What does my data say? – Bivariate analysis

Inhalt

Objectives

The Chapter briefly introduces the most popular bivariate analysis methods in the health field. As the term indicates, the bivariate analysis aims to **investigate** the *relationship* between **two variables**. This investigation can include *statistical tests* and *correlation coefficients*. The choice of the statistical test and the correlation coefficient depends on the type of variables considered. In the following sections, you will be introduced to statistical tests and correlation coefficients separately to comprehend better their difference(s) in the context of bivariate analysis.

Statistical tests

Briefly, statistical tests include calculating a test statistic that transforms the measurements into a number. This number indicates how much the relationship between the investigated variables differs from the number assumed under the null hypothesis. The ultimate goal of the statistical tests is to investigate whether we fail to reject or reject the null hypothesis for a specific error, known as the type I error. Type I error indicates the probability of rejecting the null hypothesis when this hypothesis is true. A type I error at 5% is the status quo in the published literature. However, it can be adjusted to a lower or larger value under specific circumstances.

In essence, the statistical test merely answers a question with two possible answers: can we reject the null hypothesis, yes or no? You were introduced to null and alternative hypotheses in Chapter 2 (*Einfuehrung Studiendesign - Woher kommen die Daten?*). To answer this question, we need to calculate a probability that measures the probability of observing an association when there is truly no association under the null hypothesis. This probability is known as the p-valuen.

Though it may not always be obvious, the investigated variables are distinguished into two types: **dependent** and **independent** variables. Depending on the type of variables, the Table below summarises the proper statistical tests:

Dependent variable	**Independent variable**		
	metric	**binary**	**categorical**
metric	regression analysis	two-sample t-test	one-way ANOVA
binary	regression analysis	Fisher's exact test	chi-squared test
categorical	regression analysis	chi-squared test	chi-squared test

P-value: clinicians love it, but statisticians are sceptical

P-value is the probability of observing a test value (or association) at least as large when the null hypothesis is true. This is the statistical definition of the p-value. A p-value less than a significance threshold implies a **statistically significant** test (or association). In this case, we reject the null hypothesis. On the other hand, a p-value equal to or above the significance threshold implies a statistically non-significant test (or association). In this case, we fail to reject the null hypothesis and conclude that uncertainty about the observed association increases, and the test value may not reflect the truth. The significance threshold coincides with the type I error.

P-value answers a dichotomous question: is the observed association statistically significant or not? P-value provides no information about the clinical significance of the observed association.

However, misuses of the p-value indicating evidence of clinical significance have been recorded extensively in medical research. Furthermore, some researchers misinterpret the p-value as measuring the degree of statistical significance—for instance, a p-value of 0.1% implies higher statistical significance than a p-value of 1%. In March 2016, the American Statistical Association published a statement on the statistical association and p-values.

Metric dependent variable

Independent two-sample Student's t-test

William Sealy Gosset is the inventor of the t-test, and Student is his pseudonym. Wikipedia offers some interesting historical facts about the t-test. There are various 'versions' of the t-test. This Chapter focuses on **the independent, two-sample t-test**, which investigates whether two independent groups (a binary variable) differ on average regarding a metric variable. Hence, the (independent, two-sample) t-test is proper for a **binary independent variable** and a **metric dependent variable**, and it has the following simple formula:

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

where $\bar{x}_k$ is the sampling mean of group $k = A, B$, n_k is the sample size of group k and s_p is the *pooled* sampling standard deviation defined as follows

$$s_p = \sqrt{\frac{(n_A - 1) \times s_A^2 + (n_B - 1) \times s_B^2}{n_A + n_B - 2}}$$

with s_k^2 being the sampling variance of group k.

The null hypothesis states that both groups have the same **population** mean. In contrast, the alternative hypothesis states that the compared groups have a different **population** mean. In statistical terms, the null hypothesis is written as

$$\mu_A = \mu_B \text{ which is equivalent to } \mu_A - \mu_B = 0,$$

and the alternative hypothesis is written as

$$\mu_A \neq \mu_B \text{ which is equivalent to } \mu_A - \mu_B \neq 0.$$

We use the population mean (μ_k) to specify the hypothesis because we aim to draw inferences for the whole population. In practice, we use a representative sample of the target population to approximate μ_k (via $\bar{x}_k$) and perform the statistical test to decide whether we fail to reject or reject the null hypothesis with some (type I) error.

Our conclusions can be subject to error since we use a random collection of subjects (i.e., a sample). We usually keep this error at 5% and compare the p-value with this threshold. We reject the null hypothesis when the p-value is less than 5% and conclude that the compared groups may differ on average. Otherwise, we fail to reject the null hypothesis and conclude that **there is insufficient evidence to infer** whether the compared groups may differ on average.

Student's t-test is (actually) a two-step procedure

The formula above is based on a series of assumptions:

a) The metric variable follows a normal distribution in each group. Hence, the t-test is coined as a ***parametric test***. Normal distribution requires a sufficiently large sample to be approximated.
b) The compared groups are independent; namely, they comprise two different populations;
c) The sample variances do not differ beyond the sampling error.

Step 1

We need to apply **Levene's test** for two groups to investigate the third assumption. It tests the null hypothesis of equal population variances ($H_0: \sigma_A^2 = \sigma_B^2$) against the alternative hypothesis of unequal population variance ($H_a: \sigma_A^2 \neq \sigma_B^2$). Note that s_A^2 and s_B^2 are the sampling variances that approximate the corresponding population variances.

Step 2

When the p-value is less than 0.05, we reject the null hypothesis. Then, we conclude that **there is insufficient evidence** to infer whether the compared groups have the same population variance. In this case, the Student's t-test is ***not*** reliable and should ***not*** be applied. Instead, we apply **Welch's t-test**, which has the following simple formula,

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{V(\bar{x}_A - \bar{x}_B)}} = \frac{\bar{x}_A - \bar{x}_B}{\frac{s_A}{\sqrt{n_A}} + \frac{s_B}{\sqrt{n_B}}}$$

where $\bar{x}_k$ is the sampling mean of group $k = A, B$, s_k is the sampling standard deviation of group k, and n_k is the sample size of the group k.

Brainstorming

How do you investigate the first two assumptions of the Student's t-test? Spoiler alert for the first assumption: read the first part of the section on the Mann-Whitney U test.

Verdict on the two-stage approach

It may be argued that Welch's t-test prevails over Student's t-test for two reasons:

(1) it is not a two-stage approach as it does not require applying Levene's test to access whether the compared groups have the same population variance;
(2) Welch's t-test boils down to Student's t-test when the variances are equal.

Therefore, you can use Welch's t-test directly to compare the mean value of the two groups.

Application in the R software

We will use the `ToothGrowth` R built-in dataset to illustrate the Student's t-test and Welch's t-test.

```
> head(ToothGrowth)  # Show the first 6 rows.
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
```

The variables `len`, `supp` and `dose` have been measured in 60 guinea pigs and refer to tooth length (metric), supplement type (binary), and dose in milligrams/day (metric). We will compare the supplement type OJ with the VC regarding the average tooth length.

First, we apply Levene's test to investigate whether the supplement types have the same variance of tooth length. To perform Levene's test, we use the `leveneTest` function from the `car` package. The arguments for Levene's test are the following:

```
install.packages("car")  # Install the ‘car’ package
library(car)             # Load the ‘car’ package

leveneTest(len ~ supp,
           center = "mean",
           data = ToothGrowth)
```

We obtain the following results:

```
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  1.0973 0.2992
      58
```

The p-value is equal to 0.2992; hence, we **fail to reject** the null hypothesis. The next step is to perform Student's test using the `t.test` function from the `stats` package.

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

We will consider `alternative = "two.sided"` because we have not specified the direction of the difference in the alternative hypothesis. Furthermore, we keep `mu = 0`, which refers to the null hypothesis, `var.equal = TRUE`, and `conf.level = 0.95` (we want a 5% significance level). The argument `paired = FALSE` is the default, and we can safely remove it.

```
t.test(len ~ supp,
       alternative = "two.sided",
       mu = 0,
       var.equal = TRUE,
       conf.level = 0.95,
       data = ToothGrowth)
```

Then the function yields the following results for the Student's t-test:

```
                             Two Sample t-test
```

```
data:  len by supp
t = 1.9153, df = 58, p-value = 0.06039
alternative hypothesis: true difference in means between group OJ and
group VC is not equal to 0
95 percent confidence interval:
 -0.1670064  7.5670064
sample estimates:
mean in group OJ mean in group VC
        20.66333         16.96333
```

The p-value is 0.06039; hence, we **fail to reject** the null hypothesis. We conclude that **there is insufficient evidence** to infer whether the compared groups have different tooth lengths on average. Note that the range of the 95% confidence interval for the mean difference (i.e., $\bar{x}_A - \bar{x}_B = 20.66333 - 16.96333 = 3.70$) includes the zero value of no difference. Hence, we fail to reject the null hypothesis because the 95% confidence interval contains evidence for and against one group being heavier than the other.

We would reject the null hypothesis if the 95% confidence interval did *not* include the zero value. This is because the 95% confidence interval would contain consistent evidence of the mean difference (positive across the 95% confidence interval).

Brainstorming

Use the `t.test` function to perform Welch's t-test. As a hint, use `var.equal = FALSE`. Does it provide different results from the Student's t-test? Justify your response.

Paired-sample t-test

Contrary to the independent, two-sample t-test, the **paired-sample t-test** deals with one population where each subject is measured twice (e.g., before and after a medication). Hence, these before-after observations are *not* independent. The formula to perform a paired-sample t-test is the following:

$$t = \frac{d}{\frac{s_d}{\sqrt{n}}}$$

with d being the mean of the differences calculated for each pair of observations, s_d being the standard deviation of these differences, and n is the number of participants.

Application in the R software

The `ToothGrowth` dataset is not proper to demonstrate the paired-sample t-test. We will use the `mice2` dataset from the `datarium` package. This dataset contains the weight of 10 mice measured before and after the treatment. Make sure to install and load the package first.

```
paired <- data("mice2", package = "datarium")

> head(mice2)
  id before after
1  1  187.2 429.5
2  2  194.2 404.4
3  3  231.7 405.6
4  4  200.5 397.2
5  5  201.7 377.9
6  6  235.0 445.8
```

We want to investigate whether the weight changes on average aftet the medication. Hence, we compare the average weight after with that before the medication. To perform the paired-sample t-test in R using the `t.test` function, we include the argument `paired = TRUE` and we remove the argument `var.equal = TRUE` as it is redundant:

```
t.test(x = mice2$after, y = mice2$before,
       alternative = "two.sided",
       mu = 0,
       paired = TRUE,
       conf.level = 0.95)
```

We obtain the following results:

```
                              Paired t-test

data:  mice2$after and mice2$after
t = 25.546, df = 9, p-value = 1.039e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 181.8158 217.1442
sample estimates:
mean of the differences
                199.48
```

Two hints in the results indicate rejection of the null hypothesis. First, the p-value is less than 0.05. Second, the 95% confidence interval does not include the zero value of no difference (the mean difference is positive across the 95% confidence interval). We conclude that the weight changes on average after the medication. Specifically, the weight increased by 199.48 on average after the medication.

Mann-Whitney U test: the non-parametric independent, two-sample t-test

Testing for normality

To decide between a **parametric** and **non-parametric statistical test**, we must investigate whether the metric variable follows the normal distribution. The following actions need to be taken:

- check whether the mean and median are similar (subjective conclusions);
- look at the histogram and the Q-Q plot (subjective conclusions);
- use a statistical test (objective conclusions).

The proper statistical test is the **Kolmogorov-Smirnov test** when our sample is very large (> 2000). Otherwise, apply the **Shapiro-Wilk test**, which is more powerful than the Kolmogorov-Smirnov test even for small samples (< 50). Both test the null hypothesis of a normally distributed variable.

We use the `ToothGrowth` dataset to illustrate both tests in the R software. First, we will calculate the mean and median for the tooth length separately for OJ and VS. We will use the `aggregate` function.

```
mean.group <- aggregate(len ~ supp, data = ToothGrowth, mean)
median.group <- aggregate(len ~ supp, data = ToothGrowth, median)

results <- data.frame(mean.group$supp, mean.group$len, median.group$len)
colnames(results) <- c("group", "mean", "median")

> results
```

```
  group     mean median
1    OJ 20.66333   22.7
2    VC 16.96333   16.5
```

The mean and median are relatively similar for OJ and almost identical for VC. Note that 30 guinea pigs per supplement type constitute a very small sample. Hence, defending that `len` is normally distributed using descriptive statistics or a histogram for each supplement type is challenging.

The Kolmogorov-Smirnov test is ***not*** appropriate for this dataset. Hence, we will apply the Shapiro-Wilk test for each supplement type separately using the `shapiro.test` function.

```
# Subset the dataset per supplement type
oj.group <- subset(ToothGrowth, supp == "OJ")
vc.group <- subset(ToothGrowth, supp == "VC")

# Apply Shapiro-Wilk test in each supplement type
shapiro.test(oj.group$len)
shapiro.test(vc.group$len)
```

The results for each group are shown below:

```
> shapiro.test(oj.group$len)

      Shapiro-Wilk normality test

data:  oj.group$len
W = 0.91784, p-value = 0.02359

> shapiro.test(vc.group$len)

      Shapiro-Wilk normality test

data:  vc.group$len
W = 0.96567, p-value = 0.4284
```

The Shapiro-Wilk test rejects the null hypothesis for OJ (p-value < 0.05) but fails to reject the null hypothesis for VC (p-value > 0.05). The proper test to compare the supplement types in terms of tooth length would be non-parametric.

Note that the Kolmogorov-Smirnov test is performed using the ks.test function (also from the `stats` package) with the following arguments, for instance, for OJ:

```
ks.test(oj.group$len, "pnorm")
```

When the Shapiro-Wilk test rejects the null hypothesis for at least one group or the size of each group is very small, we must resort to the **Mann-Whitney U** test to compare the groups. The Mann-Whitney U test is also called Wilcoxon–Mann–Whitney test or the Wilcoxon Rank-Sum test. In essence, this test investigates the null hypothesis that the groups have the same distribution against the alternative hypothesis that the groups have different distributions.

We use the `wilcox.test` function from the `stats` package to perform the Mann-Whitney U test. We use the argument `pair = FALSE` to indicate that the compared groups are independent samples.

```
wilcox.test(len ~ supp, exact = FALSE, pair = FALSE, data = ToothGrowth)
```

The results are shown below:

```
      Wilcoxon rank sum test with continuity correction

```

```
data:  len by supp
W = 575.5, p-value = 0.06449
alternative hypothesis: true location shift is not equal to 0
```

The p-value is above 5%; thus, we **fail to reject** the null hypothesis. We conclude that **there is insufficient evidence** to infer whether the supplement types have a different distribution or, in other words, whether the supplement types differ in terms of tooth length. Note that we drew the same conclusions with the Student's t-test.

Initially, we used the `wilcox.test` function without the argument `exact` (`exact = NULL` by default) and received the following warning message:

```
Warning message:
In wilcox.test.default(x = c(15.2, 21.5, 17.6, 9.7, 14.5, 10, 8.2,  :
  cannot compute exact p-value with ties
```

To override this warning, we indicate `exact = FALSE`. Ties imply that there are identical values in the dataset. Identical test values cannot have unique ranks, and thus, an exact p-value cannot be calculated.

Both the Kolmogorov-Smirnov test and the Mann-Whitney U test are based on ranks. Are you curious about how this statistical test works? Then, read the site from the University of Virginia Library.

Wilcoxon signed-rank test: the non-parametric paired-sample t-test

Like the `ToothGrowth` dataset, the `mice2` dataset is too small to defend the normality assumption. Hence, a parametric test (here, the paired-sample t-test) would ***not*** be appropriate. The **Wilcoxon signed-rank test** is the non-parametric alternative to the paired-sample t-test.

To perform the Wilcoxon signed-rank test in R, we also use the `wilcox.test` function, and we set the argument `pair = TRUE`.

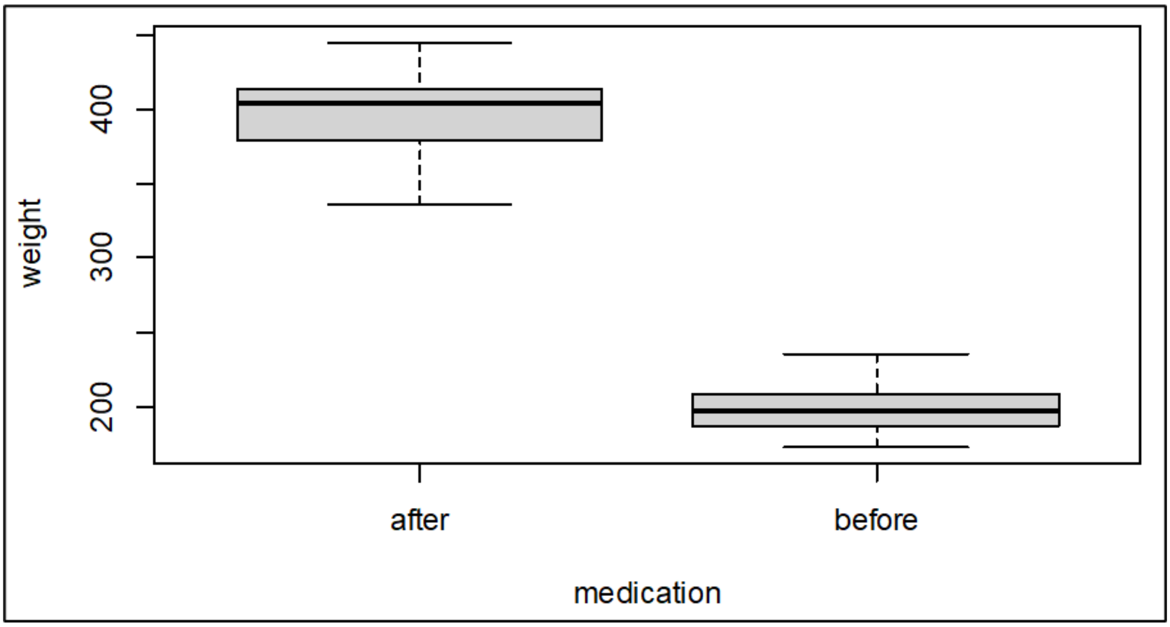
```
wilcox.test(x = mice2$after, y = mice2$before, pair = TRUE)
```

The results are shown below:

```
      Wilcoxon rank sum test with continuity correction

data:  mice2$after and mice2$before
V = 55, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

The p-value is below 5%, and we reject the null hypothesis. We conclude that the weight has a different distribution after medication, or in other words, the weight changes after medication. We can draw the boxplot to visualise the weight distribution before and after the medication.

```
# Turn the dataset in the long format
weight <- c(mice2$before, mice2$after)
medication <- rep(c("before", "after"), each = length(mice2$before))

# Create the boxplot
boxplot(weight ~ medication)
```



The median weight differs considerably before and after the medication. Note also that the two groups have a completely different range of values.

We use the `summary` function to obtain the descriptive statistics for the weight before and after medication.

```
> summary(mice2$before)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  172.4   187.8   197.3   200.6   206.9   235.0 

> summary(mice2$after)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  337.0   384.5   405.0   400.0   412.8   445.8 
```

The weight is consistently much larger after the medication as compared to before the medication.

One-way ANOVA

The **one-way ANOVA (ANalysis Of VAriance)** is the extension of the independent, two-sample Student's t-test when more than two groups are compared regarding a metric variable. Hence, the null hypothesis states that all compared groups have the same population mean. The alternative hypothesis states that at least one pair of groups differs in the population mean. The assumptions considered for the independent, two-sample Student's t-test are also relevant for the one-way ANOVA. However, simulation studies have shown that the one-way ANOVA is immune to departures from the normal distribution for each group.

One-way ANOVA distinguishes the (total) variation of the metric variable into the between-group variation and within-group variation. It uses the F-statistic to measure which of these two variations contributes more to the total variation. Specifically, $F > 1$ implies that the between-group variation contributes more to the total variation than the within-group variation (also known as an unexplained variation). Hence, at least one pair of groups may have different population means. The opposite is implied when $0 < F < 1$, and usually, it corresponds to a p-value above 0.05.

Why not perform multiple Student's t-tests?

When we reject the null hypothesis according to the F-statistic, the next step is determining which pair(s) of groups 'contributed' to reject the null hypothesis. In other words, we are interested in detecting the pair(s) of groups with different means. One may argue that applying the independent, two-sample Student's t-test for each possible pair of groups suffices. Besides, the Student's t-test is more straightforward than the one-way ANOVA. The problem is that applying the same test (here, the Student's t-test) multiple times on the same dataset increases the type I error. This is not desired because our conclusions would be misleading and possibly harmful to the end-user of our analysis. Increasing the type I error, we inevitably increase the threshold to compare with the p-value. Consequently, we make it easier to find spuriously statistically significant associations.

For a categorical variable with K levels, we observed $K(K-1)/2$ possible pairs of groups. Assuming a significance level of 0.05 and performing the Student's t-test at each pair will yield a type I error equal to $1-(1-0.05)^K$. For instance, if $K=3$, we have three possible pairs of groups: performing Student's t-test at each pair will yield a type I error equal to $1-(1-0.05)^3=0.14$ (after rounding at the second decimal). The increase in type I error due to multiple testing is known as **multiplicity**.

This issue is tackled by performing **posthoc tests**, which keep the type I error at the desired level (usually at 0.05). There are many post hoc tests to perform after running a one-way ANOVA. The Tukey test is the most commonly used posthoc test, which we will consider in this Chapter.

Application in the R software

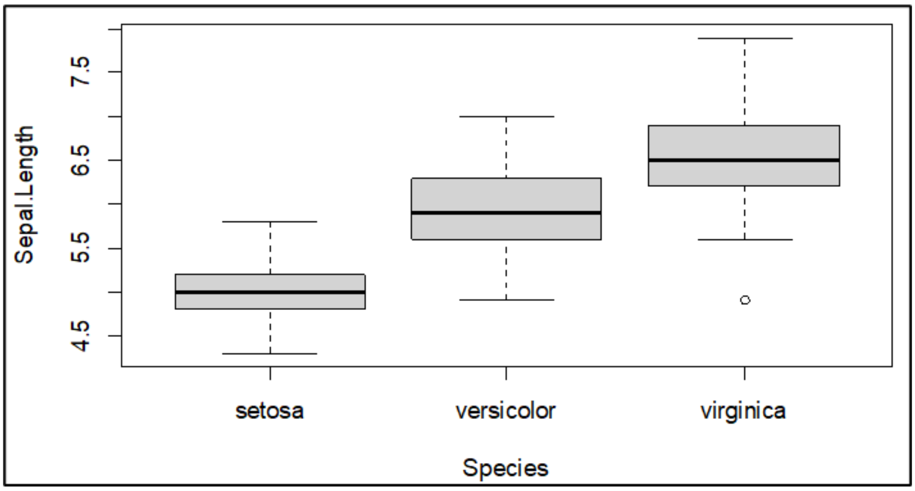
We will use the `iris` R built-in dataset to perform one-way ANOVA.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

The dataset includes 150 measurements in centimetres stemming from 50 flowers for each of the three `Species` of iris. There are four metric variables: `Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width`. For illustrative purposes, we will consider the `Sepal.Length`.

We use the boxplot function to visualise the distribution of `Sepal.Length` for each level of the variable `Species`.

```
boxplot(Sepal.Length ~ Species, data = iris)
```



The mean sepal length differs across the species. Furthermore, the range of values hardly overlaps for setosa and virginica. We want to investigate whether at least one species has a different mean of sepal length.

Like with Student's t-test, we need to test the assumption of equal variances across the groups (homoscedasticity) to decide whether we should pursue the common one-way ANOVA or Welch's ANOVA. We use the `bartlett.test` function to perform Bartlett's test of homoscedasticity:

```
bartlett.test(Sepal.Length ~ Species, data = iris)
```

We obtain the following results:

```
      Bartlett test of homogeneity of variances

data:  Sepal.Length by Species
Bartlett's K-squared = 16.006, df = 2, p-value = 0.0003345
```

The p-value is below 0.05, and we reject the null hypothesis that homoscedasticity exists across the different species. We conclude that the variance of sepal length differs in at least one species. Because the assumption of homoscedasticity is violated, we should proceed with Welch's ANOVA. However, we will apply both types of ANOVA for illustration.

The `oneway.test` function performs both types of ANOVA. We will use the following arguments for Welch's ANOVA:

```
welch.anova <- oneway.test(Sepal.Length ~ Species,
                           data = iris,
                           var.equal = FALSE)
```

We obtain the following results:

```
      One-way analysis of means

data:  Sepal.Length and Species
F = 119.26, num df = 2, denom df = 147, p-value < 2.2e-16
```

The F-test is accompanied by a p-value below 0.05; hence, we reject the null hypothesis. We conclude that there is at least one pair of species with a statistically significant mean difference.

In the case of one-way ANOVA, we need to specify the argument `var.equal = TRUE`. We obtain the following results:

```
             One-way analysis of means

data:  Sepal.Length and Species
F = 119.26, num df = 2, denom df = 147, p-value < 2.2e-16
```

We yield the same conclusions with Welch's ANOVA.

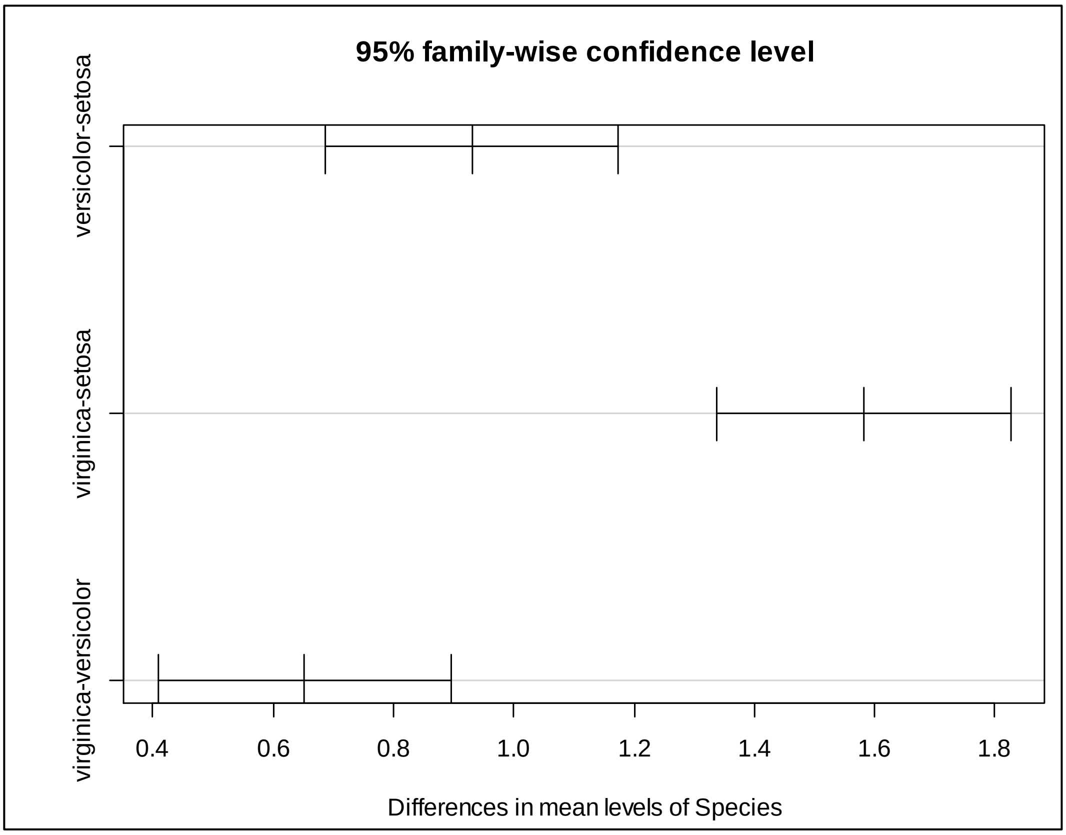
Next, we use the `TukeyHSD` function to conduct Tukey's test:

```
Tukey.test <- TukeyHSD(aov(Sepal.Length ~ Species, data = iris))

> Tukey.test
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sepal.Length ~ Species, data = iris)

$Species
                      diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

The zero value of no difference is not included in the 95% confidence interval (`lwr`, `upr`). Hence, we reject the null hypothesis that the compared species have the same sepal length on average.

The next line produces an interval plot that illustrates the results from Tukey's test:

```
plot(Tukey.test)
```



Each horizontal line refers to a pairwise comparison. The first and third vertical lines indicate the lower (`lwr`) and upper (`upr`) bound of the 95% confidence interval for each line. The central vertical line is the mean difference (`diff`).

Kruskal-Wallis test: the non-parametric one-way ANOVA

Kruskal-Wallis test is the **non-parametric** alternative to the one-way ANOVA. It investigates whether the compared groups have the same (null hypothesis) or different distribution (alternative hypothesis) regarding a metric variable. It is the extension of the Mann-Whitney U-test for more than two groups.

We use the `kruskal.test` function to perform the Kruskal-Wallis test for the above `iris` dataset.

```
kruskal.test(Sepal.Length ~ Species, data = iris)
```

We obtain the following results:

```
      Kruskal-Wallis rank sum test

data:  Sepal.Length by Species
Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

The p-value is below 0.05; hence, we reject the null hypothesis. We conclude that at least one pair of species differ in the distribution of the sepal length.

Dunn's posthoc test

The next is to perform a posthoc test to detect the pair(s) with different distributions of the sepal length. Dunn's test is the most commonly used posthoc test for the Kruskal-Wallis test. The `dunnTest` function from the `FSA` package performs Dunn's test. First, install and load the `FSA` package.

```
dunnTest(Sepal.Length ~ Species, data = iris, method = "bh")
```

We obtain the following results:

```
              Comparison         Z      P.unadj        P.adj
1    setosa - versicolor -6.106326 1.019504e-09 1.529257e-09
2     setosa - virginica -9.741785 2.000099e-22 6.000296e-22
3 versicolor - virginica -3.635459 2.774866e-04 2.774866e-04
```

We consider the results under `P.adj` as it adjusts the p-value for performing multiple comparisons among the species. The comparison `setosa - virginica` yields the higher difference according to `Z` (`Z = -9.741785`). The summary statistics for the sepal length per species confirm that virginica has the largest median sepal length, followed by versicolor, and setosa (see also boxplot):

```
> summary(subset(iris$Sepal.Length, iris$Species == "setosa"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   4.800   5.000   5.006   5.200   5.800

> summary(subset(iris$Sepal.Length, iris$Species == "versicolor"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   5.600   5.900   5.936   6.300   7.000

> summary(subset(iris$Sepal.Length, iris$Species == "virginica"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   6.225   6.500   6.588   6.900   7.900
```

Nominal dependent variable

The term nominal comprises binary (two categories) and categorical (more than two categories) variables. When we investigate a pair of nominal variables, we can select between two statistical tests: (i) Pearson's chi-squared test and (ii) Fisher's exact test. Both test the null hypothesis that there is no relationship between two nominal variables against the alternative hypothesis that there is a relationship between two nominal variables.

Pearson's chi-squared test

A $k \times n$ contingency table

We use a $k \times n$ contingency table, namely, a matrix to display the combinations of the levels of a variable with k categories and a variable with n categories. A $k \times n$ matrix has k rows and n columns. We will use the `survey` R built-in dataset that contains 237 students and 12 variables. First, install and load the `MASS` library.

```
> head(survey)
     Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I    Age
1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric 18.250
2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80 Imperial 17.583
3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA     <NA> 16.917
4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric 20.333
5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric 23.667
6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial 21.000
```

Suppose we investigate the relationship between smoking (`Smoke`) and exercise habits (`Exer`). We display the contingency table for these two categorical variables:

```
> Table (survey$Smoke, survey$Exer)

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

`Smoke` appears on the rows and has four levels: `Heavy`, `Never`, `Occas`, and `Regul`. `Exer` appears on the columns and has three levels: `Freq`, `None`, and `Some`. The most frequent combination is non-smokers who exercise frequently (87 students) or some (84 students). Surprisingly, the least frequent combination is heavy and regular smokers who do not exercise (1 student each). There is no clear pattern in the data. Let us see what we can conclude with the chi-square test. We use the `chisq.test` function from the `stats` package with the following arguments:

```
(res <- chisq.test(x = survey$Smoke, y = survey$Exer))
```

We obtain the following results accompanied with a warning (in red):

```
       Pearson's Chi-squared test

data:  survey$Smoke and survey$Exer
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(x = survey$Smoke, y = survey$Exer) :
  Chi-squared approximation may be incorrect
```

The p-value is above 0.05, and we **fail to reject** the null hypothesis, as **there is insufficient evidence** to infer whether there is a relationship between smoking and exercise habits. However, we received a warning that the chi-squared test is inappropriate for this dataset.

Uncovering the warning message

Under the null hypothesis, the chi-squared test follows a chi-squared distribution, provided that:

(1) the total number of units (here, students) is large enough, and
(2) the expected frequencies are at least as large as 5.

Pearson's chi-squared is **improper** if these conditions are not met, as it may provide misleading results.

To calculate the expected frequencies in each table cell, multiply the total in the corresponding row by the total in the corresponding column and divide by the total number of units (here, 237 students). For instance, the expected value for a heavy smoker (`Heavy`) who never exercises (`None`) is

$$\frac{(7 + 1 + 3) \times (1 + 18 + 3 + 1)}{237} = 1.07$$

after rounding at the second decimal.

We look at the expected frequencies to understand why we received this warning message:

```
> round(res$expected, 2)
            survey$Exer
survey$Smoke  Freq  None  Some
       Heavy  5.36  1.07  4.57
       Never 92.10 18.42 78.48
       Occas  9.26  1.85  7.89
       Regul  8.28  1.66  7.06
```

There are four cells with expected frequencies of less than 5. Hence, we cannot trust the results of the chi-squared test.

Reacting to the warning message

To address this limitation, some researchers proceed with a posthoc merging of the categories of the variables. The goal is *not* to observe expected frequencies below 5 in the new contingency table. For instance, we may merge the columns `None` with `Some`:

```
# This is the initial table
cont.tab <- table(survey$Smoke, survey$Exer)

# Proceed with the column-merging
new.tab <- cbind(cont.tab[, "Freq"],
                 cont.tab[, "None"] + cont.tab[, "Some"])

# Re-name the columns
colnames(new.tab) <- c("Freq", "Infreq")
```

The new Table is the following:

```
> new.tab
      Freq Infreq
```

```
Heavy    7      4
Never   87    102
Occas   12      7
Regul    9      8
```

We perform the chi-squared test on the new Table, and we obtain the following results:

```
> chisq.test(new.tab)

      Pearson's Chi-squared test

data:  new.tab
X-squared = 3.2328, df = 3, p-value = 0.3571
```

The good news is that we tackled the warning message successfully. The bad news is that we partially manipulated the data to make the test work. Merging may be pursued when:

(1) the variables have many categories as the risk of observing at least one expected frequency below five increases, and
(2) the information is not distorted; namely, the merging of categories is plausible.

However, **applying a multivariable regression model that includes many (pre-specified) independent variables is the best approach**. The many benefits of this strategy include:

(1) no data manipulation like merging of categories;
(2) accountability of inherent relationships among the independent variables that the bivariate analysis (by definition) misses;
(3) finding statistically significant relationships with more than one independent variable, and
(4) use the results to develop guidelines.

Fisher's exact test

While Pearson's chi-square test is relevant to contingency tables of any dimension, **Fisher's exact test** is relevant only for 2 × 2 contingency tables, namely, a table of two binary variables. Contrary to Pearson's chi-square test, Fisher's exact test follows a hypergeometric distribution, performs well for small sample sizes and calculates an exact p-value. Furthermore, suppose we perform Pearson's chi-square test for two binary variables and obtain a warning message that the chi-squared approximation may be incorrect. This issue can be addressed by applying Fisher's exact test.

We will use the `Melanoma` R built-in dataset from the `MASS` package to illustrate Fisher's exact test.

```
> head(Melanoma)
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

We will investigate whether the presence of an ulcer (`ulcer`: $1 =$ presence, $0 =$ absence) may be associated with gender (`sex`: $1 =$ male, $0 =$ female) using a sample of 205 patients in Denmark with malignant melanoma. We display the contingency table of these two binary variables:

```
> (melan <- table(Melanoma$sex, Melanoma$ulcer))

     0  1
```

```
  0 79 47
  1 36 43
```

To understand whether men or women are more likely to develop an ulcer, we will also display the contingency table with the proportion of frequencies using the `prop.table` function:

```
> prop.table(melan , margin = 1)

            0         1
  0 0.6269841 0.3730159
  1 0.4556962 0.5443038
```

Men are more likely to develop an ulcer than women (54% versus 37%), and women are more likely not to develop an ulcer than men (63% versus 46%).

We use the `fisher.test` function from the `stats` package to perform with the following arguments:

```
fisher.test(x = Melanoma$sex, y = Melanoma$ulcer)
```

which is equivalent to the following usage:

```
fisher.test(melan)
```

We obtain the following results:

```
      Fisher's Exact Test for Count Data

data:  melan
p-value = 0.02061
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.089871 3.700065
sample estimates:
odds ratio
  2.000701
```

The p-value is less than 0.05; hence, we reject the null hypothesis that there is no relationship between gender and the presence of an ulcer. The odd ratio value indicates that the odds of an ulcer are twice higher in men than in women. The evidence is statistically significant because the 95% confidence interval does not include the value of no difference (i.e., an odd ratio of 1).

Correlation coefficients

The Chapter concludes with (probably) the three most frequently used correlation coefficients in the literature: (i) Pearson's correlation coefficient, (ii) Spearman's correlation coefficient, and (iii) Kendall's correlation coefficient. Below, we briefly delineate their distinct differences and illustrate their use in R.

Pearson's correlation coefficient

Pearson's correlation coefficient, also known as **Pearson's r**, is proper for two metric variables and measures the strength of **linear** association between these two variables. It takes values from -1 to 1, with -1 implying a perfect negative linear correlation, 0 implying no linear correlation and 1 implying a perfect positive linear correlation. The following guidelines have been proposed to interpret Pearson's r values:

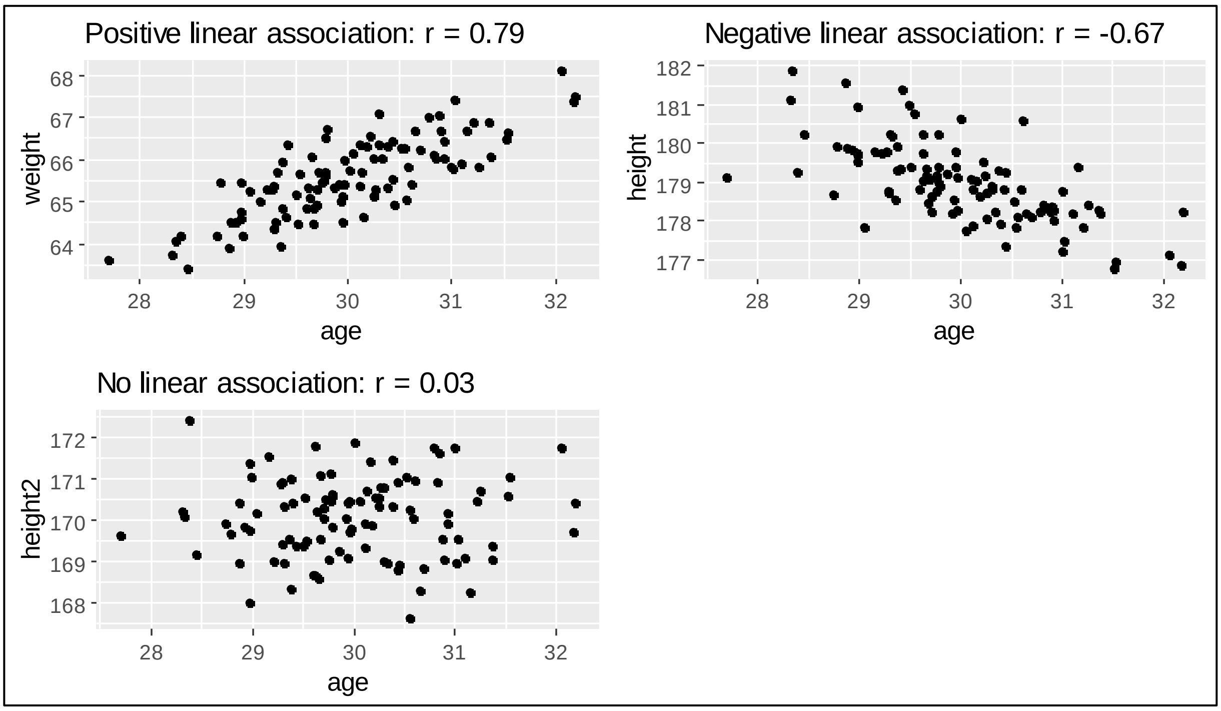
- high linear correlation: r lies between ±0.5 and ±1;
- moderate linear correlation: r lies between ±0.3 and ±0.49, and
- low linear correlation: r lies below ±0.29.

We will simulate a fictional example of positive, negative and no linear association. Pearson's correlation coefficient does not require the variables to be normally distributed. For the simulation purposes, we considered bivariate normal distribution with $r = 0.85$, $r = -0.70$, and $r = 0$ for positive, negative and no linear association, respectively.

```
set.seed(123)     # For reproducibility of the results

# Positive linear association of weight with age
age <- rnorm(100, 30, 1)
weight <- rnorm(100, 40 + 0.85*age, sqrt(1 - (0.85)^2))

# Negative linear association of height with age
height <- rnorm(100, 200 - 0.70*age, sqrt(1 - (-0.70)^2))

# No linear association between age and height
height2 <- rnorm(100, 170, 1)
```

We will use the `ggplot` function from the `ggplot2` package to create the scatterplots for each scenario. With the `ggarrange` function from the `ggpubr` package, we will plot all three scatterplots in one page. First, install and load both packages.

```
# Positive linear association of weight with age
positive <- ggplot(data.frame(age, weight), aes(x = age, y = weight)) +
              geom_point() +
              labs(title = paste("Positive linear association: r =",
                                 round(cor(age, weight), 2)))

# Negative linear association of height with age
negative <- ggplot(data.frame(age, height), aes(x = age, y = height)) +
              geom_point() +
              labs(title = paste("Negative linear association: r =",
                                 round(cor(age, height), 2)))

# No linear association between age and height
no.assoc <- ggplot(data.frame(age, height2), aes(x = age, y = height2)) +
              geom_point() +
              labs(title = paste("No linear association: r =",
                                 round(cor(age, height2), 2)))

# Bring all plots together
ggarrange(positive, negative, no.assoc)
```

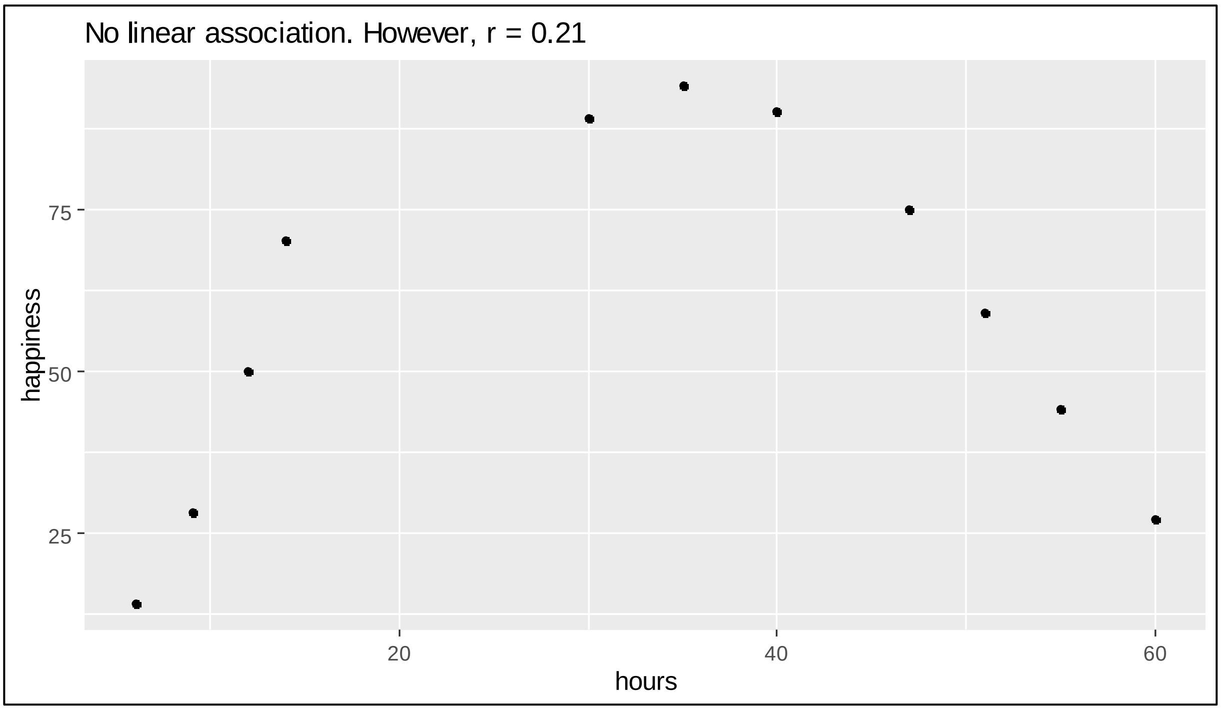


Next, we will illustrate with another fictional example that Pearson's r is misleading when the variables have a non-linear association. The source of the data is here.

```
data <- data.frame(hours = c(6, 9, 12, 14, 30, 35, 40, 47, 51, 55, 60),
                   happiness = c(14, 28, 50, 70, 89, 94, 90, 75, 59, 44, 27))
```

The following lines produce the scatterplot:

```
ggplot(data, aes(x = hours, y = happiness)) +
  geom_point() +
  labs(title = paste("No linear association. However, r =",
                     round(cor(data$hours, data$happiness), 2)))
```



According to the scatterplot, happiness is associated with the number of hours worked per week, but not linearly. Happiness increases to 38 worked hours and decreases as the hours continue to increase. Pearson's r indicates a spurious low linear association.

The **take-home message** of this example is that we should always create the scatterplot to decide whether we should apply Pearson's r.

Spearman's correlation coefficient

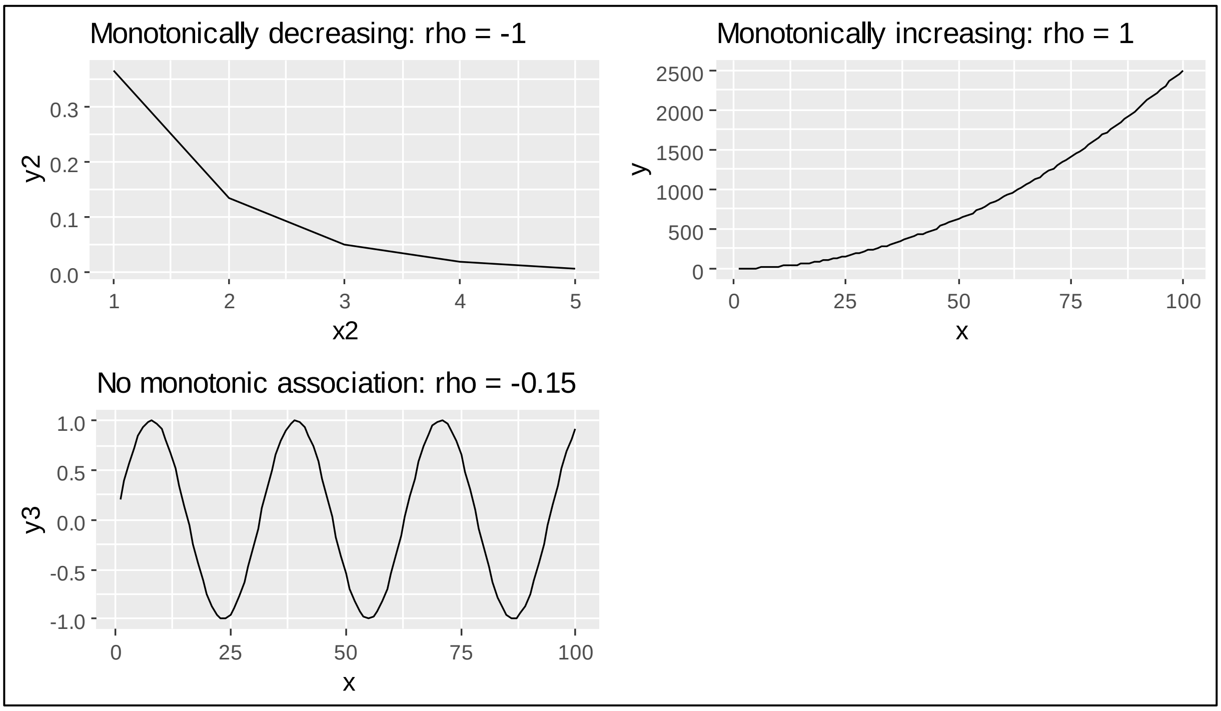
Pearson's r considers the raw data to measure the linear association of two metric variables. On the contrary, **Spearman's correlation coefficient** (also known as **Spearman's ρ**) converts the raw data into ranks to measure the strength of the monotonic association of two metric variables. Like Pearson's r, Spearman's ρ takes values from -1 to 1:

- -1 implies a perfect negative monotonic association;
- 0 implies no monotonic association, and
- 1 implies a perfect positive monotonic association.

Monotonic association implies that the data points are scattered in the same or different directions across the values of both variables, and the pattern does not need to be linear.

We will illustrate two examples where Spearman's ρ is appropriate and one example where neither Spearman's ρ nor Pearson's r should be applied. The source of the data is here.

```
x <- 1:100
x2 <- 1:5

y <- (1/4) * x^2 # Monotonically Increasing Function
y2 <- exp(-x2)   # Montonically Decreasing
y3 <- sin(x / 5) # Not Monotonic

# Monotonic decreasing association
mon.dec <- ggplot(data.frame(x2, y2), aes(x = x2, y = y2)) +
             geom_line() +
             labs(title = paste("Monotonically decreasing: rho =",
                                round(cor(x2, y2, method = "spearman"), 2)))

# Monotonic increasing association
mon.inc <- ggplot(data.frame(x, y), aes(x = x, y = y)) +
             geom_line() +
             labs(title = paste("Monotonically increasing: rho =",
                                round(cor(x, y, method = "spearman"), 2)))

# *No* monotonic association
not.mon <- ggplot(data.frame(x, y3), aes(x = x, y = y3)) +
             geom_line() +
             labs(title = paste("No monotonic association: rho =",
                                round(cor(x, y3, method = "spearman"), 2)))

# Bring all plots together
ggarrange(mon.dec, mon.inc, not.mon)
```



We used the `cor` function again and included the argument `method = "spearman"` to obtain the Spearman's ρ.

Kendall's correlation coefficient

Kendall's correlation coefficient (also known as **Kendall's tau**) is preferred to Spearman's ρ when there are many ties in the ranks or the dataset is small. Like Pearson's r and Spearman's ρ, Kendall's tau takes values from -1 to 1.

We use the `cor` function and include the argument `method = "kendall"` to obtain Kendall's tau.

Brainstorming

Calculate Kendall's tau using the dataset considered to illustrate Spearman's ρ. What do you conclude?

Perspectives on Bivariate Analysis

The Chapter concludes with some perspective on the usability of bivariate analysis for a (retrospective or prospective) research study that aims to detect associations between a dependent variable and multiple independent variables. Then, the main inference tool shall be a regression model than a bivariate analysis alone.

It is intriguing to check every pair of variables using the corresponding statistical tests or correlation coefficients. However, you risk digging very deep into the rabbit hole. What may start as a genuine interest in getting to know our data may result in losing the forest for the trees. The latter can happen if you use the bivariate analysis to decide on the subset of independent variables to base the regression analysis.

Provided the independent variables under investigation have been carefully pre-specified in a protocol, any retrospective deviations from the protocol shall be documented and thoroughly justified. Retrospective decisions *not* to include a variable due to missing data in two-thirds of the sample units may be rational. Parameter estimation becomes unstable from a certain point onward. Removing this variable may allow the statistical model to provide better predictions using the remaining variables that do not suffer considerable missingness.

Not to include a subset of variables because the bivariate analysis did not reveal any statistically significant associations constitutes **data dredging** (also known as **data snooping**).

It is rational to wonder how we can make the best of a bivariate analysis without conducting data dredging. Use the bivariate analysis to check the distribution of the remaining variables to the levels of a nominal target-independent variable.

This bivariate investigation is highly relevant in observational studies where randomisation is not applicable. Consequently, observed and unobserved characteristics are unlikely to be balanced in the compared groups of the target-independent variable.

Suppose we are interested in investigating whether diet A or B reduces blood pressure substantially. An observational study is a proper design for this investigation since randomising participants to either diet may be highly unethical. Hence, any differences in the compared diets regarding blood pressure will not be the diet alone but an amalgamation of diet, demographic, social and clinical characteristics. A bivariate analysis will investigate whether the compared diets are similar concerning demographic, social and clinical characteristics (the null hypothesis). Then, for each characteristic, perform the proper statistical test at a significance level of 0.05. Tabulate the descriptive statistics of the corresponding characteristics per diet and present the results from the bivariate analysis. The following Table is a recommended tabulation of the characteristics (also known as the Table of characteristics). It illustrates the distribution of several characteristics per diet and the p-value from the corresponding bivariate analyses.

Characteristic	Diet A	Diet B
Demographic Characteristics		
Age		
Mean (StD)	value (value)	value (value)
Median (IQR)	value (value – value)	value (value – value)
Min. – Max.	value – value	value – value
Missing data	n (%)	n (%)
Welch's t-test	p-value	
Gender		
Male	n (%)	n (%)
Female	n (%)	n (%)
Fisher's exact test	p-value	
Clinical characteristics		
…	…	…

Usually, the unbalanced characteristics in the compared groups (here, diet) may be responsible for convergence issues and warnings later in the regression analysis. Knowing which characteristics are seriously unbalanced in the compared groups may help you better tune your model and obtain plausible results.