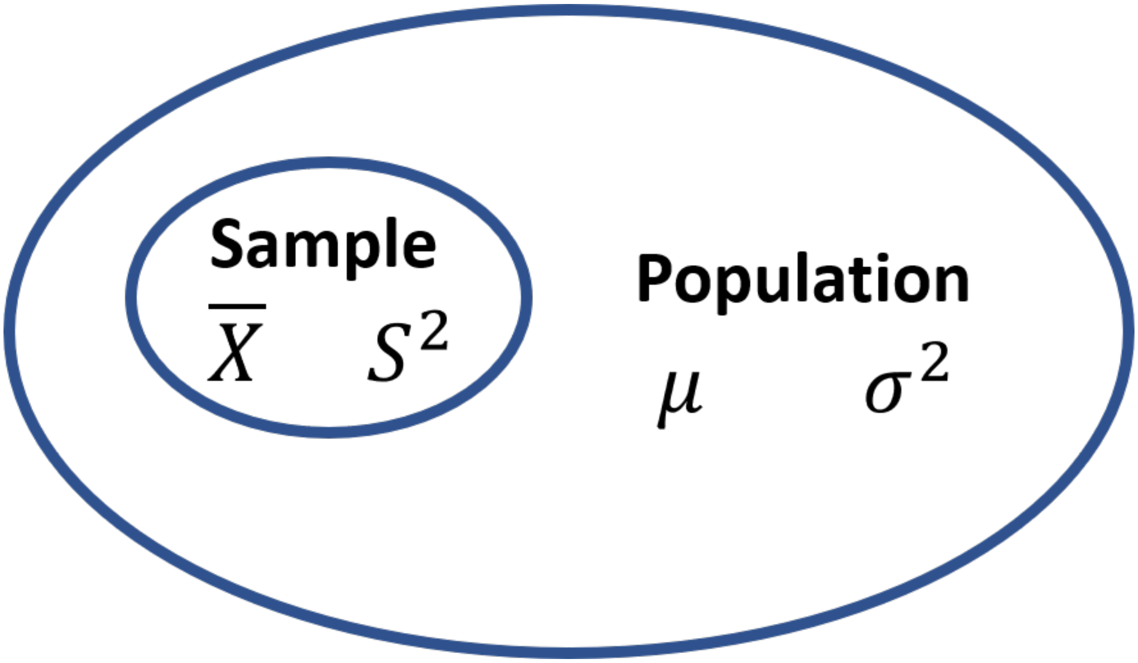
Chapter 4: Confidence intervals

Contents

1 General idea



A **confidence interval (CI)** for a population parameter is an interval with an associated probability p, such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample, a proportion p of the confidence intervals would contain the population parameter in question.

The **confidence level** of a confidence interval gives the probability that the calculated interval includes the true value of the parameter. [1] If the sampling was repeated numerous times and the confidence interval recalculated from each sample, a proportion p of the confidence intervals would contain the population parameter in question. This proportion corresponds to the confidence level.
In other words, if independent samples are taken repeatedly from the same population and a confidence interval is calculated for each sample, then a certain percentage (the confidence level) of the intervals will include the unknown population parameter.

The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter.

Confidence intervals are used in conjunction with point estimates to convey information about the uncertainty of the estimates.

2 Construction of a 95% confidence interval for μ, in case σ is known.

Example *Temperature*:
Import the data set *temp_warm.txt* as `temperature`
Can you give an estimate of the average October temperature in Western Europe?

We know that, based on our sample, the summary statistics for the October temperature in Western Europe are

```
westT <- subset(temperature, temperature$Area=="West", select = October)
summary(westT) # descriptive statistics
```

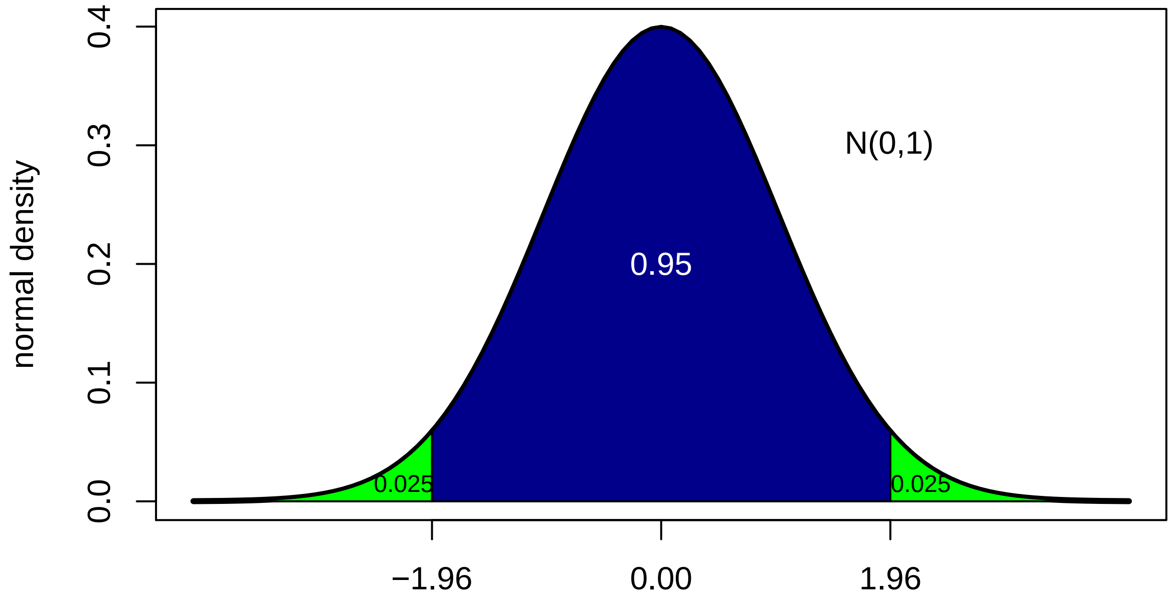
```
##     October
##  Min.   : 8.90
##  1st Qu.: 9.80
##  Median :11.10
##  Mean   :10.94
##  3rd Qu.:11.50
##  Max.   :13.50
```

The average October temperature in Western Europe is 10.94°C. What can we say about the population average μ?

2.1 Derivation

1. Use of the CLT, for n large enough:
   $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
   $\Rightarrow \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \approx N(0,1)$
2. Property of $N(0,1)$

[1] http://www.stat.yale.edu/Courses/1997-98/101/confint.htm



Hence, $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ are between -1.96 and 1.96 with a probability of 0.95.

$P(-1.96 \leq \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \leq 1.96) = 0.95$

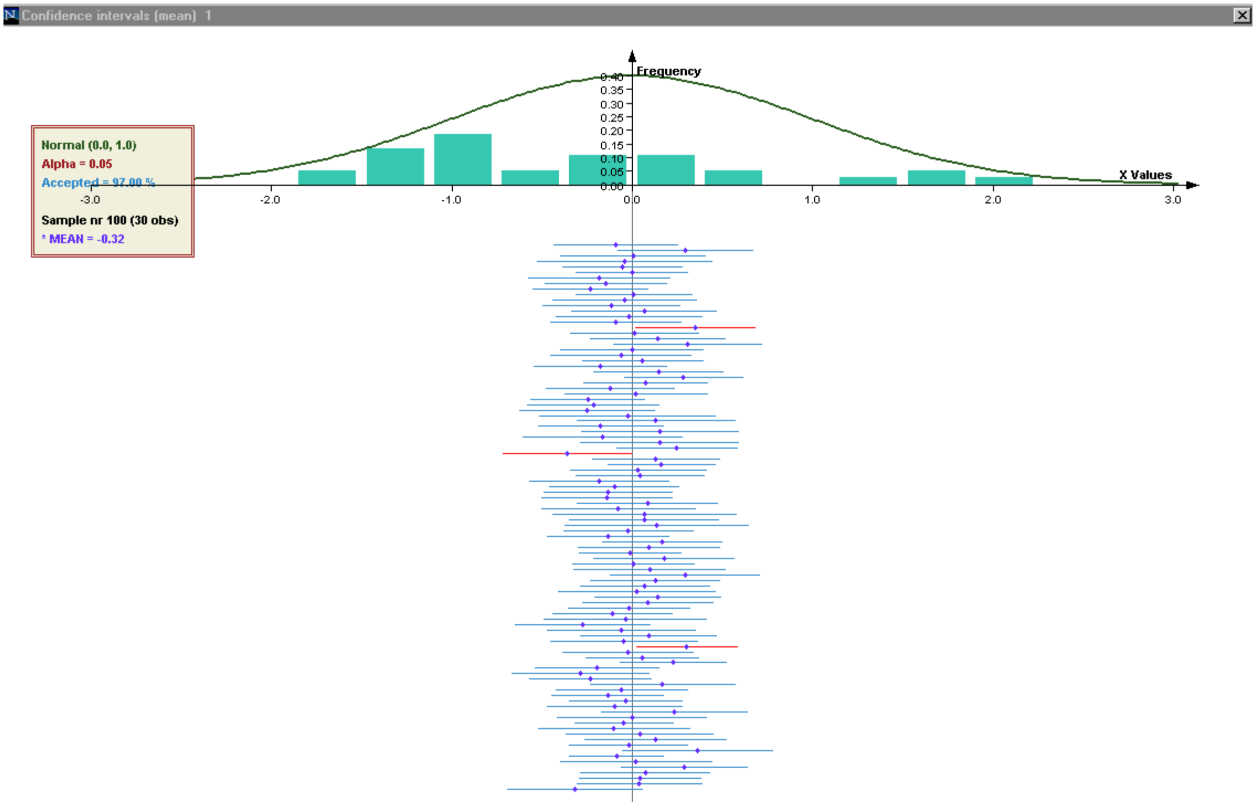
$\Leftrightarrow P(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$

$\Leftrightarrow P(-\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$

$\Leftrightarrow P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$

95% confidence interval for μ (in case σ is known) is given by $\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

To illustrate the concept of a confidence interval, we look at an applet http://lstat.kuleuven.be/newjava/vestac/

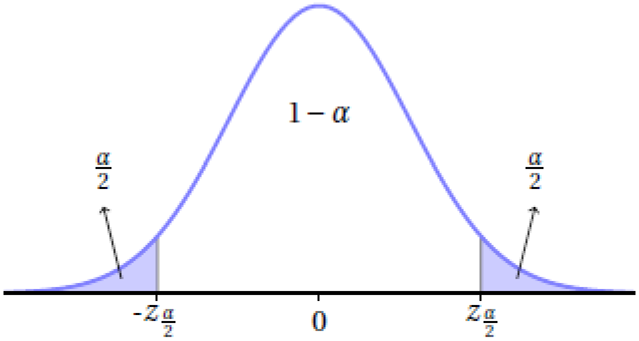

Confidence intervals (mean) 1

Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 97.00 %
Sample nr 100 (30 obs)
* MEAN = -0.32

Remember:
A **confidence interval (CI)** for a population parameter is an interval with an associated probability p, such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample, a proportion p of the confidence intervals would contain the population parameter in question.

2.2 Changing the confidence level (in case σ is known)

We will change the confidence level of 0.95 to a value $1-\alpha$. The uncertainty α is divided equally at the two sides of the normal distribution so that



A $(1-\alpha)\cdot 100\%$ - confidence interval for μ (in case σ is known), is given by $\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$

Here is $z_{\frac{\alpha}{2}}$ the value for which
$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$
$\Leftrightarrow P(Z \leq -z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$
$\Leftrightarrow P(Z \leq z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$

with $Z \sim N(0,1)$.

Control that $z_{0.025} = 1.96$.

Summary: Confidence interval for μ when σ is known:

Confidence level	Confidence interval
0.90	$\overline{x} \pm 1.645 \cdot \frac{\sigma}{\sqrt{n}}$
0.95	$\overline{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$
0.99	$\overline{x} \pm 2.576 \cdot \frac{\sigma}{\sqrt{n}}$

2.2.1 In R

Example *Temperature*:
Import the data set *temp_warm.txt* as `temperature`
Construct a 95% confidence interval for the average October temperature in Western Europe. We assume that σ is known to be 1.5°. We have a sample of nine cities.

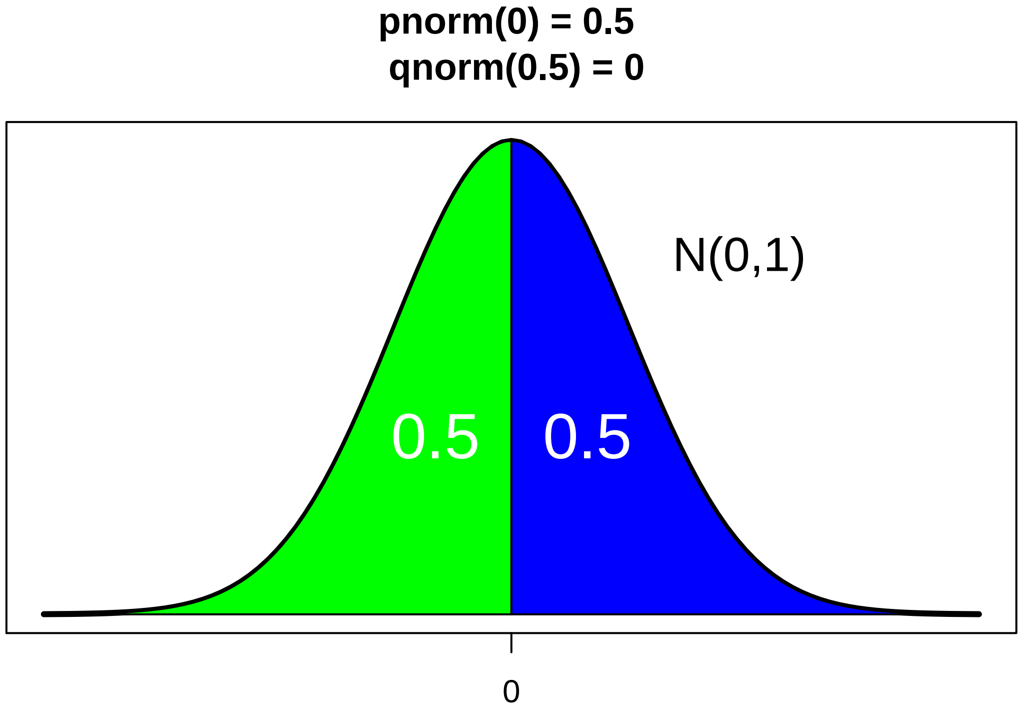
95% CI is given by $\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

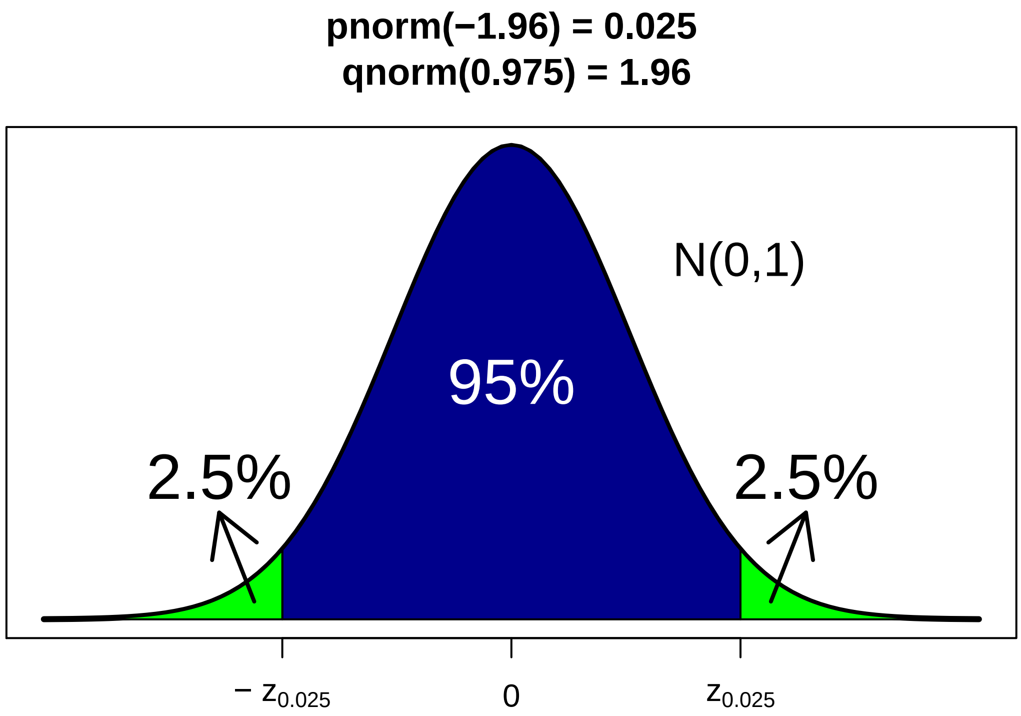
1. Compute the critical values for the $N(0,1)$ distribution.

Function in R	Description
`pnorm(x, mean = 0, sd = 1)`	Normal cumulative distribution function $P(X \leq x)$ when `x` is given
`qnorm(p, mean = 0, sd = 1)`	Quantile function when `p` is given

We use `qnorm` to obtain the quantiles of the $N(0,1)$.

$P(X \leq x)$	Quantile function
`pnorm(x = 0) = 0.5`	`qnorm(p = 0.5) = 0`
`pnorm(x = -1.96) = 0.025`	`qnorm(p = 0.025) = -1.96`
`pnorm(x = 1.96) = 0.975`	`qnorm(p = 0.975) = 1.96`


pnorm(0) = 0.5
qnorm(0.5) = 0
N(0,1)
0.5
0.5
0



2. Compute the lower and upper bound of the CI

```
confidence_level <- 0.95
alpha <- 1 - confidence_level
sigma <- 1.5
n <- 9
westT <- subset(temperature, temperature$Area=="West", select = October)
xmean <- mean(westT$October)
lcl <- xmean - qnorm(1-alpha/2)*sigma/sqrt(n)
ucl <- xmean + qnorm(1-alpha/2)*sigma/sqrt(n)
result <- list(mean = xmean, lcl = lcl, ucl = ucl)
result
```

```
## $mean
## [1] 10.94444
##
## $lcl
## [1] 9.964462
##
## $ucl
## [1] 11.92443
```

Remark:
You can also use the `z.test` in `BSDA` package.

```
library(BSDA)
z.test(westT$October, sigma.x = 1.5)
```

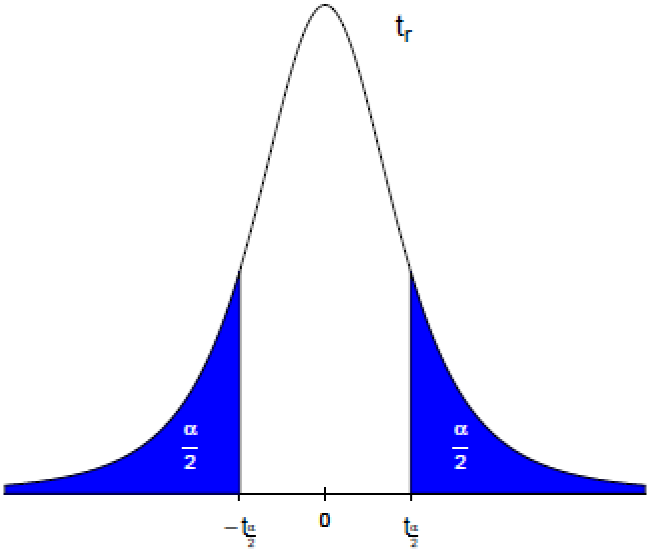
```
##
##  One-sample z-Test
##
## data:  westT$October
## z = 21.889, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.964462 11.924426
## sample estimates:
## mean of x
##  10.94444
```

3 Confidence interval in case σ is not known (real life)

3.1 construction

1. When σ is known and n is large,
   then $\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \approx N(0,1)$,
   then, a $(1-\alpha)\cdot 100\%$- confidence interval for μ is given by $\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$

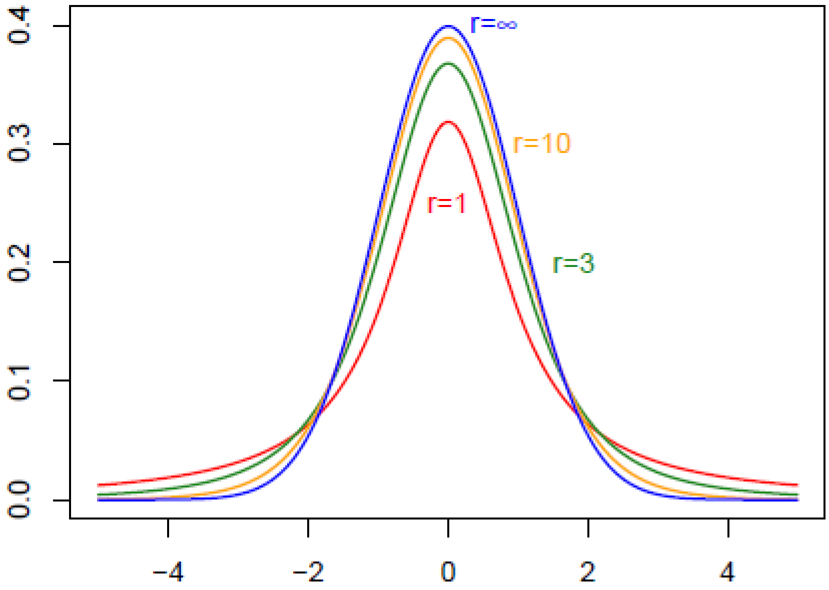
2. When σ is not known, then we are going to estimate σ by its sample estimate S, the sample standard deviation. When n is large, we then can prove that $\frac{\overline{X}-\mu}{S}\sqrt{n} \approx t_{(n-1)}$, where $t_{(n-1)}$ is a Student's t-distribution with $n-1$ degrees of freedom.



We then have
$P(-t_{n-1,\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{S}\cdot\sqrt{n} \le t_{n-1,\frac{\alpha}{2}}) = 1-\alpha$

We say, $\left[\overline{x} - t_{n-1,\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}, \overline{x} + t_{n-1,\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}\right]$ is a $(1-\alpha)\cdot 100\%$- confidence interval for μ (in case σ is not known).

T distribution with r degrees of freedom.



Estimating σ by S has induced an error, hence the heavier tails.

3.2 Exercise

Example *Temperature*:

Import the data set *temp_warm.txt* as `temperature`

Construct a 95% confidence interval for the average October temperature in Western Europe

```
westT <- subset(temperature, temperature$Area=="West", select = October)
t.test(westT$October)
```

```
##
##  One Sample t-test
##
## data:  westT$October
## t = 22.404, df = 8, p-value = 1.667e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.817938 12.070950
## sample estimates:
## mean of x
##  10.94444
```

The sample mean is given by 10.94.

The 95% confidence interval for the average October temperature in Western Europe is then $[9.82, 12.07]$.

Remark:

The construction of a confidence interval is based on the CLT. Hence it can be used whenever the CLT is valid.

Hence, we can use it when the sample size is large (> 25) or when the underlying data is normally distributed. Here, we have a sample of 9 observations. Therefore, we have to check for normality.

```
shapiro.test(westT$October)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  westT$October
## W = 0.95625, p-value = 0.7584
```

Since $p - value = 0.76 > 0.05$, we do not reject H_0.

4 Calculating sample sizes (in case σ is known)

Engineers often need to collect data to estimate a parameter of interest within a given precision. Suppose we wish to estimate a population mean when we have at least a reasonable idea of the population standard deviation. In this case, we may treat σ as known. Thus the form of the confidence interval is $\overline{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

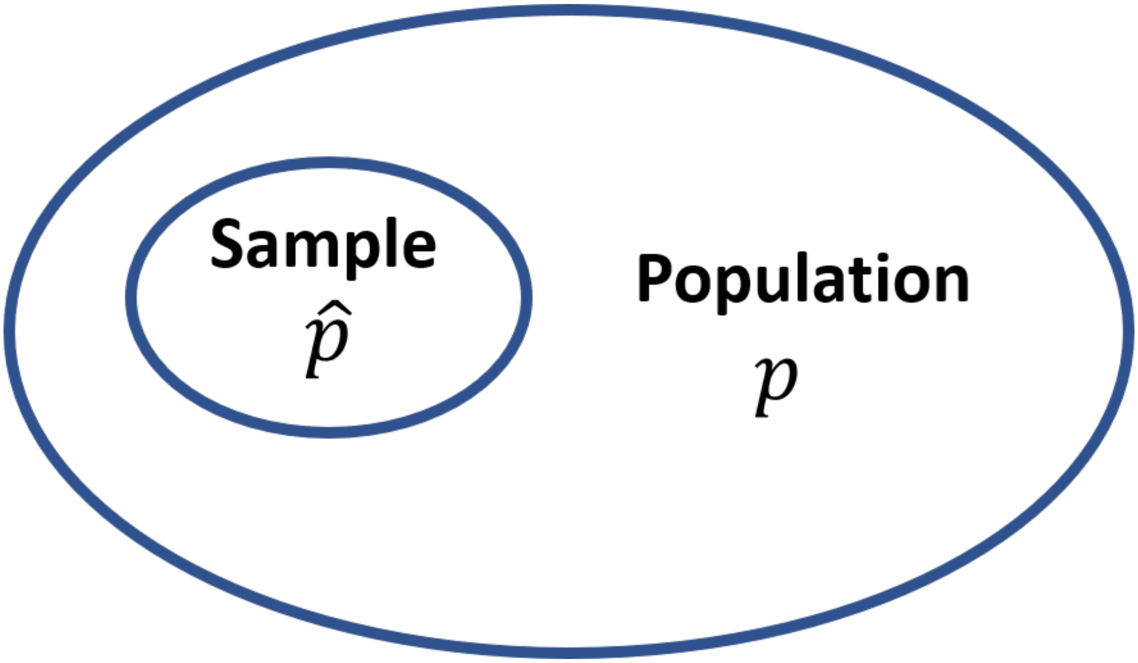
We observe that $\frac{\sigma}{\sqrt{n}}$ controls the width of this interval. Since we know σ, we can make the width of the interval as small as we wish by an appropriate choice of n.

Suppose we wish to estimate this mean to within $\pm B$ units. We need to choose n such that
$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq B$
$\Leftrightarrow B\sqrt{n} \geq z_{\frac{\alpha}{2}}\sigma$
$\Leftrightarrow \sqrt{n} \geq z_{\frac{\alpha}{2}} \frac{\sigma}{B}$
$\Leftrightarrow n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{B}\right)^2$
since n must be an integer, we should always round up.

Example *October temperature Western Europe*:
Suppose we want to have a 95% confidence interval with a width of ± 0.5°. We already know (assume) that the population standard deviation is 1.5°.

By using the previous equation, we obtain:
$n \geq \left(\frac{1.96 \cdot 1.5}{0.5}\right)^2$
$\Leftrightarrow n \geq 34.57$
Hence, we take $n = 35$ because the sample size must be an integer.

5 Confidence interval for a proportion



Example:

240 out of a random sample of 400 people said they preferred to fly to United Kingdom rather than go by Eurostar. Construct a 95% confidence interval of the proportion of people who prefer to fly.

X = number of people who prefer to fly to United Kingdom
$X \sim B(n = 400, p)$, with p the probability one prefers to fly to United Kingdom

In general:

Using the central limit theorem, it can be shown that the binomial distribution with parameters n and p can be approximated by a normal distribution with mean np and variance $np(1-p)$. The approximation works well when $np \geq 5$ and $n(1-p) \geq 5$. In other words,

> If $X \sim B(n, p)$, with X **the number of successes**, $np \geq 5$ and $n(1-p) \geq 5$,
> then $X \sim N(np, np(1-p))$.

The **proportion of successes** $\frac{X}{n}$ can then be approximated as a normal distribution with mean p and variance $\frac{p(1-p)}{n}$, or $\frac{X}{n} \sim N(p, \frac{p(1-p)}{n})$

A confidence interval can then be constructed as follows:

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

This is a $(1-\alpha) \cdot 100\%$-confidence interval for the population proportion p.

Summary: confidence interval for p

Confidence level	Confidence interval
0.90	$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
0.95	$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
0.99	$\hat{p} \pm 2.576\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Example:

240 out of a random sample of 400 people said they preferred to fly to United Kingdom rather than go by Eurostar. Construct a 95% confidence interval of the proportion of people who prefer to fly.

A 95% CI for p is given by $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

$\hat{p} = \frac{success}{n} = \frac{240}{400} = \frac{3}{5} = 0.6$

95% CI:
$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$= 0.6 \pm 1.96\sqrt{\frac{0.6 \cdot 0.4}{400}}$
$= 0.6 \pm 0.04801$

95% Confidence Interval for proportion is $[0.552 \leq p \leq 0.648]$.

5.1 In R: use `prop.test(x, n)`

Confidence interval for a proportion

```
prop.test(240, 400)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  240 out of 400, null probability 0.5
## X-squared = 15.602, df = 1, p-value = 7.815e-05
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5499965 0.6480518
## sample estimates:
##   p
## 0.6
```

6 Exercises

6.1 Bio-informatics

In a bio-informatics context, one knows that a certain biometric parameter has a historic standard deviation of 8.

1. A recent sample of 4 observations yielded a sample mean of 101.4. Construct a 99% confidence interval for the true mean width.
2. Find the sample size required to estimate the true mean width to within ± 2 units using a 99% confidence interval.

6.2 Solutions

```r
conf <- 0.99
alpha <- 1-conf
sigma <- 8
n <- 4
xmean <- 101.4
B <- 2
lcl <- xmean-qnorm(1-alpha/2)*sigma/sqrt(n)
ucl <- xmean+qnorm(1-alpha/2)*sigma/sqrt(n)
n <- (qnorm(1-alpha/2)*sigma/B)^2
result <- list(lcl = lcl, ucl = ucl, n = n)
result
```

```
## $lcl
## [1] 91.09668
##
## $ucl
## [1] 111.7033
##
## $n
## [1] 106.1583
```

Use `zsum.test` from package `BSDA`

```r
zsum.test(mean.x = 101.4, sigma.x = 8, n.x = 4, conf.level = 0.99)
```

```
##
##  One-sample z-Test
##
## data:  Summarized x
## z = 25.35, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##   91.09668 111.70332
## sample estimates:
```

```
## mean of x
##     101.4
```