots + I_n}{n}$$

From CLT we know that for large samples $\hat{p} \overset{d}{\approx} N\left(p, \frac{pq}{n}\right)$, so

$$z := \frac{p - \hat{p}}{\sqrt{\frac{pq}{n}}} \overset{d}{\to} Z,\ Z \in N(0,1)$$

We can use the standard normal curve and to conclude that

$$\mathbb{P}\left(-z^* < \frac{p-\hat{p}}{\sqrt{\frac{p(1-p)}{n}}} < z^* \right) \to 1-\alpha$$

where z^* quantile of $N(0,1)$

These inequalities are quadratic with respect to p. When we rise it to the second power and solve it with respect to p we obtain

$$\mathbb{P}\left(\frac{1}{1+\frac{z^*}{n}}(\hat{p}+\Delta_-) < p < \frac{1}{1+\frac{z^*}{n}}(\hat{p}+\Delta_+) \right) \to 1-\alpha$$

where

$$\Delta_\pm = \frac{(z^*)^2}{2n} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(z^*)^2}{4n^2}}$$

As far as z^* is a finite number $\left(\text{usually } z^* \in (-3,3)\right)$

$$\frac{(z^*)^2}{2n} \underset{n\to\infty}{\longrightarrow} 0, \frac{(z^*)^2}{4n^2} \underset{n\to\infty}{\longrightarrow} 0$$

therefore,

$$\Delta_\pm \approx \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\mathbb{P}\left(\frac{1}{1+\frac{z^*}{n}}\left(\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) < p < \frac{1}{1+\frac{z^*}{n}}\left(\hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \right) \to 1-\alpha$$

However

$$\frac{z^*}{n} \underset{n\to\infty}{\longrightarrow} 0$$

so

$$\frac{1}{1+\frac{z^*}{n}} \underset{n\to\infty}{\longrightarrow} 1$$

and

$$\left(\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

and the large sample confidence interval is

$$\left(\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Then a $(1-\alpha)100$ confidence interval is given by

$$(\hat{p} - z^*SE, \hat{p} + z^*SE)$$

where $SE = \sqrt{\frac{p(1-p)}{n}}$. Obviously as n gets bigger we have shorter interval. Because SE gets smaller as far as there is a $\sqrt{n}$ in its denominator.

We have observed that when α decreases z^* increases. Therefore, if we need a bigger interval we have to choose smaller α, which corresponds to a bigger z^*.

We can use the exact distribution of $n\hat{p} \approx Bi(n,p)$ therefore

$$\mathbb{P}\left(b^*_{\frac{\alpha}{2}} \leq n\hat{p} \leq b^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

where $b^*_{\frac{\alpha}{2}}$ is $\frac{\alpha}{2}$ quantile of $Bi(n,p)$ and $b^*_{1-\frac{\alpha}{2}}$ is $1-\frac{\alpha}{2}$ quantile of $Bi(n,p)$.

Therefore,

$$\mathbb{P}\left(\frac{b^*_{\frac{\alpha}{2}}}{n} \leq \hat{p} \leq \frac{b^*_{1-\frac{\alpha}{2}}}{n}\right) = 1-\alpha$$

Due to the statistical definition of probability for large samples $p \approx \tilde{p}$. Therefore in the above considerations we can replace p with $\hat{p}$ and vise verca. And we obtain

$$\mathbb{P}\left(\frac{b^*_{\frac{\alpha}{2}}}{n} \leq p \leq \frac{b^*_{1-\frac{\alpha}{2}}}{n}\right) = 1 - \alpha$$

where $b^*_{\frac{\alpha}{2}}$ is $\frac{\alpha}{2}$ quantile of $Bi(n, \hat{p})$ and $b^*_{1-\frac{\alpha}{2}}$ is $1 - \frac{\alpha}{2}$ quantile of $Bi(n, \hat{p})$.

In this way we obtain the confidence interval

$$\left(\frac{b_{\frac{\alpha}{2}^*}}{n}, \frac{b^*_{1-\frac{\alpha}{2}}}{n}\right)$$

This confidence interval is implemented in the `prop.test` function in R.

Example:

It is reported that 100 people were surveyed and 42 of them liked brand X.

```
> x <- rbinom(100, 1, 0.42)
```

We have to choose α.

If we would like to construct $68\,\%$ – confidence interval, then $1 - \alpha = 0.68$ and $\alpha = 0.32$, therefore $z^* \approx 1$. As we know we can calculate this more precisely in R using

```
> alpha <- 0.32
> qnorm(alpha / 2)
[1] -0.9944579
> qnorm(1 - alpha/2)
[1] 0.9944579
```

$$\mathbb{P}\left(-1 < \frac{p - \hat{p}}{SE} < 1\right) = 0.68$$

Or in particular, on average $68\,\%$ of the intervals $(\hat{p} - SE, \hat{p} + SE)$ contains the true value of p.

Analogously if $\alpha = 0.05$

```
> alpha <- 0.05
> qnorm(alpha / 2)
[1] -1.959964
> qnorm(1 - alpha/2)
[1] 1.959964
```

therefore $z^* \approx 2$ and

$$\mathbb{P}\left(-2 < \frac{p - \hat{p}}{SE} < 2 \right) = 0.95$$

Or in particular, on average $95\,\%$ of the intervals $(\hat{p} - 2SE, \hat{p} + 2SE)$ contain the true value of p.

If $\alpha = 0.01$

```
> alpha <- 0.01
> qnorm(alpha / 2)
[1] -2.575829
> qnorm(1 - alpha/2)
[1] 2.575829
```

therefore $z^* = 3$ and

$$\mathbb{P}\left(-3 < \frac{p - \hat{p}}{SE} < 3 \right) = 0.998$$

Or in particular, on average $99.8\,\%$ of the intervals $(\hat{p} - 3SE, \hat{p} + 3SE)$ contain the true value of p.

For any chosen **confidence level /ниво на доверие/** $\alpha \in (0,1)$ we can find $z^* = z_{1-\frac{\alpha}{2}}$ or $(1-\alpha)100$ confidence interval

$$\mathbb{P}(-z^* < z < z^*) = 1 - \alpha$$

We can see the value of z^* for different α in the following way

```
> alpha <- c(0.2, 0.1, 0.05, 0.001)
> zstar <- qnorm(1 - alpha/2)
> zstar
[1] 1.281552 1.644854 1.959964 3.290527
```

Notice the value $z^* = 1.96$ corresponds to $\alpha = 0.05$ or a $95\,\%$ confidence interval.

If we know z^* and would like to see the value of α we can do this via the `pnorm` function

```
> 2 * (1 - pnorm(zstar))
[1] 0.200 0.100 0.050 0.001
```

We can compute the $99\,\%$ confidence interval as

```
> n <- length(x)
> phat <- mean(x)
> alpha <- 0.01
> zstar <- qnorm(1 - alpha / 2)
> SE <- sqrt(phat * (1 - phat) / n)
> phat + c(-zstar*SE, zstar*SE)
[1] 0.2833121 0.5366879
```

We can compute the $95\,\%$ confidence interval as

```
> n <- length(x)
> phat <- mean(x)
> alpha <- 0.05
> zstar <- qnorm(1 - alpha / 2)
> SE <- sqrt(phat * (1 - phat) / n)
> phat + c(-zstar*SE, zstar*SE)
[1] 0.3136024 0.5063976
```

We can compute the $68\,\%$ confidence interval as

```
> n <- length(x)
> phat <- mean(x)
> alpha <- 0.32
> zstar <- qnorm(1 - alpha / 2)
> SE <- sqrt(phat * (1 - phat) / n)
> phat + c(-zstar*SE, zstar*SE)
```

```
[1] 0.3610892 0.4589108
```

In order to compute

$$\left(\frac{b^*_{\frac{\alpha}{2}}}{n}, \frac{b^*_{1-\frac{\alpha}{2}}}{n} \right)$$

confidence interval we can use the `prop.test` function in R.

```
> prop.test(42, 100, conf.level = 0.99)

    1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.2972701 0.5530641
sample estimates:
   p
0.42
```

99 % confidence interval is (0.297, 0.553)

```
> prop.test(42, 100)

    1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3233236 0.5228954
sample estimates:
   p
0.42
```

95 % confidence interval is (0.323, 0.523)

```
> prop.test(42, 100, conf.level = 0.68)
```

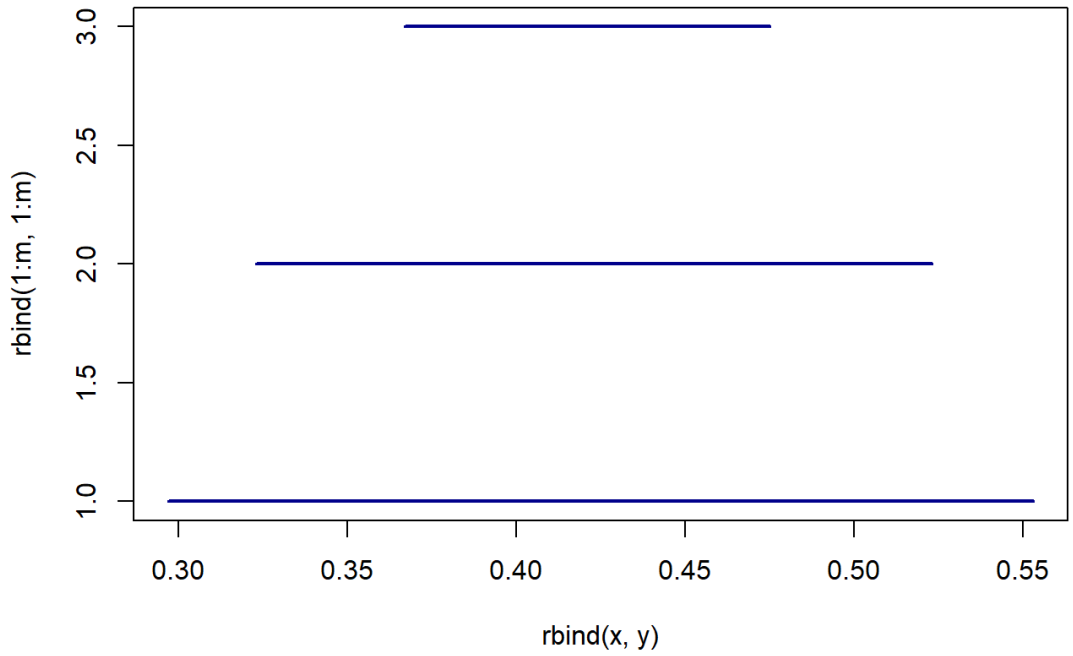
```
    1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
68 percent confidence interval:
 0.3670666 0.4746590
sample estimates:
   p
0.42
```

68 % confidence interval is (0.367, 0.475)

Notice that when we want to be more confident we need wider confidence interval.

```
> m <- 3
> x <- c(0.297, 0.323, 0.367)
> y <- c(0.553, 0.523, 0.475)
> matplot(rbind(x, y),
+         rbind(1:m, 1:m),
+         type="l",
+         lty=1,
+         lwd = 2,
+         col = "darkblue")
```



Confidence interval for the median

Wilcoxon signed rank exact test

Is a **non-parametric** confidence interval for the median. First we arrange the data increasingly. The first ordered statistic is the minima. The n-th ordered statistic is the maxima. Then, we compute the median Me and center the sample with the median $X_{(i)} - Me$. Let R_i is the rank of the i-th observation in this sequence and Z_i is an indicator variable $Z_i = 0$ if $X_{(i)} - Me < 0$ and $Z_i = 1$ otherwise. Wilcoxon uses the statistic

$$W = \sum_{i=1}^{n} Z_i R_i$$

The corresponding theoretical distribution and quantiles are implemented in function `wilcox.test`. Therefore, we compute the corresponding confidence interval by using this function.

Example:

Pay of CEO's in America in 2001 dollars is

```
> x <- c(110, 12, 2.5, 98, 1017, 540, 54, 4.3, 150, 432)
```

Unlike the above tests, we need to specify that we want a confidence interval computed.

```
> range(x)
[1]    2.5 1017.0
> wilcox.test(x, conf.int = TRUE)

    Wilcoxon signed rank exact test

data:  x
V = 55, p-value = 0.001953
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
  33.0 514.5
sample estimates:
(pseudo)median
           150
```

Confidence interval is enormous as the size of the sample is small and the range is huge.

We couldn't have used a t-test as the data isn't normal.

Confidence interval for the standard deviation

Chi-square test

We already know that if the observed random variable is $X \in N(\mu, \sigma^2)$ then the observations on it are $X_i \in N(\mu, \sigma^2)$ and

$$\frac{X_i - \mu}{\sigma} \in N(0,1),\ i = 1, 2, \ldots, n$$

There are two cases

- If μ is known. Then

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \in \mathcal{X}^2(n)$$

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \in \mathcal{X}^2(n)$$

Here we can build two-sided or one-sided confidence interval

- Two-sided CI

As far as

$$\mathbb{P}\left(x^*_{\frac{\alpha}{2}} < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} < x^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

where $x^*_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ of $\mathcal{X}^2(n)$ distribution and $x^*_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ of $\mathcal{X}^2(n)$.

We take the reciprocal values in the inequalities and as far as $\frac{1}{x}$ is a decreasing function the inequalities turn in the opposite direction and we obtain

$$\mathbb{P}\left(\frac{1}{x^*_{\frac{\alpha}{2}}} > \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\mu)^2} > \frac{1}{x^*_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

Now we leave only σ^2 in the middle

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{\frac{\alpha}{2}}} > \sigma^2 > \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

So we obtain the confidence interval for the variance

$$\left(\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{\frac{\alpha}{2}}} \right)$$

As far as $\sqrt{x}$ is an increasing function

$$\mathbb{P}\left(\sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{\frac{\alpha}{2}}}} > \sigma > \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{1-\frac{\alpha}{2}}}} \right) = 1 - \alpha$$

So we obtain the confidence interval for the standard deviation

$$\left(\sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{x^*_{\frac{\alpha}{2}}}} \right)$$

Example:

Let's continue the example for the height of the students. If the mean of the population height of the students is known and equal to 172 and $\alpha = 0.05$ then the two-sided confidence interval for the standard deviation can be computed in the following way.

```
> height <- survey$Height[!is.na(survey$Height)]
> n <- length(height)
> s <- sqrt(sum((height - 172)^2))
> alpha <- 0.05
> xstar1 <- qchisq(1 - alpha/2, n)
> xstar2 <- qchisq(alpha/2, n)
> leftend <- sqrt(s^2 / xstar1)
> rightend <- sqrt(s^2 / xstar2)
> c(leftend, rightend)
[1]  8.972384 10.873545
```

- One-sided CI

As far as

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} < x^*_{1-\alpha}\right) = 1-\alpha$$

where $x^*_{1-\alpha}$ is the $1-\alpha$ of $\mathcal{X}^2(n)$ distribution.

We take the reciprocal values in the inequalities and as far as $\frac{1}{x}$ is a decreasing function the inequalities turn in the opposite direction and we obtain

$$\mathbb{P}\left(\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\mu)^2} > \frac{1}{x^*_{1-\alpha}}\right) = 1-\alpha$$

Now we leave only σ^2

$$\mathbb{P}\left(\sigma^2 > \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{1-\alpha}}\right) = 1-\alpha$$

So, we obtain the following one-sided confidence interval for the variance

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{1-\alpha}}, \infty\right)$$

As far as $\sqrt{x}$ is an increasing function

$$\mathbb{P}\left(\sigma > \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{1-\alpha}}}\right) = 1-\alpha$$

So, we obtain the following one-sided confidence interval for the standard deviation

$$\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{x^*_{1-\alpha}}}, \infty\right)$$

Example:

Let's continue the example for the height of the students. If the mean of the population height of the students is known and equal to 172 and $\alpha = 0.05$ then the one-sided confidence interval for the standard deviation can be computed in the following way.

```
> n <- length(height)
> s <- sqrt(sum((height - 172)^2))
> alpha <- 0.05
> xstar <- qchisq(1 - alpha, n)
> leftend <- sqrt(s^2 / xstar)
> leftend
[1] 9.104016
```

- If μ is unknown, then we replace μ with the average of the sample and the degrees of freedom of $\mathcal{X}^2$ distribution become $n - 1$.

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \in \mathcal{X}^2(n-1)$$

$$\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} \in \mathcal{X}^2(n-1)$$

Here we can build two-sided or one-sided confidence interval

- Two-sided CI

$$\mathbb{P}\left(x^*_{\frac{\alpha}{2}} < \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} < x^*_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

where $x^*_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ of $\mathcal{X}^2(n-1)$ distribution and $x^*_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ of $\mathcal{X}^2(n-1)$.

We take the reciprocal values in the inequalities and as far as $\frac{1}{x}$ is a decreasing function the inequalities turn in the opposite direction and we obtain

$$\mathbb{P}\left(\frac{1}{x^*_{\frac{\alpha}{2}}} > \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} > \frac{1}{x^*_{1-\frac{\alpha}{2}}} \right) = 1 - \alpha$$

Now we leave only σ^2 in the middle

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{\frac{\alpha}{2}}} > \sigma^2 > \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

So we obtain the confidence interval for the variance

$$\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{\frac{\alpha}{2}}}\right)$$

As far as $\sqrt{x}$ is an increasing function

$$\mathbb{P}\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x_{\frac{\alpha}{2}}}} > \sigma > \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{1-\frac{\alpha}{2}}}}\right) = 1-\alpha$$

So we obtain the confidence interval for the standard deviation

$$\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x^*_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{x_{\frac{\alpha}{2}}}}\right)$$

Example:

Let's continue the example for the height of the students. If the mean of the population height of the students is unknown and $\alpha = 0.05$ then the two-sided confidence interval for the standard deviation can be computed in the following way.

```
> n <- length(survey$Height)
> s <- sd(survey$Height, na.rm = TRUE)
> alpha <- 0.05
> xstar1 <- qchisq(1 - alpha/2, n - 1)
> xstar2 <- qchisq(alpha/2, n - 1)
> leftend <- sqrt(s^2 * (n-1) / xstar1)
> rightend <- sqrt(s^2 * (n-1) / xstar2)
```

```
> c(leftend, rightend)
[1]  9.033603 10.823901
```

Here we have used that

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- One-sided CI

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} < x^*_{1-\alpha} \right) = 1 - \alpha$$

where $x^*_{1-\alpha}$ is the $1-\alpha$ of $\mathcal{X}^2(n-1)$ distribution.

We take the reciprocal values in the inequalities and as far as $\frac{1}{x}$ is a decreasing function the inequalities turn in the opposite direction and we obtain

$$\mathbb{P}\left(\frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} > \frac{1}{x^*_{1-\alpha}} \right) = 1 - \alpha$$

Now we leave only σ^2

$$\mathbb{P}\left(\sigma^2 > \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{x^*_{1-\alpha}} \right) = 1 - \alpha$$

So, we obtain the following one-sided confidence interval for the variance

$$\left(\sigma^2 > \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{x^*_{1-\alpha}}, \infty \right)$$

As far as $\sqrt{x}$ is an increasing function

$$\mathbb{P}\left(\sigma^2 > \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{x^*_{1-\alpha}}}\right) = 1 - \alpha$$

So, we obtain the following one-sided confidence interval for the standard deviation

$$\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{x^*_{1-\alpha}}}, \infty\right)$$

Example:

Let's continue the example for the height of the students. If the mean of the population height of the students is unknown and $\alpha = 0.05$ then the one-sided confidence interval for the standard deviation can be computed in the following way.

```
> n <- length(survey$Height)
> s <- sd(survey$Height, na.rm = TRUE)
> alpha <- 0.05
> xstar <- qchisq(1 - alpha, n - 1)
> leftend <- sqrt(s^2 * (n-1) / xstar)
> leftend
[1] 9.158673
```

Here we have used that

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

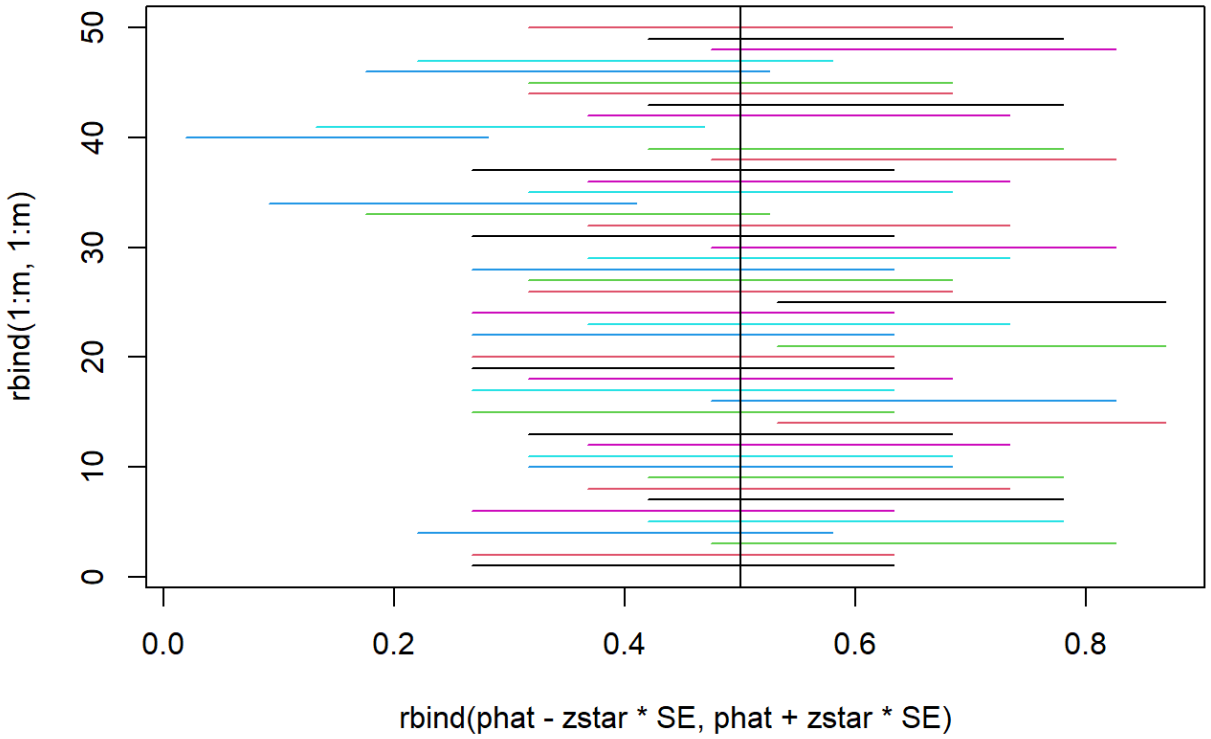
$$(n-1)s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Confidence interval isn't always right

The fact that not all confidence intervals contain the true value of the parameter is often illustrated by plotting a number of random confidence intervals at once.

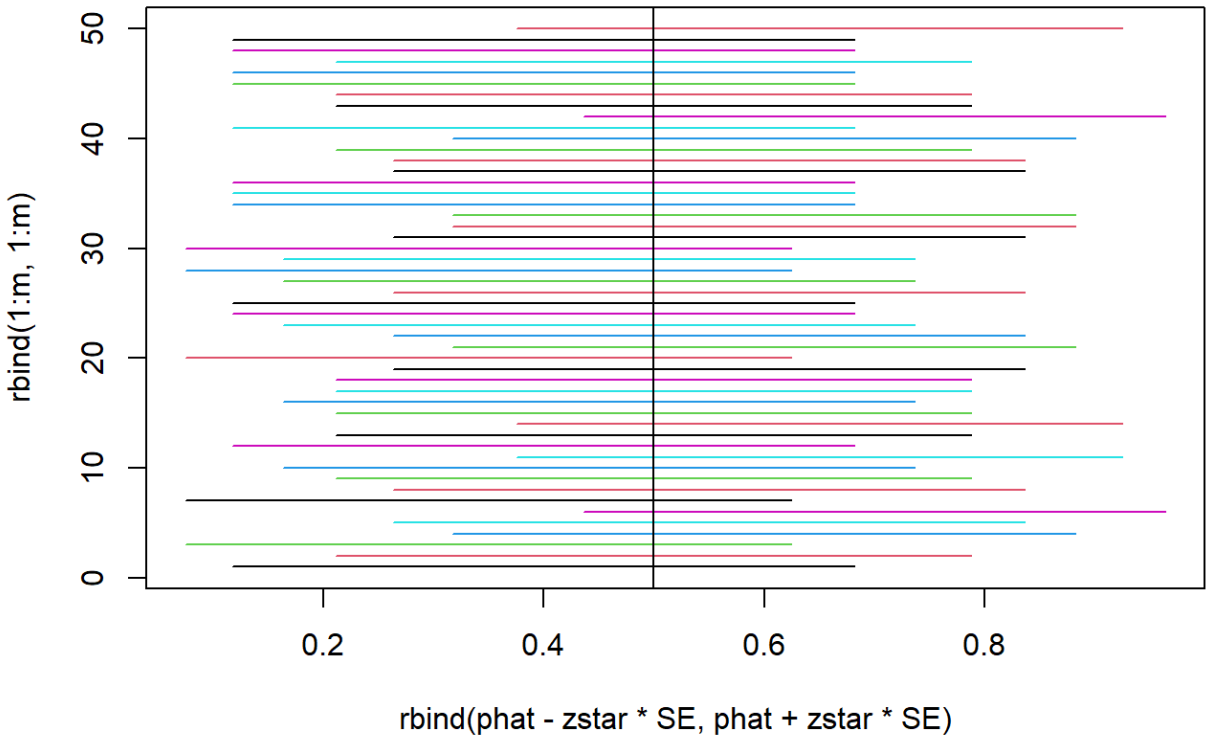
Let's toss 20 coins 50 times and make a CI for the proportions.

```
> m <- 50; n <- 20; p <- 1/2;
> phat <- rbinom(m, n, p) / n
> SE <- sqrt(phat * (1-phat)/n)
> alpha <- 0.10
> zstar <- qnorm(1 - alpha/2)
> matplot(rbind(phat - zstar*SE, phat + zstar*SE),
+         rbind(1:m, 1:m),
+         type = "l",
+         lty = 1)
> abline(v = p) # draw line for p = 1/2
```



This means that in $90\,\%$ of the time population proportion will be inside the CI.

```
> m <- 50; n <- 20; p <- 1/2;
> phat <- rbinom(m, n, p) / n
> SE <- sqrt(phat * (1-phat)/n)
> alpha <- 0.01
> zstar <- qnorm(1 - alpha/2)
> matplot(rbind(phat - zstar*SE, phat + zstar*SE),
+         rbind(1:m, 1:m),
+         type = "l",
+         lty = 1)
> abline(v = p) # draw line for p = 1/2
```



This means that in $99\,\%$ of the time population proportion will be inside the CI.