Chapter 5: Hypothesis testing

Contents

1 Introduction

Questions which will be dealt with in this chapter are:

- Is the average October temperature in Western Europe lower then 12°?
- Is there a significant difference between the average October temperature in West and Eastern Europe?
- Is the variance of the October temperature the same for Western Europe as for Eastern Europe?

These are examples of statistical hypotheses.

A **statistical hypothesis** is a statement concerning a population parameter.

The hypothesis being tested is referred to as the **null hypothesis** and is denoted by H_0. Any hypothesis which differs from the null hypothesis is called an **alternative hypothesis** and is denoted by H_1.

In general, it is not possible to prove that a null hypothesis is true or false. *If the data supports the null hypothesis, then the null hypothesis is* ***not rejected****. If the data does not support the null hypothesis then the null hypothesis is* ***rejected*** *in favor of the alternative.*

2 Error probabilities

2.1 Example

Example ***Criminal court***
To fix ideas, consider a criminal court case. The null hypothesis and alternative hypothesis are
H_0: The accused is innocent
H_1: The accused is guilty

		Decision jury	
		Convict	**Acquit**
State of the accused	**Innocent**	**Type I error**	**OK**
	Guilty	**OK**	**Type II error**

Two types of errors can be made:

- *Type I error*: Convicting an innocent person
- *Type II error*: Acquitting a guilty person

2.2 In general

		Decision of hypothesis test	
		Reject H_0	**Do not reject H_0**
Population	**H_0 true**	**Type I error** **P[reject H_0|H_0 true] = α**	**OK** P[not reject H_0|H_0 true] = 1- α
	H_1 true	**OK** P[reject H_0|H_1 true] = 1-β = power	**Type II error** P[not reject H_0|H_1 true] = β

Type I error, α, is the probability of rejecting the null hypothesis when the null hypothesis is true.
Type II error, β, is the probability of not rejecting the null hypothesis when the null hypothesis is false.

The type I error, α, **is called the *significance level* of the test**. Choose $\alpha = 0.05$ as a rule of thumb.

3 One-sided and two-sided tests

3.1 One-sided test

Example *Temperature*
Is the average October temperature lower than 12°?

$H_0 : \mu = 12$
versus
$H_1 : \mu < 12$

μ is here the average October temperature.
The alternative hypothesis is one-sided and the test is therefore called a **one-sided test**.

3.2 Two-sided test

Example *Temperature*
Is the average October temperature in Western Europe different from 10°?

$H_0 : \mu = 10$
versus
$H_1 : \mu \neq 10$

μ is here the average October temperature in Western Europe.
The alternative hypothesis is two-sided and the test is therefore called a **two-sided test**.

4 Structure of a hypothesis test: the five steps

The general procedure used for testing a hypothesis assumes that

- your sample is large enough so that we can use CLT ($n \geq 25$),
- or (in case you have a limited sample) that your data follows a normal distribution.

The following **general outline** is recommended for hypotheses testing:

Step 1: State the appropriate hypotheses
Null hypothesis: H_0
Alternative hypothesis: H_1

Step 2: Choose a level of significance for the test
This is usually $\alpha = 0.05$, but α can be changed to fit the needs of the problems.

Step 3: State the appropriate test statistics
For example, for hypothesis test around the mean:

- If the variance is know, then use
  $Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$
- If the variance is unknown, then use
  $T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \approx t_{n-1}$

Step 4: Compute the observed value of the test statistic and the $p-value$
Compute the observed value of the test statistic (Z_{obs} or T_{obs}) and compute the corresponding $p-value$.

> $p-value$ = the probability, assuming H_0 is true, of obtaining a value for the test statistic as observed or more extreme.

More extreme is defined in the direction of the alternative hypothesis.
For instance, the $p-value$ could be calculated as: $p-value = P(T \geq T_{obs}|H_0)$.

Step 5: Draw conclusion based on the $p-value$

- If $p-value < \alpha$, then we **reject** H_0 at level α.
  If this $p-value$ is 'very small' then we reject the H_0 and conclude that the H_1 is true.
- If $p-value > \alpha$, then we do **not reject** H_0 at level α.
  If this $p-value$ is 'not small' then the H_0 is not rejected.

> We reject the H_0 in favor of the H_1 if the $p-value$ is smaller than the chosen significance level α.

5 Hypothesis test for the mean, when standard deviation is assumed to be known

5.1 The Z-test

The **Z-test** is a hypothesis test for the mean when the standard deviation is assumed to be known and the normality assumption is satisfied.

5.1.1 Example one-sided Z-test

Example *Temperature*
Is the average October temperature in Western Europe less than 12°? We assume σ to be known and equal to 1.5.

Step 1
The **hypothesis** is stated as

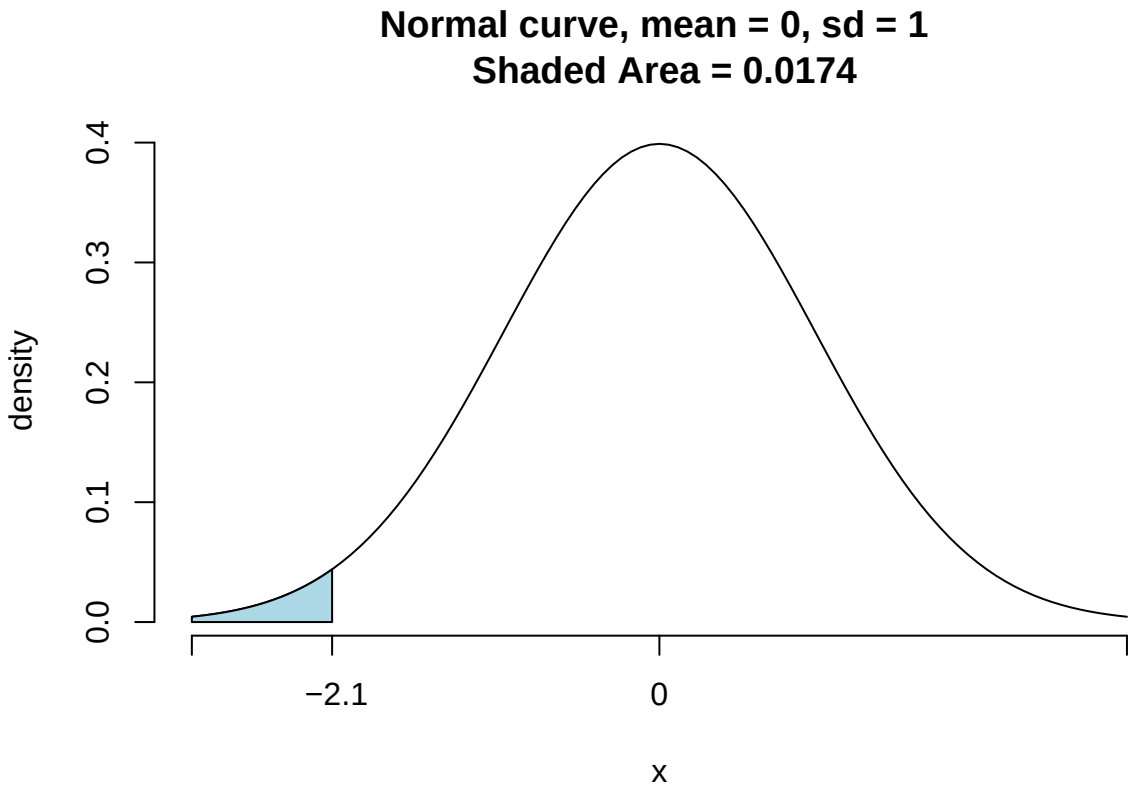
$H_0 : \mu = 12$ versus $H_1 : \mu < 12$
with μ the average October temperature in Western Europe.

Step 2
Select $\alpha = 0.05$

Step 3
Select the test statistic.
We assume here that the standard deviation σ is known to be 1.5 (thus case σ known).
As test statistic we choose $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Step 4

1. **Compute** Z_{obs}
   $Z_{obs} = \frac{10.94-12}{\frac{1.5}{\sqrt{9}}} = -2.1$
2. **Compute** $p - value$
   one-sided $p - value$:
   $p - value$
   $= P(Z \leq z_{obs} | H_0 \text{ is true})$
   $= P(Z \leq -2.1 | H_0 \text{ is true})$
   $= 0.017$



Step 5
Conclusion:
Since the $p - value < 0.05$, we reject H_0.
We have sufficient evidence to conclude that the average October temperature is significant smaller than 12.

5.1.1.1 One-sided Z-test in R

Example *Temperature* in R

```
westT <- subset(temperature, temperature$Area=='West', select=October)
sigma <- 1.5

# Step 1
mu0 <- 12

# Step 2
alpha <- 0.05

# Step 3: Select the test statistic

# Step 4
# Step 4.1: Compute test statistic
n <- length(westT$October)
xmean  <-mean(westT$October)
z  <- (xmean-mu0)/(sigma/sqrt(n))

# Step 4.2: Compute p-value
onesided.pvalue  <- pnorm(z)
result1 <- list(z = z, pvalue= onesided.pvalue)
result1
```

```
## $z
## [1] -2.111111
##
## $pvalue
## [1] 0.01738138
```

Remark:
right one sided $p-value$ is calculated as `right.onesided.pvalue  <- (1-pnorm(z))`

left one-sided $p-value$ is calculated as `left.onesided.pvalue <- pnorm(z)`

Another way:

```
library(BSDA)
z.test(westT$October, alternative="less", mu=12, sigma.x=1.5, conf.level=0.95)
```

```
##
##  One-sample z-Test
##
## data:  westT$October
## z = -2.1111, p-value = 0.01738
## alternative hypothesis: true mean is less than 12
## 95 percent confidence interval:
##        NA 11.76687
## sample estimates:
## mean of x
##  10.94444
```

Remark:
In case you don't have the raw data (only averages and standard deviations), then you can use the `zsum.test` (from package `BSDA`)

Usage of `zsum.test`*:*

```
zsum.test(mean.x, sigma.x = NULL, n.x = NULL, mean.y = NULL,  sigma.y = NULL, n.y = NULL,
          alternative = "two.sided", mu = 0,  conf.level = 0.95)
```

5.1.2 Example two-sided Z-test

Example ***Temperature***
Is the average October temperature in Western Europe different from 10°? We assume σ to be known and equal to 1.5.

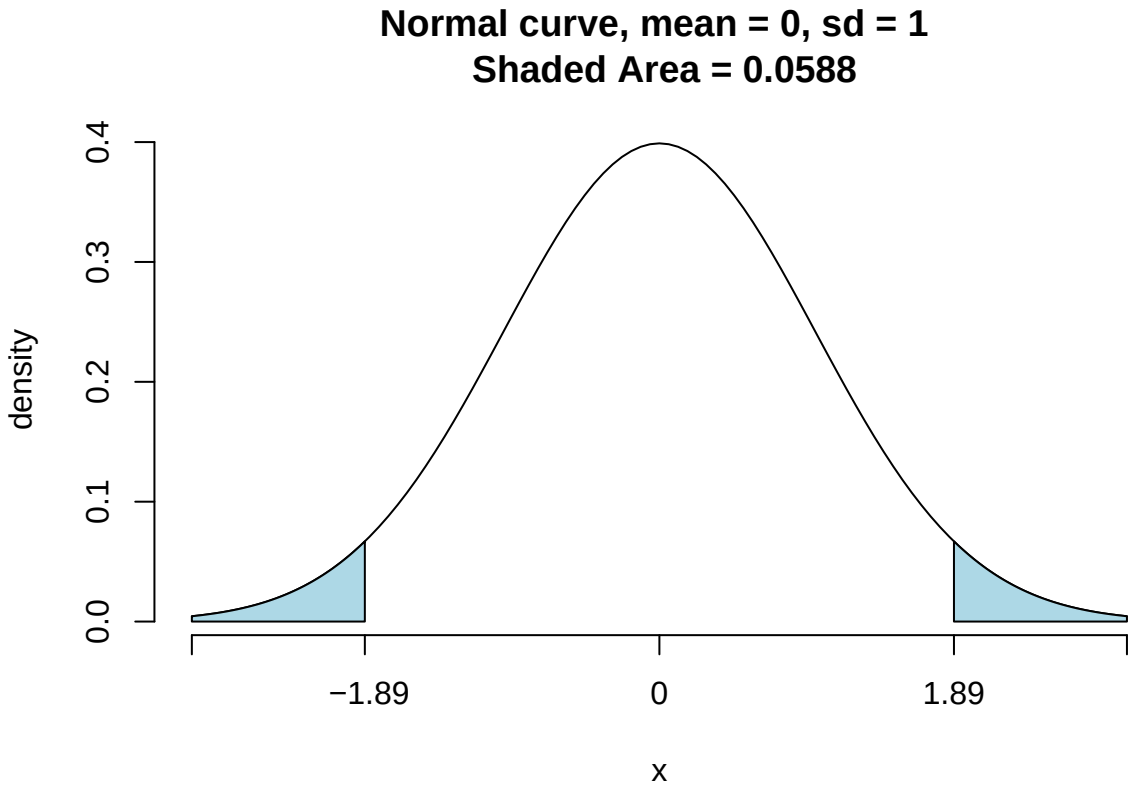
Step 1
The **hypothesis** is stated as
$H_0 : \mu = 10$ versus $H_1 : \mu \neq 10$
with μ the average October temperature in Western Europe.

Step 2
Select $\alpha = 0.05$

Step 3
Select the test statistic.
We assume here that the standard deviation σ is known to be 1.5 (thus case σ known).
As test statistic we choose $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ from which we know that $Z \sim N(0,1)$ under H_0.

Step 4

1. **Compute** Z_{obs}
   $Z_{obs} = \frac{10.94-10}{\frac{1.5}{\sqrt{9}}} = 1.89$
2. **Compute** $p - value$
   two-sided $p - value$:
   $p - value$
   $= 2 \cdot P(Z > |z_{obs}| \ \ |H_0 \text{ is true})$
   $= 2 \cdot P(Z > 1.89 | H_0 \text{ is true})$
   $= 0.059$



Step 5
Conclusion:
Since the $p - value > 0.05$, we do not reject H_0.
We have sufficient evidence to conclude that the average October temperature is not significant different from 10.

5.1.2.1 Two-sided Z-test in R

Example *Temperature* in R

```
sigma <- 1.5

# Step 1
mu0 <- 10

# Step 2
alpha <- 0.05

# Step 3: Select the test statistic

# Step 4
# Step 4.1: Compute test statistic
n <- length(westT$October)
xmean  <-mean(westT$October)
z  <- (xmean-mu0)/(sigma/sqrt(n))
```

```r
# Step 4.2: Compute p-value
twosided.pvalue  <- 2*(1-pnorm(abs(z)))
result2 <- list(z = z, pvalue = twosided.pvalue)
result2
```

```
## $z
## [1] 1.888889
##
## $pvalue
## [1] 0.05890672
```

Another way:

```r
z.test(westT$October, alternative="two.sided", mu=10, sigma.x=1.5, conf.level=0.95)
```

```
##
##  One-sample z-Test
##
## data:  westT$October
## z = 1.8889, p-value = 0.05891
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.964462 11.924426
## sample estimates:
## mean of x
##  10.94444
```

5.2 Relation confidence interval - hypothesis test

Example ***Temperature***

- In case σ is known, then a 95% CI for the average October temperature in Western Europe is given by $[9.96, 11.9]$.
- Assume we are interested in testing
  $H_0 : \mu = 10$ versus $H_1 : \mu \neq 10$ with a significance level of 0.05.
  With μ the average October temperature in Western Europe.

Consider a hypothesis test on significance level α
$H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

If you construct a $(1 - \alpha) \cdot 100\%$ confidence interval for μ, then:

> If μ_0 does not belong to the CI $\Leftrightarrow$ reject H_0
> If the CI contains $\mu_0 \Leftrightarrow$ do not reject H_0

Example ***Temperature***

In previous chapter, we constructed 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:

```r
conf <- 0.95
sigma <- 1.5
n <- 9
xmean <- mean(westT$October)
alpha <- 1 - conf
lcl <- xmean-qnorm(1-alpha/2)*sigma/sqrt(n)
```

```r
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
```

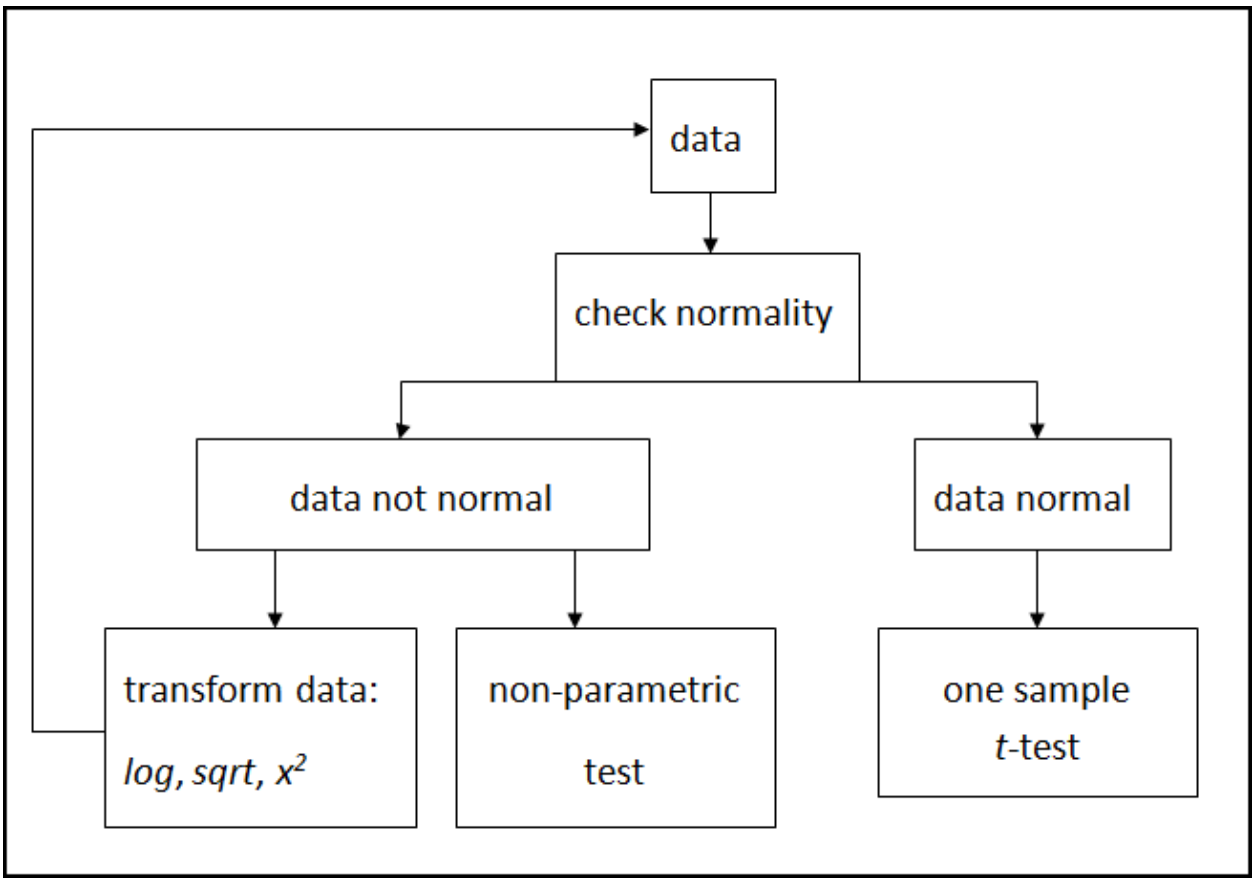
Assume we are interested in testing
$H_0 : \mu = 10$ versus $H_1 : \mu \neq 10$
with significance level of 0.05.

In case σ is known, then a 95% CI for the average October temperature in Western Europe is given by $[9.96; 11.9]$.
This 95% CI contains the value 10, hence $H_0 : \mu = 10$ will not be rejected on a significance level of 0.05.

6 Hypothesis test for the mean when σ is unknown

The scheme of a one sample hypothesis test for the mean when σ is unknown is as follows:

1. When sample size is large enough, we can use the CLT which assures that the average is normally distributed. In this situation, the **one-sample t-test** can be used.
2. When sample size is small, then we have to use the scheme below. Check for normality (e.g. Shapiro-Wilk test). If normality is not rejected, we can use the one sample t-test. When normality is rejected, a **non-parametric alternative or a transformation** can be used.



6.1 The t-test

Example ***Temperature in Western Europe***

Is the average October temperature in Western Europe less than 12°?

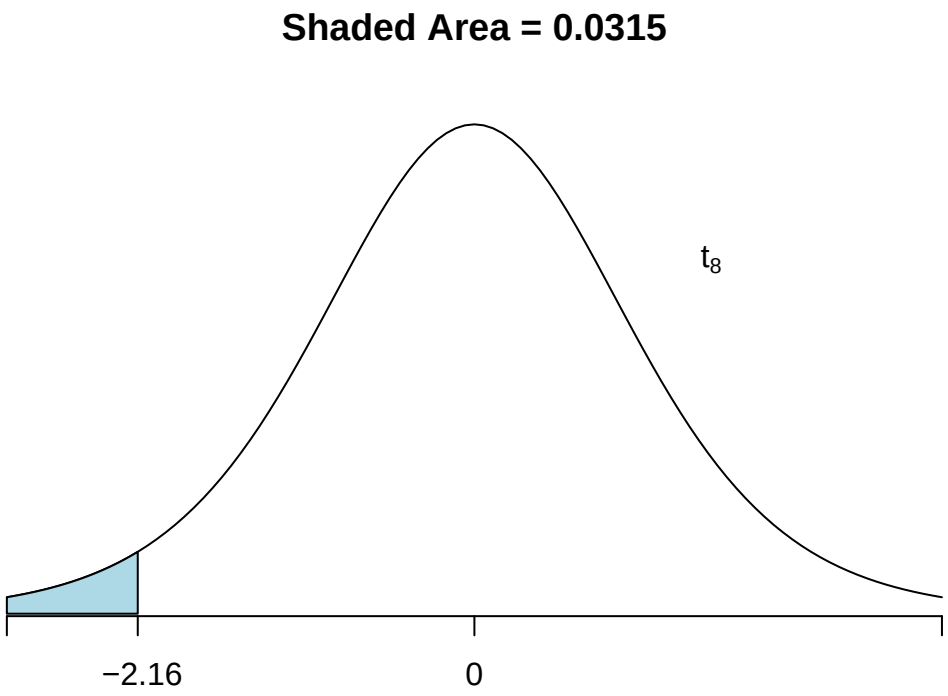
Step 1
The hypothesis is stated as
$H_0 : \mu = 12$ versus $H_1 : \mu < 12$
with μ the average October temperature in Western Europe.

Step 2
Select $\alpha = 0.05$

Step 3
Select the test statistic.
The standard deviation σ is unknown , hence as test statistic we choose $T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$
from which we know the distribution under H_0.

Step 4

1. **Compute** T_{obs}
   $T_{obs} = -2.16$
2. **Compute** $p - value$
   Because we have a one-sided test, we need to compute the one-sided $p - value$: $p - value$
   $= P(T \leq T_{obs} | H_0 \text{ is true})$
   $= P(T \leq -2.16 | H_0 \text{ is true})$
   $= 0.0315$



Step 5
Conclusion:
Since the $p - value < 0.05$, we reject H_0.
We have sufficient evidence to conclude that the average October temperature is significant smaller than 12.

```
t.test(westT$October, mu = 12, conf.level = 0.95, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  westT$October
## t = -2.1608, df = 8, p-value = 0.03136
## alternative hypothesis: true mean is less than 12
## 95 percent confidence interval:
##      -Inf 11.85285
## sample estimates:
## mean of x
##  10.94444
```

6.2 A non-parametric alternative: the sign test

Remember:
The general procedure used for testing a hypothesis assumes that

- Your sample is large enough so that we can use CLT ($n \geq 25$)
- Or (in case you have a limited sample) that your data follows a normal distribution

Example *Temperature in Western Europe*

Here we have a small data set with only 9 observations, hence normality should be checked.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  westT$October
## W = 0.95625, p-value = 0.7584
```

Since $p - value = 0.76 > 0.05$, the null hypothesis of normality is not rejected.

What to do when you have a small data set and **normality does not hold**? Then it is better to **use a non-parametric test**. We will illustrate how the sign test is working with the well-known example of the temperature data (although here it was not necessary to go for a non-parametric test).

Illustration of the sign test: *Temperature in Western Europe*

The **hypothesis** is then reformulated as:
Is the median October temperature less than 12°?

Step 1
The hypothesis is stated as
H_0 : **median** $= 12$ versus H_1 : **median** < 12

Step 2
Select $\alpha = 0.05$

Step 3
Select an appropriate **test statistic**.

The test statistic X is the number of positive differences.

Step 4

1. **Compute the test statistic** X

The sign test starts by computing the vector of differences `October - 12` and keeps the sign of this difference.

City	`westT$October`	Sign of the difference (`October - 12`)
1	11.4	-
2	10.0	-
3	11.1	-
4	12.5	+
5	11.5	-
6	13.5	+
7	9.8	-
8	9.8	-
9	8.9	-

Here almost all differences are negative. The sign test only looks at the sign of the differences. Here, $X = 2$. Under the H_0 that the median is 12, $X \sim B(9, 0.5)$ and the null hypothesis can be formulated as: $H_0 : p = 0.5$ versus $H_1 : p < 0.5$

2. **Compute** $p - value$
   We now can compute the corresponding $p - value$: $P(X \leq 2) = 0.089$.

Step 5
Conclusion:
There seems to be not enough evidence to reject the H_0. Hence, based on this sample, we cannot say that

the median of the October temperature is significant smaller than 12.

In R

Perform a **Binomial Test** by using the function `binom.test` from the package `stats`

```
nr.success <- sum(westT$October - 12 > 0)
binom.test(nr.success, n = 9, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  nr.success and 9
## number of successes = 2, number of trials = 9, p-value = 0.08984
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.5496416
## sample estimates:
## probability of success
##              0.2222222
```

Perform a **sign test** by using the function `SIGN.test` from the package `BSDA`

```
library(BSDA)
SIGN.test(westT$October, md = 12, alternative = "less")
```

```
##
##  One-sample Sign-Test
##
## data:  westT$October
## s = 2, p-value = 0.08984
## alternative hypothesis: true median is less than 12
## 95 percent confidence interval:
##      -Inf 12.06667
## sample estimates:
## median of x
##        11.1
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt  U.E.pt
## Lower Achieved CI     0.9102   -Inf 11.5000
## Interpolated CI       0.9500   -Inf 12.0667
## Upper Achieved CI     0.9805   -Inf 12.5000
```

Remark:
Non-parametric tests are not so powerful as parametric tests.

7 Inference for two independent samples: two-sample t-test

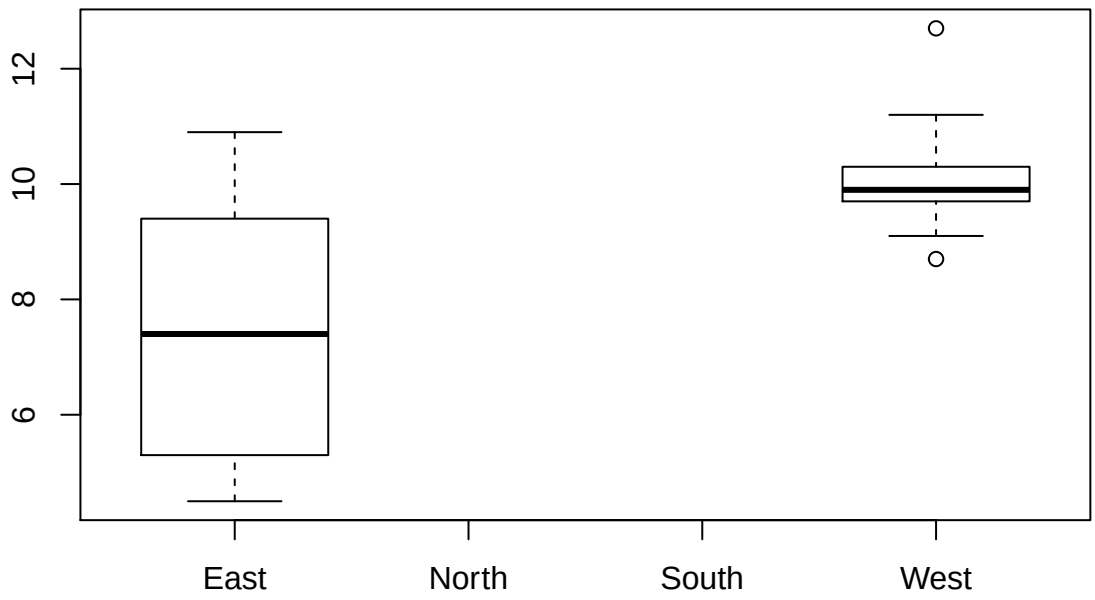
7.1 Hypothesis test: Compare means between two independent groups

Example: ***Temperature***
Compare the average annual temperature between the East-European cities and the West-European cities.

Comparison of samples by descriptive statistics:

```r
westA <- subset(temperature, temperature$Area=='West', select = annual)
eastA <- subset(temperature, temperature$Area=='East', select = annual)
meanW <- mean(westA$annual)
meanE <- mean(eastA$annual)
varW <- var(westA$annual)
varE <- var(eastA$annual)
nW <- length(westA$annual)
nE <- length(eastA$annual)
result <- list(meanW = meanW, meanE = meanE, varW = varW, varE = varE, nW = nW, nE = nE)
result
```

```
## $meanW
## [1] 10.18889
##
## $meanE
## [1] 7.45
##
## $varW
## [1] 1.403611
##
## $varE
## [1] 5.4
##
## $nW
## [1] 9
##
## $nE
## [1] 8
```



Comparison of samples by hypothesis test:

Step 1:
Formulation of H_0 and H_1:
$H_0 : \mu_E = \mu_W$ versus $H_1 : \mu_E \neq \mu_W$
Where μ_E is the average annual temperature for the East-European cities and μ_W is the average annual temperature for the West-European cities.

Step 2
We take significance level $\alpha = 0.05$.

Step 3
Test statistic T

Denote by X the annual temperature for East-European cities and by Y the annual temperature for West-European cities. These two groups are independent.
Assume $X \sim N(\mu_E, \sigma_E^2)$ and $Y \sim N(\mu_W, \sigma_W^2)$.
In general, we do not know σ_E^2 and σ_W^2 and these are estimated by S_E^2 and S_W^2.

- **In case of homogeneity of variances** (when $H_0 : \sigma_E^2 = \sigma_W^2$ is not rejected), then we use as test statistic:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_E - \mu_W)}{S_p \sqrt{\frac{1}{n_E} + \frac{1}{n_W}}} \sim t_{n_E+n_w-2} T \sim t_{n_E+n_w-2} \text{ under } H_0 \text{ and} S_p = \sqrt{\frac{\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2}{n_E + n_W - 2}} \tag{1}$$

- **In case of no homogeneity of variances** (when $H_0 : \sigma_E^2 = \sigma_W^2$ is rejected), then we use as test statistic:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_E - \mu_W)}{\sqrt{\frac{S_E^2}{n_E} + \frac{S_W^2}{n_W}}} \sim t_r T \sim t_r \text{ under } H_0 \text{ and} r = \frac{\left(\frac{S_E^2}{n_E} + \frac{S_W^2}{n_W}\right)^2}{\frac{(S_E^2/n_E)^2}{n_E - 1} + \frac{(S_W^2/n_W)^2}{n_W - 1}} \tag{2}$$

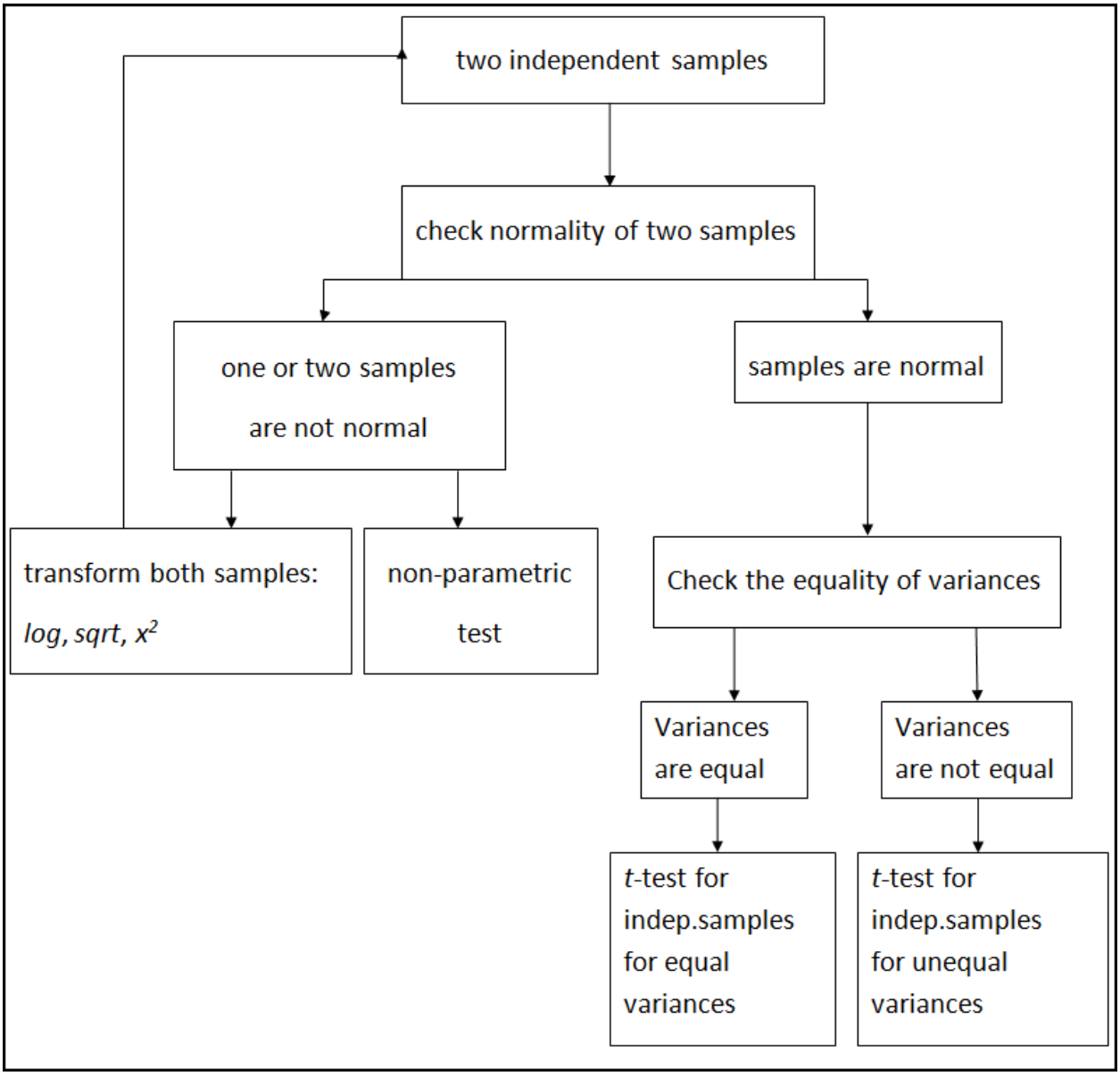
Step 4

1. **Compute the test statistic** T_{obs}

2. **Compute** $p - value$
   Based on T_{obs}, which follows a t-distribution under H_0, we compute the corresponding $p - value$

Step 5
Based on $p - value$, we formulate our conclusion.

7.2 Overview



7.3 In R

Example: *Temperature*

7.3.1 Checking the assumptions

1. **Test for normality** in both groups (see later)
   H_0 : **data are normally distributed**
   versus
   H_1 : **data are not normally distributed**

```
shapiro.test(westA$annual)
```

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

```
##
## data:  westA$annual
## W = 0.91314, p-value = 0.3385
```

Because $p - value = 0.34 > 0.05$, we do not reject H_0.

```
shapiro.test(eastA$annual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  eastA$annual
## W = 0.94527, p-value = 0.6635
```

Because $p - value = 0.66 > 0.05$, we do not reject H_0.

2. **Check homogeneity of variances** (see later)
   $H_0 : \sigma^2_{East} = \sigma^2_{West}$
   versus
   $H_1 : \sigma^2_{East} \neq \sigma^2_{West}$

```
var.test(westA$annual, eastA$annual)
```

```
##
##  F test to compare two variances
##
## data:  westA$annual and eastA$annual
## F = 0.25993, num df = 8, denom df = 7, p-value = 0.07814
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.05305367 1.17710003
## sample estimates:
## ratio of variances
##           0.259928
```

Since $p - value = 0.078 > 0.005$, H_0 is not rejected.

7.3.2 t-test

Using t-test for equal variances to compare the average in the two independent groups
$H_0 : \mu_E = \mu_W$ versus $H_1 : \mu_E \neq \mu_W$

```
t.test(westA$annual, eastA$annual, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  westA$annual and eastA$annual
## t = 3.1177, df = 15, p-value = 0.007057
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8664241 4.6113537
## sample estimates:
## mean of x mean of y
##  10.18889   7.45000
```

Remark:

- **In case the variances are not equal**, apply the t-test for unequal variances to compare the averages in the two independent groups. (Although the variances are homogeneous here, we only illustrate how

it works).
$H_0 : \mu_E = \mu_W$ versus $H_1 : \mu_E \neq \mu_W$

```
t.test(westA$annual, eastA$annual, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  westA$annual and eastA$annual
## t = 3.0046, df = 10.135, p-value = 0.01305
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7114477 4.7663301
## sample estimates:
## mean of x mean of y
##  10.18889   7.45000
```

- What if the **data do not follow a normal distribution**?
  - If n_1 is large enough (> 25), then the CLT can be used for group 1.
    If n_2 is large enough (> 25), then the CLT can be used for group 2.
  - When the normality assumption is not satisfied, then transformation of the data or the use of a non-parametric test can be used.

7.4 F-test for testing equality of variances

We have to **test whether the variances in both populations are the same**.

$H_0 : \sigma_E^2 = \sigma_W^2$ versus $H_1 : \sigma_E^2 \neq \sigma_W^2$
Or
$H_0 : \frac{\sigma_E^2}{\sigma_W^2} = 1$ versus $H_1 : \frac{\sigma_E^2}{\sigma_W^2} \neq 1$

Therefore, we use the **F test**. This test makes use of the sample variances.

The test statistic is $F = \frac{S_E^2}{S_W^2}$ which has an F-distribution with $n_E - 1$ and $n_W - 1$ degrees of freedom.

In R:

```
var.test(westA$annual, eastA$annual)
```

```
##
##  F test to compare two variances
##
## data:  westA$annual and eastA$annual
## F = 0.25993, num df = 8, denom df = 7, p-value = 0.07814
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.05305367 1.17710003
## sample estimates:
## ratio of variances
##           0.259928
```

7.5 Test for normality

The Shapiro-Wilk test is a normality test. The **Shapiro-Wilk test** is based on the degree of linearity is a QQ plot.

We want to test the hypothesis that the annual temperature (for West-European and East-European cities separately) are normally distributed.

H_0 : **data are normally distributed**
versus
H_1 : **data are not normally distributed**

In R:

```
shapiro.test(westA$annual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  westA$annual
## W = 0.91314, p-value = 0.3385
```

```
shapiro.test(eastA$annual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  eastA$annual
## W = 0.94527, p-value = 0.6635
```

Since the $p-value$ is larger than 0.05, we cannot reject the H_0. We may assume that the annual temperature data for West-European cities follow a normal distribution.
The same can be said for East-European cities.

8 Distribution free test: Wilcoxon rank sum test

8.1 Introduction

Construction of confidence intervals and hypothesis testing all assume that the data are sampled from a normal population (or that the sample size is large enough so that the CLT can be used). Therefore, these methods are **parametric statistical tests**.
When the data are normal, these tests are most powerful.

In case the data are not normal, **non-parametric tests** have to be used. Non-parametric tests do not depend on the distribution of the sampled population. Thus, they are called **distribution-free tests**.

In case the data are non-normal, or the data are ranked, the non-parametric tests are more powerful than their corresponding parametric counterparts.

	t-test for one or two means	Non-parametric alternative
1 group	*One sample t-test*	*Sign test*
Two independent groups	*t-test* for comparing means - equal variances - unequal variances	*Wilcoxon rank sum test* for testing same distributions in two independent groups
Two dependent groups	*Paired t-test*	*Signed test*
More independent groups	*Anova*	*Kruskal Wallis test* for testing same distributions in k (k>2) independent groups

8.2 Wilcoxon rank sum test

In the previous topic, we were interested to know whether there was a significant difference between the average annual temperature in Western- and Eastern-European cities.

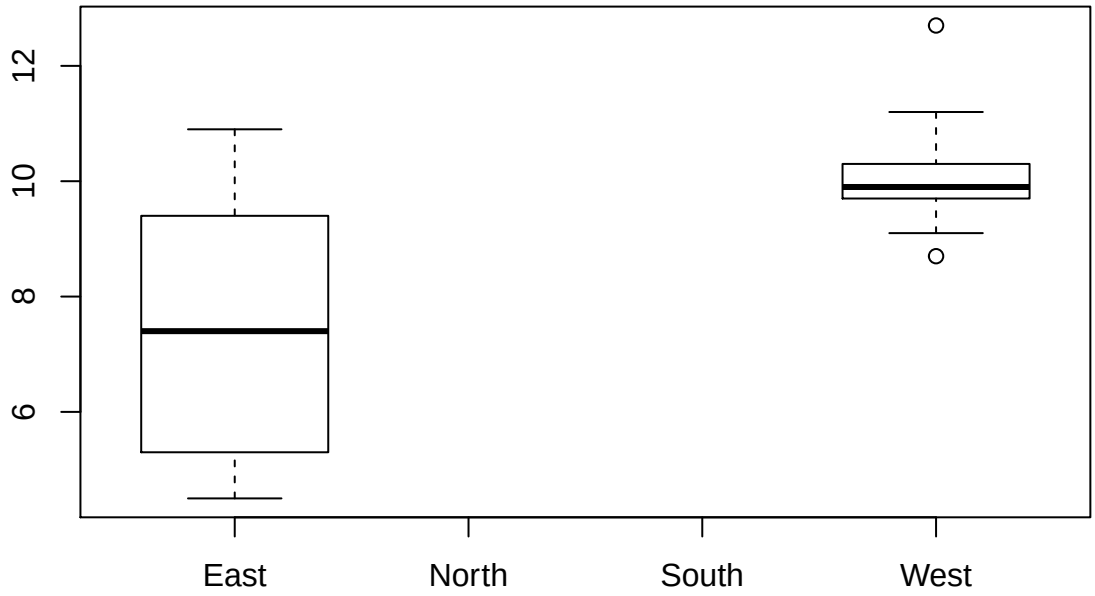
In case *data are normally distributed*, the hypothesis $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ is tested by a t-test (assuming equal or unequal variances). (see previous pages)

In case *data are not normally distributed*, we can use the non-parametric **Wilcoxon rank sum test**.

Example: ***Temperature***
We want to compare the annual temperature in East- and West-European cities. (Although normality is satisfied here, we use the same data for illustrative purposes).

We will use here the Wilcoxon rank sum test to compare the distributions of the annual temperature between two regions of Europe.

Step 1:
Formulation of H_0 and H_1:
H_0 : **The center value of D_0 and D_1 are identical**
versus
H_1 : **The center value of D_0 and D_1 are shifted**
And here, D_0 is the distribution of the Eastern European annual temperature. D_1 is the distribution of the Western European annual temperature.



Step 2
We take significance level $\alpha = 0.05$.

Step 3

- If we have small samples, then we can calculate the $p - value$ based on the exact distribution of T_{East}
- In the opposite case, if we have large samples, then we will derive a new statistic based on T_{East}, which will be approximately normal distributed.

Step 4

Compute the test statistic and compute $p - value$

First steps in computing the test statistic and computing the $p - value$ are:

1. **Order all annual temperatures** from small to large

```
westeastA <- subset(temperature, temperature$Area == "West" | temperature$Area == "East",
                    select = c(City, annual, Area))
library(doBy)
sortannual <- orderBy( ~ annual, data = westeastA)
sortannual
```

```
##             City annual Area
## 34 St_Petersburg    4.5 East
## 15        Moscow    5.1 East
## 14         Minsk    5.5 East
## 9           Kiev    7.1 East
## 10        Krakow    7.7 East
## 35        Zurich    8.7 West
## 3         Berlin    9.1 West
## 18        Prague    9.2 East
## 22         Sofia    9.6 East
## 29        Geneva    9.7 West
## 28     Frankfurt    9.8 West
## 1      Amsterdam    9.9 West
## 4       Brussels   10.3 West
## 24       Antwerp   10.3 West
## 5       Budapest   10.9 East
## 17         Paris   11.2 West
## 26      Bordeaux   12.7 West
```

2. **Rank these values**
   Give value 1 for the smallest, 17 for the largest. Tied observations are assigned ranks equal to the average of the ranks of the tied observations (e.g., Antwerp and Brussels receive the same average rank 13.5).

3. **Compute the sum of the ranks**
   For cities in Eastern Europe: 47
   For cities in Western Europe: 106
   Reasoning: If H_0 is true, then we would expect that the average rank sums of T_{East} and T_{West} are nearly equal. (With T_{West} and T_{East} respectively the Western and Eastern annual temperature.)

These steps demonstrate the idea behind the Wilcoxon test statistic.

In R

```
wilcox.test(westA$annual, eastA$annual)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  westA$annual and eastA$annual
## W = 61, p-value = 0.01833
## alternative hypothesis: true location shift is not equal to 0
```

Step 5
$p - value = 0.018 < 0.05$. Hence, H_0 of equal center values is rejected.

9 Test for paired data

On many occasions, the two samples are not independent because they involve the same sampling unit. Measuring each subject or entity twice results in pairs of observations. In such situations, tests for paired data are used to perform hypothesis tests.

Illustrative example: *diet*

If we want to know whether a diet is successful, we can select a number of persons and measure their weight before and after the diet. These two measurements are **paired**.

Person	Weight before	Weight after	Difference: (weight before) - (weight after)
1			
2			
3			
4			

The hypothesis to be tested:
$H_0 : \mu_{before} = \mu_{after}$ versus $H_1 : \mu_{before} > \mu_{after}$
Or
$H_0 : \mu_{before} - \mu_{after} = 0$ versus $H_1 : \mu_{before} - \mu_{after} > 0$
with μ the average weight.

To analyze the data, we first compute the differences:
$D_i = Y_{1,i} - Y_{2,i}$
where $Y_{1,i}$ is the weight before and $Y_{2,i}$ is the weight after the diet.
We assume $D_i \sim N(\delta, \sigma^2)$ and that they are i.i.d.

Let δ be the population difference in weight, then we can reformulate this hypothesis as a one-sample problem.
$H_0 : \delta = 0$ versus $H_1 : \delta > 0$

$\rightarrow$ This is an illustrative example of a case where a test for paired data is used to test this hypothesis.

9.1 Paired t-test

The **paired t-test** is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired t-test, each subject or entity is measured twice, resulting in pairs of observations.[1]

Step 1:
To check whether there is a significant difference between the mean of two sets of observations, we test
$H_0 : \delta = 0 (= \delta_0)$ versus $H_1 : \delta > 0$
with δ the mean difference.

Step 2
We take significance level $\alpha = 0.05$.

Step 3
Test statistic:

$$T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{n}} \text{with } \overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} \text{ and } S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} \tag{3}$$

[1] From https://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/

The test statistic T has a t-distribution with $n-1$ degrees of freedom under H_0.

Step 4
Compute the test statistic T_{obs} and compute the one sided $p-value$.

Step 5
The conclusion is based on the $p-value$.

9.1.1 Example

Example *Decathlon*
The data set for this example is *combine_decathlon.txt*. This data set contains 9 athletes who participated in both the Olympic games and the Decastar competition.

All these athletes who participated in both competitions certainly focused their physical preparation on their performances at the Olympic Games. Did they really have a better performance in the Olympic games compared to the Decastar competition?

Paired t-test
Hypothesis to be tested:
$H_0 : \mu_{decastar} = \mu_{olympic}$ versus $H_0 : \mu_{decastar} > \mu_{olympic}$
with μ the average time needed to run 1500 m.

```
summary(combine_decathlon)
```

```
##      namesD  decastar_1500    olympic_1500
##  BARRAS :1   Min.   :266.6   Min.   :264.4
##  BERNARD:1   1st Qu.:278.1   1st Qu.:269.5
##  CLAY   :1   Median :282.0   Median :276.3
##  HERNU  :1   Mean   :283.7   Mean   :274.6
##  KARPOV :1   3rd Qu.:291.7   3rd Qu.:278.1
##  NOOL   :1   Max.   :301.5   Max.   :282.0
##  (Other):3
```

```
olympic <- combine_decathlon[,3]
decastar <- combine_decathlon[,2]
t.test(decastar, olympic, paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  decastar and olympic
## t = 2.3934, df = 8, p-value = 0.02181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.020124      Inf
## sample estimates:
## mean of the differences
##                9.056667
```

The one-sided $p-value$ is 0.022. We have to reject the H_0. We have sufficient evidence to conclude that the performance of the athletes was significant better at the Olympic Games compared to the Decastar event.

Remark:
This paired t-test relies on the **assumption that the differences are normally distributed**.

```
difference <- decastar - olympic
shapiro.test(difference)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference
## W = 0.92444, p-value = 0.4302
```

- In case the sample size is large enough, the CLT can be used.
- In case the sample size is limited and normality does not hold, we can use non-parametric alternative.

9.2 Distribution-free alternative: Sign test

The **sign test** is an non-parametric alternative to the paired t-test.

9.2.1 Example

Example *Decathlon*
Although normality can be assumed for the data *combine_decathlon.txt*, we illustrate the use of the sign test on this data set.

Here, we use a very small data set with only 6 athletes

```
small_decathlon <- combine_decathlon[1:6,]
small_decathlon
```

```
##    namesD decastar_1500 olympic_1500
## 1  BARRAS         282.0       267.09
## 2 BERNARD         280.1       276.31
## 3    CLAY         301.5       282.00
## 4   HERNU         285.1       264.35
## 5  KARPOV         300.2       278.11
## 6    NOOL         266.6       276.33
```

Step 1:
Formulation of H_0 and H_1:
$H_0 : median_{decastar} - median_{olympic} = 0$
versus
$H_0 : median_{decastar} - median_{olympic} > 0$

Under the H_0 that there is no Olympic Games effect, $X \sim B(6, 0.5)$ and the null hypothesis can be formulated as:
$H_0 : p = 0.5$ versus $H_1 : p > 0.5$
with p the probability that the obtained difference is positive (or Decastar time > Olympic time)

Step 2
We take significance level $\alpha = 0.05$.

Step 3
The sign test starts by computing the vector of differences. The sign test only looks at the sign of the differences.

```
small_decathlon$difference <- sign(small_decathlon$decastar_1500 - small_decathlon$olympic_1500)
small_decathlon
```

```
##    namesD decastar_1500 olympic_1500 difference
## 1  BARRAS         282.0       267.09          1
## 2 BERNARD         280.1       276.31          1
## 3    CLAY         301.5       282.00          1
## 4   HERNU         285.1       264.35          1
## 5  KARPOV         300.2       278.11          1
```

```
## 6    NOOL         266.6       276.33         -1
```

Here, almost all differences are positive. The test statistic X is the number of positive differences.

Step 4

Compute the test statistic and compute $p - value$

Here, $X = 5$. Now, we can compute the $p - value$: $P(X \geq 5) = 0.1094$

Step 5

There seems to be not enough evidence to reject the H_0. Hence, based on this small sample, we cannot say that there is an Olympic Games effect.

In R

Use of `binom.test` (package `stats`)

```
sum.pos <- sum(small_decathlon$difference>0)
sum.pos
```

```
## [1] 5
```

```
binom.test(sum.pos, 6, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  sum.pos and 6
## number of successes = 5, number of trials = 6, p-value = 0.1094
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4181966 1.0000000
## sample estimates:
## probability of success
##              0.8333333
```

Use of the sign test in R (package `BSDA`)

```
SIGN.test(small_decathlon$decastar_1500, small_decathlon$olympic_1500,
          alternative = "greater")
```

```
##
##  Dependent-samples Sign-Test
##
## data:  small_decathlon$decastar_1500 and small_decathlon$olympic_1500
## S = 5, p-value = 0.1094
## alternative hypothesis: true median difference is greater than 0
## 95 percent confidence interval:
##  -4.772667       Inf
## sample estimates:
## median of x-y
##        17.205
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.8906  3.7900    Inf
## Interpolated CI       0.9500 -4.7727    Inf
## Upper Achieved CI     0.9844 -9.7300    Inf
```

10 Test for proportions

10.1 Hypothesis test for one proportion

Hypothesis statement:
$H_0 : p = p_0$ versus $H_1 : p \neq p_0$

X is the number of successes in n trials.
$X \sim B(n, p)$

Let $\hat{p} = \frac{X}{n}$, the observed proportion of successes.

10.1.1 Normal approximation

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
$\frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$

When $np_0 \geq 5$, $n(1-p_0) \geq 5$ and H_0 true, then we have
$\hat{p} \sim \left(p_0, \frac{p_0(1-p_0)}{n}\right)$

Then we can use as **test statistic:**
$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$ under H_0

The two-sided $p - value$ is $2 \cdot P(Z > |Z_{obs}|)$

Example: ***Temperature***
Someone states that half of the cities in Europe have an annual temperature above 10°. Verify this statement.

Let p be the proportion of European cities with annual temperature $> 10°$.

$H_0 : p = 0.5$ versus $H_1 : p \neq 0.5$
$\alpha = 0.05$

We consider X **= number of cities with an annual temperature above 10° from a group of 35**

Based on our sample, how may cities have an annual temperature above 10°?

```
n <- length(temperature$annual)
n
```

```
## [1] 35
```

```
number <- sum(temperature$annual>10)
number
```

```
## [1] 14
```

From this group of 35 cities, we observe 14 cities with an annual temperature above 10°. This is an observed proportion of 0.4.

In R
In R, the traditional Z test is not implemented, but the `prop.test` can be used which gives similar results

```
prop.test(x = 14, n = 35, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  14 out of 35, null probability 0.5
```

```
## X-squared = 1.0286, df = 1, p-value = 0.3105
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2435206 0.5779096
## sample estimates:
##   p
## 0.4
```

Two-sided $p - value = 0.31 > 0.05$ hence we do not reject the H_0. The proportion of European cities with annual temperature $> 10°$ is not significant different from 0.5.

10.1.2 Normal approximation cannot be used

The exact method used in case the normal approximation cannot be used corresponds to the binomial test.

$X \sim B(n, p)$ with X the number of successes in n trials.
$\hat{p} = \frac{X}{n}$

When the normal approximation cannot be used, a binomial test is used to evaluate the hypothesis for one proportions.

$H_0 : p = p_0$

H_1	$p - value$	Computation $p - value$
$p \neq p_0$	Two-sided	$2 \cdot P(X \leq x)$ if $\hat{p} \leq p_0$; Or $2 \cdot P(X \geq x)$ if $\hat{p} \geq p_0$
$p > p_0$	Right one-sided	$P(X \geq x)$
$p < p_0$	Left one-sided	$P(X \leq x)$

The $p - value$ is the sum of probabilities of all events that are as extreme or more extreme than the observed sample result.

Example: *Temperature*
Someone states that half of the cities in Europe have an annual temperature above 10°. Verify this statement without using the normal approximation.

Let p be the proportion of European cities with annual temperature $> 10°$.

$H_0 : p = 0.5$ versus $H_1 : p \neq 0.5$
$\alpha = 0.05$

$X \sim B(n = 35, p = 0.5)$.
What is the probability that $X = 14$ or more extreme (in the sense of the alternative)?
Since $\hat{p} = 0.4 (< 0.5)$ we compute the two sided $p - value$ as
$p - value = 2 \cdot P(X \leq 14 | H_0 true) = 2 \cdot 0.1553 = 0.3105$

```
binom.test(x = 14, n = 35, p = 0.5, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  14 and 35
## number of successes = 14, number of trials = 35, p-value = 0.3105
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2387081 0.5788823
## sample estimates:
## probability of success
```

```
##                   0.4
```

Two-sided $p - value = 0.31 > 0.05$ hence we do not reject the H_0.

We cannot reject the statement that half of the cities in Europe have an annual temperature above 10°.

10.2 Hypothesis test for two proportions

Example: ***Temperature***
Compare the proportion of cities in Northern and Western Europe which have an annual temperature of at least 9° with the proportion of cities in Southern and Eastern Europe which have an annual temperature of at least 9°.

Let p_{NW} = the proportion of Northern and Western European cities with annual temperature $\geq 9°$.

Let p_{SE} = the proportion of Southern and Eastern European cities with annual temperature $\geq 9°$.

$H_0 : p_{NW} = p_{SE}$ versus $H_1 : p_{NW} \neq p_{SE}$
$\alpha = 0.05$

Based on our sample, what are the observed proportions?

```
NW_area <- subset(temperature, temperature$Area=='North' | temperature$Area=='West',
                  select = annual)
SE_area <- subset(temperature, temperature$Area=='South' | temperature$Area=='East',
                  select = annual)
n_NW <- length(NW_area$annual)
n_SE <- length(SE_area$annual)
number_NW <- sum(NW_area >= 9)
number_SE <- sum(SE_area >= 9)
list(n_NW=n_NW, above9_NW=number_NW, n_SE=n_SE, above9_SE= number_SE,
     prop_NW=number_NW/n_NW, prop_SE=number_SE/n_SE)
```

```
## $n_NW
## [1] 17
##
## $above9_NW
## [1] 10
##
## $n_SE
## [1] 18
##
## $above9_SE
## [1] 13
##
## $prop_NW
## [1] 0.5882353
##
## $prop_SE
## [1] 0.7222222
```

In general: Two sample z-test for comparing proportions from two independent groups
A hypothesis test for two proportions is used to determine whether the difference between two proportions is significant.

The hypothesis evaluated is
$H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$

If $n_1p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2p_2 \geq 5$ and $n_2(1-p_2) \geq 5$ then $Z \sim N(0,1)$. Then, a **two-sample z-test for comparing proportions from two independent groups** can be used and the corresponding $p-value$ can be computed.

The used test statistic Z:
$Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
where $\hat{p} = \frac{n_1 \cdot \hat{p}_1 + n_2 \cdot \hat{p}_2}{n_1+n_2}$ and $\hat{q} = 1 - \hat{p}$

Remark:
When the condition $n_1p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2p_2 \geq 5$ and $n_2(1-p_2) \geq 5$ does not hold, it is better to use the **Fisher exact test**. We do not cover this topic here.

In R, the traditional z-test is not implemented, but the `prop.test` can be used.

```
x <- c(number_NW, number_SE)
n <- c(n_NW, n_SE)
prop.test(x, n)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 0.22886, df = 1, p-value = 0.6324
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.5035022  0.2355284
## sample estimates:
##    prop 1    prop 2
## 0.5882353 0.7222222
```

The two-sided $p-value$ is 0.63 hence we do not reject the H_0.
There is no significant difference between the proportion of cities in Northern and Western Europe which have an annual temperature of at least 9° and with the proportion of cities in Southern and Eastern Europe which have an annual temperature of at least 9°.

11 Chi-square goodness of fit test

The goal of this technique is to determine the fit of the distribution function.

Example: ***Birthday***
In a group of 48 students, the distribution of the birthdays is as given below:

Spring	Summer	Fall	Winter	*Total*
10	12	10	16	48

Are the birthdays uniformly distributed over the seasons?

$H_0 : p_1 = p_2 = p_3 = p_4 (= \frac{1}{4})$
versus
H_1 : **There is no uniform distribution**

In general:

The χ^2 **test** is a goodness of fit test to check the fit of observed data with any kind of distribution. The goal of this technique is to determine the fit of the distribution function.

The **hypotheses** are:
H_0 : **The data follow a specific distribution F^*.**
versus
H_1 : **The data does not follow that specific distribution F^*.**

The null hypothesis states that there is no significant difference between the measured and expected frequencies.

Class	1	2	...	k
Observed frequency	O_1	O_2	...	O_k
Probability distribution from F^*	p_1	p_2	...	p_k
Expected frequency	E_1	E_2	...	E_k

with $E_i = n \cdot p_i$ for all $i = 1, 2, ..., k$.
Class refers to the different levels of a categorical variable.

The chi-square test is based on a comparison of the observed frequencies O_i with the corresponding expected frequencies E_i under H_0.

The **test statistic** is defined as
$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}\Big|_{H_0} \approx \chi^2_{k-1}$

The value of χ^2 follows a chi-square distribution, with $k-1$ degrees of freedom. (Assume that there are k classes).

There are two **assumptions** for this test:

1. The sample is a random sample from the population
2. The sample size is reasonably large (expected number $E_i \geq 5$)

Example: ***Birthday***
In a group of 48 students, the distribution of the birthdays is as given below:

Spring	Summer	Fall	Winter	*Total*
10	12	10	16	48

Are the birthdays uniformly distributed over the seasons?

$H_0 : p_1 = p_2 = p_3 = p_4 (= \frac{1}{4})$
versus
H_1 : **There is no uniform distribution**

Test statistic: Pearson chi-square
$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}\Big|_{H_0} \approx \chi^2_{k-1}$
with $k - 1 = 3$ degrees of freedom.

Compute the observed values:

Class	Season	O	E	Cell deviation $O - E$	Cell chi-square $\frac{(O-E)^2}{E}$
1	Spring	10	12	-2	0.333
2	Summer	12	12	0	0.000

Class	Season	O	E	Cell deviation $O-E$	Cell chi-square $\frac{(O-E)^2}{E}$
3	Fall	10	12	-2	0.333
4	Winter	16	12	4	1.333
SUM		48	48	0	$\chi^2 = 1.999$

Observed chi-square: $\chi^2_{obs} = 1.999$ p-value: $p - value = P(\chi^2_3 \geq \chi^2_{obs}) = P(\chi^2_3 \geq 1.999) = 0.5726 > 0.05$

Conclusion: We do not reject the H_0. Birthdays are uniformly distributed over the several seasons.

In R

```
birthdays <- c(10, 12, 10 ,16)
predprob <- c(0.25, 0.25, 0.25, 0.25)
chisq.test(birthdays, p = predprob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  birthdays
## X-squared = 2, df = 3, p-value = 0.5724
```

12 Power and sample size (in case σ is known)

12.1 Power of the one-sample t-test

We have seen before:

		Decision of hypothesis test	
		Reject H_0	**Do not reject H_0**
Population	**H_0 true**	**Type I error** **P[reject H_0|H_0 true] = α**	**OK** P[not reject H_0|H_0 true] = 1- α
	H_1 true	**OK** P[reject H_0|H_1 true] = 1-β = power	**Type II error** P[not reject H_0|H_1 true] = β

The **power** of a test is the probability of making a correct decision (to reject H_0) when H_0 is false.

Power = P(reject $H_0|H_1$ is true)

In order to be able to determine the power of a test we have to determine

- Significance level $\alpha = 0.05$
- An idea of the variation in the population σ
- Specific alternative hypothesis H_1 and a specific alternative value μ_1

Example: ***Temperature***
We have a sample of 9 West-European cities from which we know the average October temperature. We assume standard deviation to be known and to be equal to 1.5. We test whether the average October temperature is different from 10°.

$H_0 : \mu = 10$ versus $H_1 : \mu \neq 10$

1. What is the power of the test for the following situation?
Assume that the real underlying average October temperature is 11°. How often will this test detect this difference (and hence reject H_0)?

In order to compute the power, we need:

- $\delta = |\mu_1 - \mu_0| = 1$
- Standard deviation (known): $\sigma = 1.5$
- Sample size $n = 9$
- Significance level: $\alpha = 0.05$
- Alternative: Two-sided (because the original test was two-sided)

```
westT <- subset(temperature, temperature$Area=='West', select=October)
n <- length(westT$October)
delta <- 1
sd <- 1.5 # sd is assumed to be known
sig.level <- 0.05
power.t.test(n = n, delta = delta, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 1
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.4209651
##     alternative = two.sided
```

We only have a power of 42% which is much too low! This means that the H_0 will be rejected in favor of the H_1 in 42% of the times when the true underlying average October temperature is 11°C.

2. What is the power of the test for the following situation?
Assume that the real underlying average October temperature is 9°. How often will this test detect this difference (and hence reject H_0)?

In order to compute the power, we need:

- $\delta = |\mu_1 - \mu_0| = 1$
- Standard deviation (known): $\sigma = 1.5$
- Sample size $n = 9$
- Significance level: $\alpha = 0.05$
- Alternative: Two-sided (because the original test was two-sided)

$Power = 0.42$ because the delta value δ is exactly the same for this two sided test.

3. Let's vary the values of the alternative hypothesis.
What is the result for the power? Compute the power when we let δ increase

```
delta2 <- seq(from = 1, to = 2.5, by = 0.25)
power.t.test(n =n, delta = delta2, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.4209651, 0.5930652, 0.7480155, 0.8641356, 0.9367429, 0.9747105, 0.9913504
##     alternative = two.sided
```

When you want to detect larger differences in the means, then the power of your test will increase.

Remark:
If the alternative mean is shifted farther away from μ_0, ($|\mu_1 - \mu_0|$ increases), then the power increases!

4. What is the effect of a change in standard deviation?
Assume that the real underlying average October temperature is $11°C$. How often will this test detect this difference (and hence reject the H_0)? We assume here standard deviation $\sigma = 1$.

In order to compute the power, we need:

- $\delta = |\mu_1 - \mu_0| = 1$
- Standard deviation (known): $\sigma = 1$
- Sample size $n = 9$
- Significance level: $\alpha = 0.05$
- Alternative: Two-sided (because the original test was two-sided)

```
sd <- 1
power.t.test(n =n, delta = delta, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.7480155
##     alternative = two.sided
```

What if the standard deviation would be $\sigma = 2$?

```
sd <- 2
power.t.test(n =n, delta = delta, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 1
##              sd = 2
##       sig.level = 0.05
```

```
##           power = 0.2622673
##     alternative = two.sided
```

Remark:
If the standard deviation of the distribution of the individual observations increases, (σ increases), then the power decreases!

5. What is the effect of the sample size? Assume that the real underlying average October temperature is $11°C$. How often will this test detect this difference (and hence reject the H_0)? We assume here standard deviation $\sigma = 1.5$ and a sample of size 20 (i.e., $n = 20$).
To evaluate the effect of the sample size, we consider this situation multiple times with a different sample size.

In order to compute the power, we need:

- $\delta = |\mu_1 - \mu_0| = 1$
- Standard deviation (known): $\sigma = 1.5$
- Sample size n ranges from 10 to 40
- Significance level: $\alpha = 0.05$
- Alternative: Two-sided (because the original test was two-sided)

```
sd <- 1.5
n <- seq(from = 10, to = 40, by = 5)
power.t.test(n =n, delta = delta, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 10, 15, 20, 25, 30, 35, 40
##           delta = 1
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.4691805, 0.6708562, 0.8072909, 0.8920169, 0.9415758, 0.9692876, 0.9842420
##     alternative = two.sided
```

Remark:
If the sample size increases (n increases), then the power increases.

12.2 Sample size computation

For a situation when the real underlying average October temperature is $11°C$. Assume $\sigma = 1.5$. What is the sample size needed in order to obtain a power of 0.8?

- $\delta = |\mu_1 - \mu_0| = 1$
- Standard deviation (known): $\sigma = 1.5$
- Sample size n **is unknown**
- Significance level: $\alpha = 0.05$
- Alternative: Two-sided (because the original test was two-sided)
- **Power** $= 0.8$

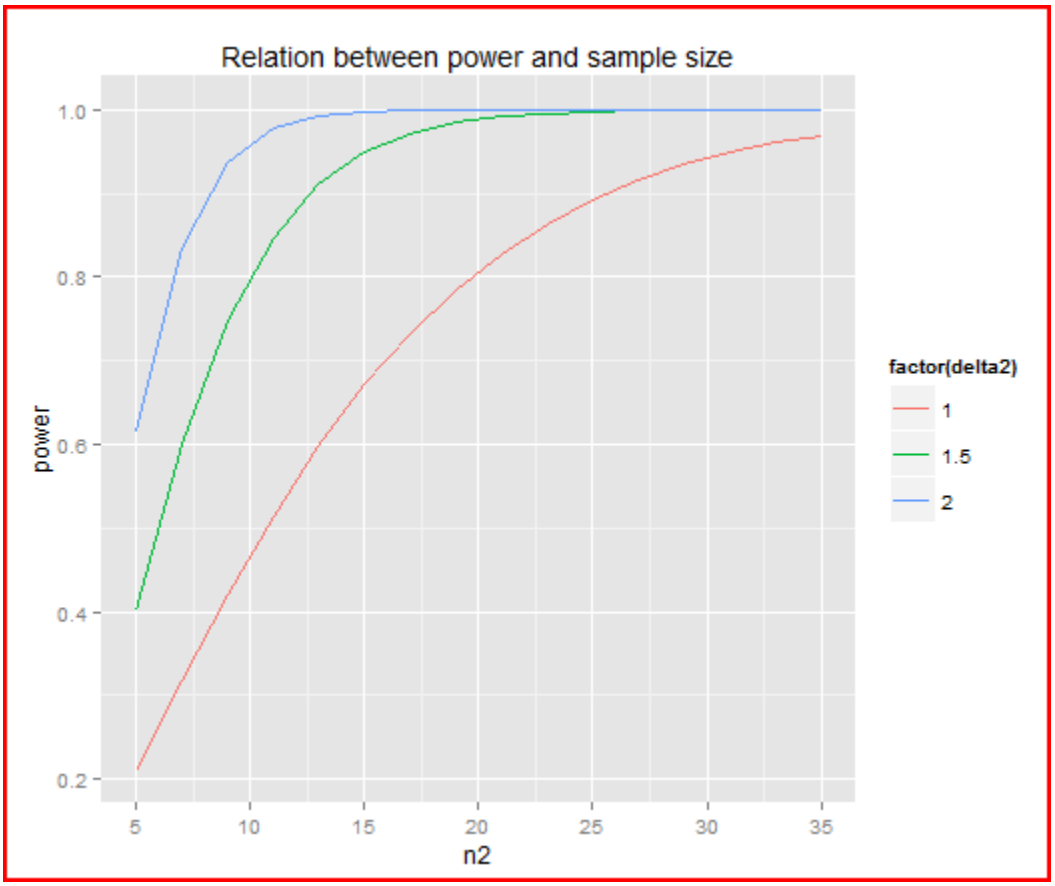
```
delta <- 1
sd <- 1.5
sig.level <- 0.05
power.t.test(n = NULL, delta = delta, sd = sd, sig.level = sig.level, power = 0.8,
             type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
```

```
##               n = 19.66697
##           delta = 1
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

We need a sample of at least 20 cities.

12.3 Relationship between power and sample size



See also: https://www.youtube.com/watch?v=kMYxd6QeAss (only first part of the movie)

Remark:

1. This `power.t.test` function can also be used to compute power (and sample sizes) for two-sided t-tests and paired t-test.
2. If you want to compute power (and sample sizes) for proportion, then use the `power.prop.test` in R.

To compute the power for the following situation:
$H_0 : p_1 = p_2$ versus $H_1 : p1 \neq p2$
with $\alpha = 0.05$
with $p_1 = 0.4$, $p_2 = 0.1$, $n_1 = 9$ and $n_2 = 8$.

`power.prob.test` assumes *equal sample sizes*. In case this is not so, we take n as $min(n_1, n_2)$.

```
power.prop.test(n = 8, p1 = 0.4, p2 = 0.1, power = NULL, alternative = "two.sided")
```

```
##
```

```
##      Two-sample comparison of proportions power calculation
##
##               n = 8
##              p1 = 0.4
##              p2 = 0.1
##       sig.level = 0.05
##           power = 0.2701926
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Example: ***Gene expression***

The expression level of a gene measured with a given technology is known to have a mean expression level of 1.5 in the normal human population. A researcher measures the expression level of this gene and has obtained the following values: $1.9; 2.5; 1.3; 2.1; 1.5; 2.7; 1.7; 1.2$ and 2.0. **We expect the gene to be up-regulated in the condition under study.**

It is known from the literature that the mean expression level of the give gene, measures with the same technology is $\mu = 1.5$ and the standard deviation is $\sigma = 0.44$.

$H_0 : \mu = 1.5$ versus $H_1 : \mu > 1.5$

We expect the expression level of this gene will have an increase of 20%.

Take $\alpha = 0.05$.
Furthermore, $\mu_1 = 1.5 + 0.2 \cdot 1.5 = 1.8$. The difference between the μ from H_0 and the μ from H_1 is hence 0.3.

1. What is the power of this test?
Assume that the real underlying population has a mean expression value of 1.8. How often will this test detect this difference (and hence reject the H_0)?

power calculation:

```
n <- 9
delta <- 0.3
sd <- 0.44
sig.level <- 0.05
power.t.test(n = n, delta = delta, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 0.3
##              sd = 0.44
##       sig.level = 0.05
##           power = 0.5886434
##     alternative = one.sided
```

The power of the test is only 0.59. This means that the H_0 will be rejected in favor of the H_1 in 58% of the times when the true underlying population mean is 1.8.

2. Lets vary the values of the alternative hypothesis.
What i the result for the power?

Compute the power when we let delta increase

```
delta2 <- seq(from = 0.3, to = 1, by  = 0.1)
power.t.test(n = n, delta = delta2, sd = sd, sig.level = sig.level, power = NULL,
             type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 9
##           delta = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
##              sd = 0.44
##       sig.level = 0.05
##           power = 0.5886434, 0.8001086, 0.9275141, 0.9808051, 0.9963357, 0.9994994, 0.9999513, 0.999
##     alternative = one.sided
```

When you want to detect a larger difference in means, then the power of your test will increase.

3. Compute size of sample needed

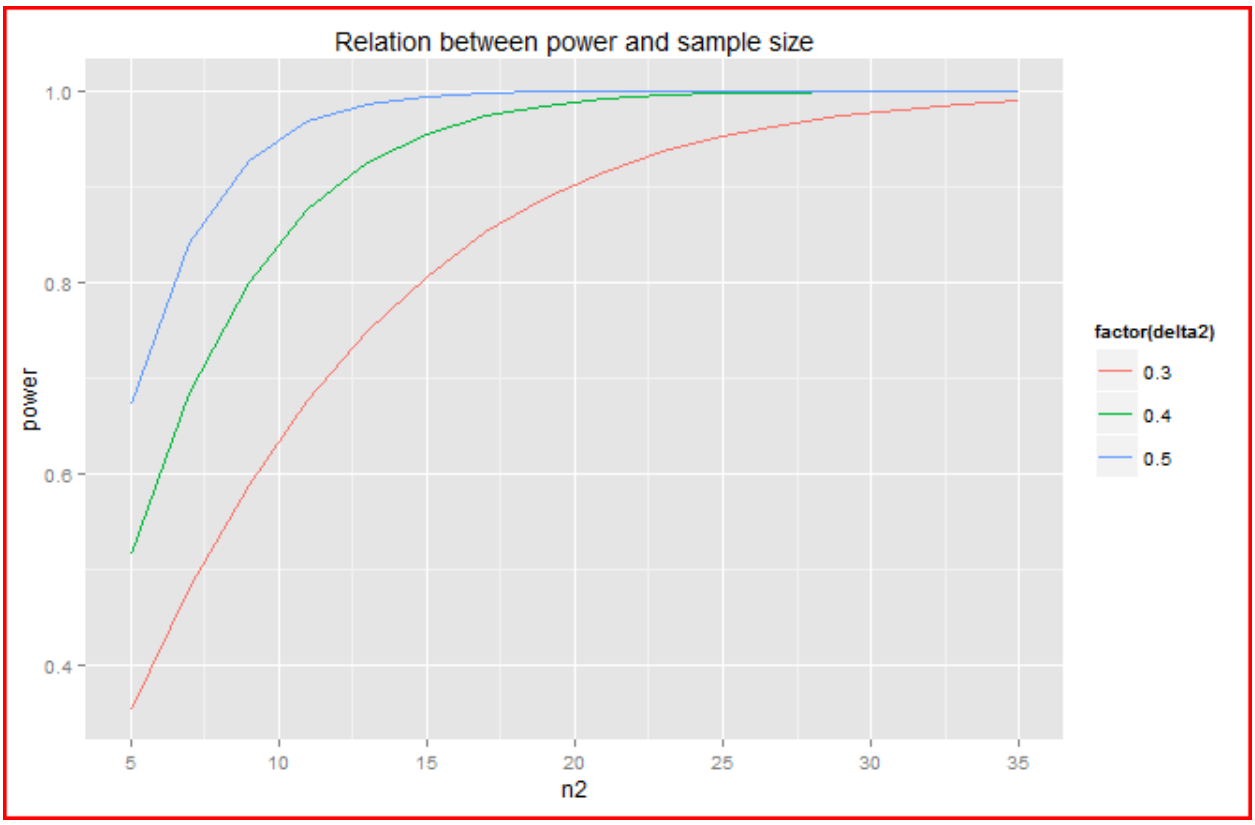
For the situation where the real underlying population average is 1.8. What is the sample size needed in order to obtain a power of 0.80?

```
delta <- 0.3
sd <- 0.44
sig.level <- 0.05
power.t.test(n = NULL, delta = delta, sd = sd, sig.level = sig.level, power = 0.8,
             type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 14.75024
##           delta = 0.3
##              sd = 0.44
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
```

We need a sample of at least size 15. This is the number of observations we need in order to detect a specific change in the gene expression, when the standard deviation is known.

4. Relationship power and sample size


Relation between power and sample size
power
1.0
0.8
0.6
0.4
5
10
15
20
25
30
35
n2
factor(delta2)
0.3
0.4
0.5