Repeated Measures ANOVA in R

The **repeated-measures ANOVA** is used for analyzing data where same subjects are measured more than once. This test is also referred to as a *within-subjects ANOVA* or *ANOVA with repeated measures*. The "within-subjects" term means that the same individuals are measured on the same outcome variable under different time points or conditions.

For example, you might have measured 10 individuals' self-esteem score (the outcome or dependent variable) on three time points during a specific diet to determine whether their self-esteem improved.

This chapter describes the different types of repeated measures ANOVA, including:

- **One-way repeated measures ANOVA**, an extension of the paired-samples t-test for comparing the means of three or more levels of a *within-subjects* variable.
- **two-way repeated measures ANOVA** used to evaluate simultaneously the effect of two within-subject factors on a continuous outcome variable.
- **three-way repeated measures ANOVA** used to evaluate simultaneously the effect of three within-subject factors on a continuous outcome variable.

The main goal of two-way and three-way repeated measures ANOVA is, respectively, to evaluate if there is a statistically significant interaction effect between two and three within-subjects factors in explaining a continuous outcome variable.

You will learn how to:

- **Compute and interpret the different repeated measures ANOVA in R**.
- **Check repeated measures ANOVA test assumptions**
- **Perform post-hoc tests**, multiple pairwise comparisons between groups to identify which groups are different
- **Visualize the data** using box plots, add ANOVA and pairwise comparisons p-values to the plot

Contents:

Related Book

Practical Statistics in R II - Comparing Groups: Numerical Variables

Assumptions

The repeated measures ANOVA makes the following assumptions about the data:

- **No significant outliers** in any cell of the design. This can be checked by visualizing the data using box plot methods and by using the function `identify_outliers()` [rstatix package].
- **Normality**: the outcome (or dependent) variable should be approximately normally distributed in each cell of the design. This can be checked using the **Shapiro-Wilk normality test** (`shapiro_test()` [rstatix]) or by visual inspection using **QQ plot** (`ggqqplot()` [ggpubr package]).
- **Assumption of sphericity**: the variance of the differences between groups should be equal. This can be checked using the **Mauchly's test of sphericity**, which is automatically reported when using the R function `anova_test()` [rstatix package]. Read more in Chapter @ref(mauchly-s-test-of-sphericity-in-r).

Before computing repeated measures ANOVA test, you need to perform some preliminary tests to check if the assumptions are met.

Note that, if the above assumptions are not met there are a non-parametric alternative (*Friedman test*) to the one-way repeated measures ANOVA.

Unfortunately, there are no non-parametric alternatives to the two-way and the three-way repeated measures ANOVA. Thus, in the situation where the assumptions are not met, you could consider running the two-way/three-way repeated measures ANOVA on the transformed and non-transformed data to see if there are any meaningful differences.

If both tests lead you to the same conclusions, you might not choose to transform the outcome variable and carry on with the two-way/three-way repeated measures ANOVA on the original data.

It's also possible to perform robust ANOVA test using the **WRS2** R package.

No matter your choice, you should report what you did in your results.

Prerequisites

Make sure that you have installed the following R packages:

- `tidyverse` for data manipulation and visualization
- `ggpubr` for creating easily publication ready plots
- `rstatix` provides pipe-friendly R functions for easy statistical analyses
- `datarium` : contains required data sets for this chapter

Start by loading the following R packages:

```
1  library(tidyverse)
2  library(ggpubr)
3  library(rstatix)
```

Key R functions:

- ```
  1  anova_test()
  ```

  [rstatix package], a wrapper around

  ```
  1  car::Anova()
  ```

  for making easy the computation of repeated measures ANOVA. Key arguments for performing repeated measures ANOVA:
    - `data` : data frame
    - `dv` : (numeric) the dependent (or outcome) variable name.
    - `wid` : variable name specifying the case/sample identifier.
    - `within` : within-subjects factor or grouping variable
- `get_anova_table()` [rstatix package]. Extracts the ANOVA table from the output of `anova_test()` . It returns ANOVA table that is automatically corrected for eventual deviation from the sphericity assumption. The default is to apply automatically the Greenhouse-Geisser sphericity correction to only within-subject factors violating the sphericity assumption (i.e., Mauchly's test p-value is significant, $p <= 0.05$). Read more in Chapter @ref(mauchly-s-test-of-sphericity-in-r).

One-way repeated measures ANOVA

Data preparation

We'll use the self-esteem score dataset measured over three time points. The data is available in the datarium package.

```
1  # Data preparation
2  # Wide format
3  data("selfesteem", package = "datarium")
4  head(selfesteem, 3)
```

```
 5  ## # A tibble: 3 x 4
 6  ##       id    t1    t2    t3
 7  ##    <int> <dbl> <dbl> <dbl>
 8  ## 1     1  4.01  5.18  7.11
 9  ## 2     2  2.56  6.91  6.31
10  ## 3     3  3.24  4.44  9.78
11  # Gather columns t1, t2 and t3 into long format
12  # Convert id and time into factor variables
13  selfesteem <- selfesteem %>%
14    gather(key = "time", value = "score", t1, t2, t3) %>%
15    convert_as_factor(id, time)
16  head(selfesteem, 3)
17  ## # A tibble: 3 x 3
18  ##    id    time   score
19  ##    <fct> <fct> <dbl>
20  ## 1 1     t1      4.01
21  ## 2 2     t1      2.56
22  ## 3 3     t1      3.24
```

The one-way repeated measures ANOVA can be used to determine whether the means self-esteem scores are significantly different between the three time points.

Summary statistics

Compute some summary statistics of the self-esteem `score` by groups (`time`): mean and sd (standard deviation)

```
1  selfesteem %>%
2    group_by(time) %>%
3    get_summary_stats(score, type = "mean_sd")
4  ## # A tibble: 3 x 5
5  ##    time  variable     n  mean    sd
6  ##    <fct> <chr>    <dbl> <dbl> <dbl>
7  ## 1 t1    score       10  3.14 0.552
8  ## 2 t2    score       10  4.93 0.863
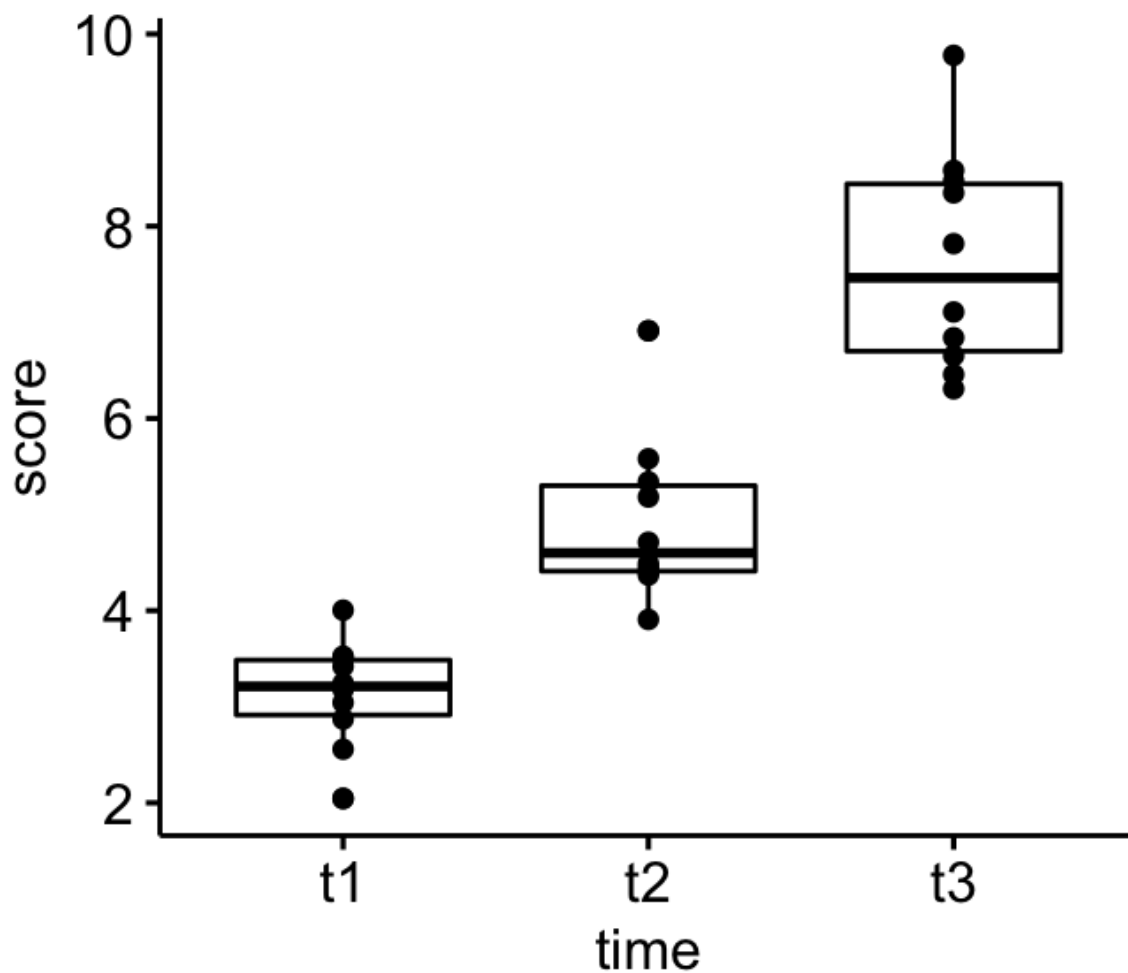9  ## 3 t3    score       10  7.64 1.14
```

Visualization

Create a box plot and add points corresponding to individual values:

```
1  bxp <- ggboxplot(selfesteem, x = "time", y = "score", add = "point")
2  bxp
```

Check assumptions

Outliers

Outliers can be easily identified using box plot methods, implemented in the R function `identify_outliers()` [rstatix package].

```
1  selfesteem %>%
2    group_by(time) %>%
3    identify_outliers(score)
4  ## # A tibble: 2 x 5
5  ##    time  id    score is.outlier is.extreme
6  ##    <fct> <fct> <dbl> <lgl>      <lgl>
7  ## 1 t1    6      2.05 TRUE       FALSE
8  ## 2 t2    2      6.91 TRUE       FALSE
```

There were no extreme outliers.

Note that, in the situation where you have extreme outliers, this can be due to: 1) data entry errors, measurement errors or unusual values.

You can include the outlier in the analysis anyway if you do not believe the result will be substantially affected. This can be evaluated by comparing the result of the ANOVA with and without the outlier.

It's also possible to keep the outliers in the data and perform robust ANOVA test using the WRS2 package.

Normality assumption

The normality assumption can be checked by computing Shapiro-Wilk test for each time point. If the data is normally distributed, the p-value should be greater than 0.05.

```
selfesteem %>%
  group_by(time) %>%
  shapiro_test(score)
## # A tibble: 3 x 4
##   time  variable statistic     p
##   <fct> <chr>        <dbl> <dbl>
## 1 t1    score        0.967 0.859
## 2 t2    score        0.876 0.117
## 3 t3    score        0.923 0.380
```

The self-esteem score was normally distributed at each time point, as assessed by Shapiro-Wilk's test (p > 0.05).

Note that, if your sample size is greater than 50, the normal QQ plot is preferred because at larger sample sizes the Shapiro-Wilk test becomes very sensitive even to a minor deviation from normality.

QQ plot draws the correlation between a given data and the normal distribution. Create QQ plots for each time point:

```
ggqqplot(selfesteem, "score", facet.by = "time")
```
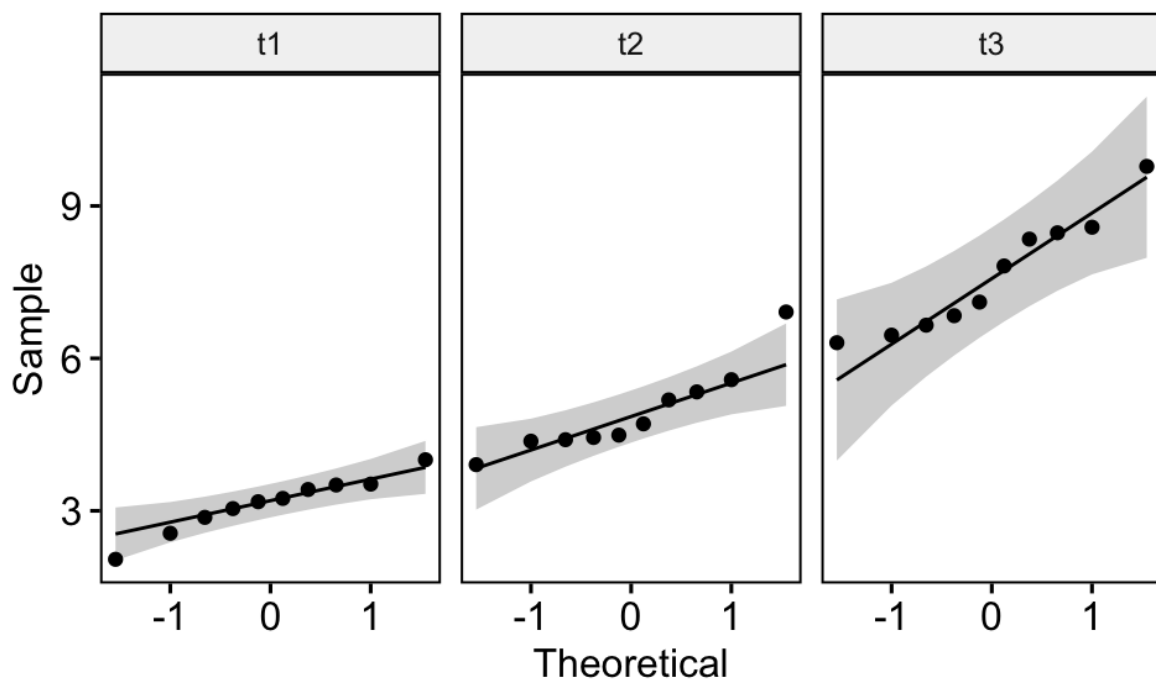

From the plot above, as all the points fall approximately along the reference line, we can assume normality.

Assumption of sphericity

As mentioned in previous sections, the assumption of sphericity will be automatically checked during the computation of the ANOVA test using the R function `anova_test()` [rstatix package]. The Mauchly's test is internally used to assess the sphericity assumption.

By using the function `get_anova_table()` [rstatix] to extract the ANOVA table, the Greenhouse-Geisser sphericity correction is automatically applied to factors violating the sphericity assumption.

Computation

```
1   res.aov <- anova_test(data = selfesteem, dv = score, wid = id, within = time)
2   get_anova_table(res.aov)
3   ## ANOVA Table (type III tests)
4   ##
5   ##   Effect DFn DFd    F        p p<.05   ges
6   ## 1   time   2  18 55.5 2.01e-08     * 0.829
```

The self-esteem score was statistically significantly different at the different time points during the diet, F(2, 18) = 55.5, p < 0.0001, eta2[g] = 0.83.

where,

- `F` Indicates that we are comparing to an F-distribution (F-test); `(2, 18)` indicates the degrees of freedom in the numerator (DFn) and the denominator (DFd), respectively; `55.5` indicates the obtained F-statistic value
- `p` specifies the p-value
- `ges` is the generalized effect size (amount of variability due to the within-subjects factor)

Post-hoc tests

You can perform multiple **pairwise paired t-tests** between the levels of the within-subjects factor (here `time`). P-values are adjusted using the Bonferroni multiple testing correction method.

```
1   # pairwise comparisons
2   pwc <- selfesteem %>%
3     pairwise_t_test(
4       score ~ time, paired = TRUE,
5       p.adjust.method = "bonferroni"
6       )
7   pwc
8   ## # A tibble: 3 x 10
9   ##   .y.   group1 group2    n1    n2 statistic    df          p    p.adj
        p.adj.signif
10  ## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>      <dbl>    <dbl>
        <chr>
11  ## 1 score t1     t2        10    10     -4.97     9 0.000772     0.002      **

12  ## 2 score t1     t3        10    10     -13.2     9 0.000000334 0.000001
        ****
13  ## 3 score t2     t3        10    10     -4.87     9 0.000886     0.003      **
```

All the pairwise differences are statistically significant.

Report

We could report the result as follow:

The self-esteem score was statistically significantly different at the different time points, F(2, 18) = 55.5, p < 0.0001, generalized eta squared = 0.82.

Post-hoc analyses with a Bonferroni adjustment revealed that all the pairwise differences, between time points, were statistically significantly different (p <= 0.05).

```
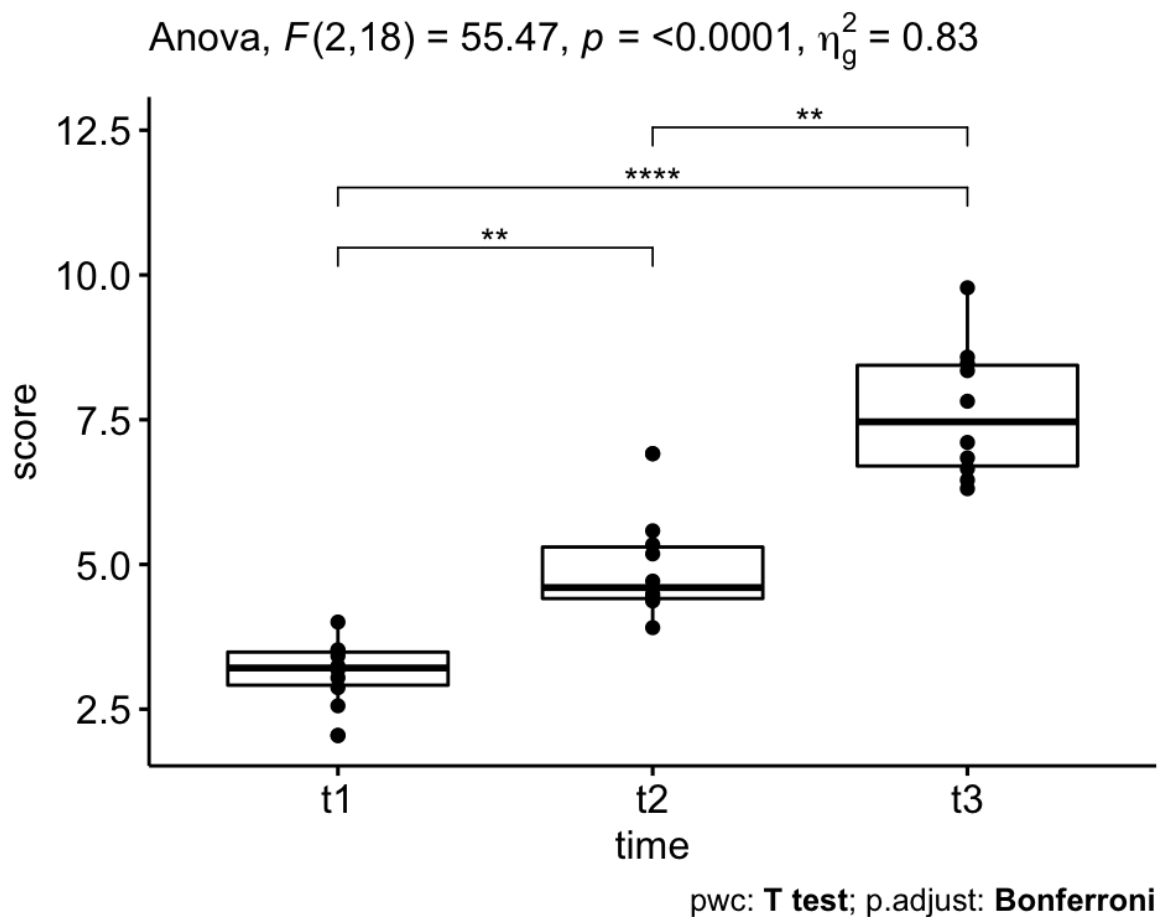1  # Visualization: box plots with p-values
2  pwc <- pwc %>% add_xy_position(x = "time")
3  bxp +
4    stat_pvalue_manual(pwc) +
5    labs(
6      subtitle = get_test_label(res.aov, detailed = TRUE),
7      caption = get_pwc_label(pwc)
8    )
```


Anova, $F(2,18) = 55.47$, $p = {<}0.0001$, $\eta_g^2 = 0.83$

pwc: **T test**; p.adjust: **Bonferroni**

Two-way repeated measures ANOVA

Data preparation

We'll use the `selfesteem2` dataset [in datarium package] containing the self-esteem score measures of 12 individuals enrolled in 2 successive short-term trials (4 weeks): control (placebo) and special diet trials.

Each participant performed all two trials. The order of the trials was counterbalanced and sufficient time was allowed between trials to allow any effects of previous trials to have dissipated.

The self-esteem score was recorded at three time points: at the beginning (t1), midway (t2) and at the end (t3) of the trials.

The question is to investigate if this short-term diet treatment can induce a significant increase of self-esteem score over time. In other terms, we wish to know if there is significant interaction between `diet` and `time` on the self-esteem score.

The two-way repeated measures ANOVA can be performed in order to determine whether there is a significant interaction between diet and time on the self-esteem score.

Load and show one random row by treatment group:

```
# Wide format
set.seed(123)
data("selfesteem2", package = "datarium")
selfesteem2 %>% sample_n_by(treatment, size = 1)
## # A tibble: 2 x 5
##    id    treatment    t1    t2    t3
##    <fct> <fct>      <dbl> <dbl> <dbl>
## 1 4      ctr          92    92    89
## 2 10     Diet         90    93    95
# Gather the columns t1, t2 and t3 into long format.
# Convert id and time into factor variables
selfesteem2 <- selfesteem2 %>%
  gather(key = "time", value = "score", t1, t2, t3) %>%
  convert_as_factor(id, time)
# Inspect some random rows of the data by groups
set.seed(123)
selfesteem2 %>% sample_n_by(treatment, time, size = 1)
## # A tibble: 6 x 4
##    id    treatment time  score
##    <fct> <fct>     <fct> <dbl>
## 1 4      ctr       t1      92
## 2 10     ctr       t2      84
## 3 5      ctr       t3      68
## 4 11     Diet      t1      93
## 5 12     Diet      t2      80
## 6 1      Diet      t3      88
```

In this example, the effect of "time" on self-esteem score is our **focal variable**, that is our primary concern.

However, it is thought that the effect "time" will be different if treatment is performed or not. In this setting, the "treatment" variable is considered as **moderator variable**.

Summary statistics

Group the data by `treatment` and `time`, and then compute some summary statistics of the `score` variable: mean and sd (standard deviation).

```
1  selfesteem2 %>%
2    group_by(treatment, time) %>%
3    get_summary_stats(score, type = "mean_sd")
4  ## # A tibble: 6 x 6
5  ##   treatment time  variable     n  mean    sd
6  ##   <fct>     <fct> <chr>    <dbl> <dbl> <dbl>
7  ## 1 ctr       t1    score       12  88     8.08
8  ## 2 ctr       t2    score       12  83.8  10.2
9  ## 3 ctr       t3    score       12  78.7  10.5
10 ## 4 Diet      t1    score       12  87.6   7.62
11 ## 5 Diet      t2    score       12  87.8   7.42
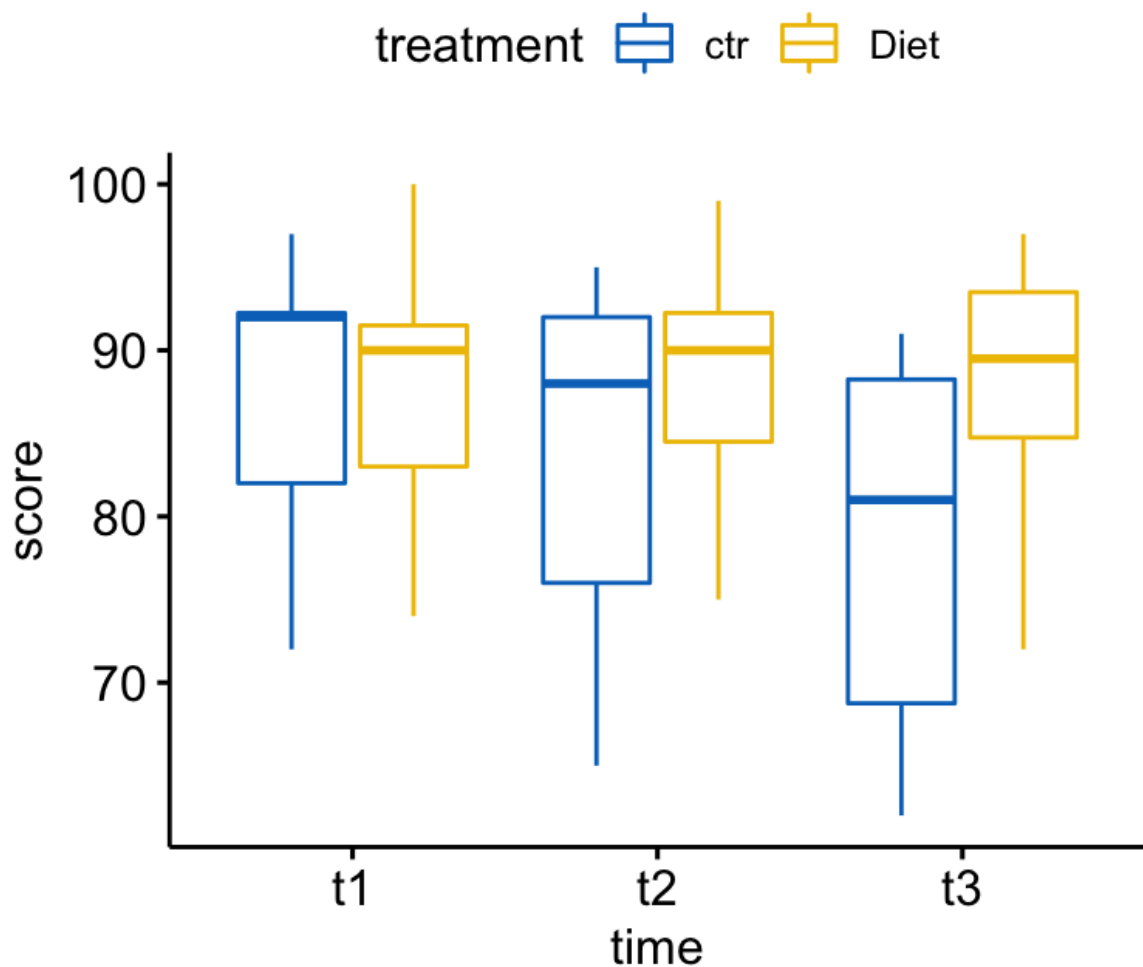12 ## 6 Diet      t3    score       12  87.7   8.14
```

Visualization

Create box plots of the score colored by treatment groups:

```
1  bxp <- ggboxplot(
2    selfesteem2, x = "time", y = "score",
3    color = "treatment", palette = "jco"
4    )
5  bxp
```

Chek assumptions

Outliers

```
1  selfesteem2 %>%
2    group_by(treatment, time) %>%
3    identify_outliers(score)
4  ## [1] treatment  time      id        score     is.outlier is.extreme
5  ## <0 rows> (or 0-length row.names)
```

There were no extreme outliers.

Normality assumption

Compute Shapiro-Wilk test for each combinations of factor levels:

```
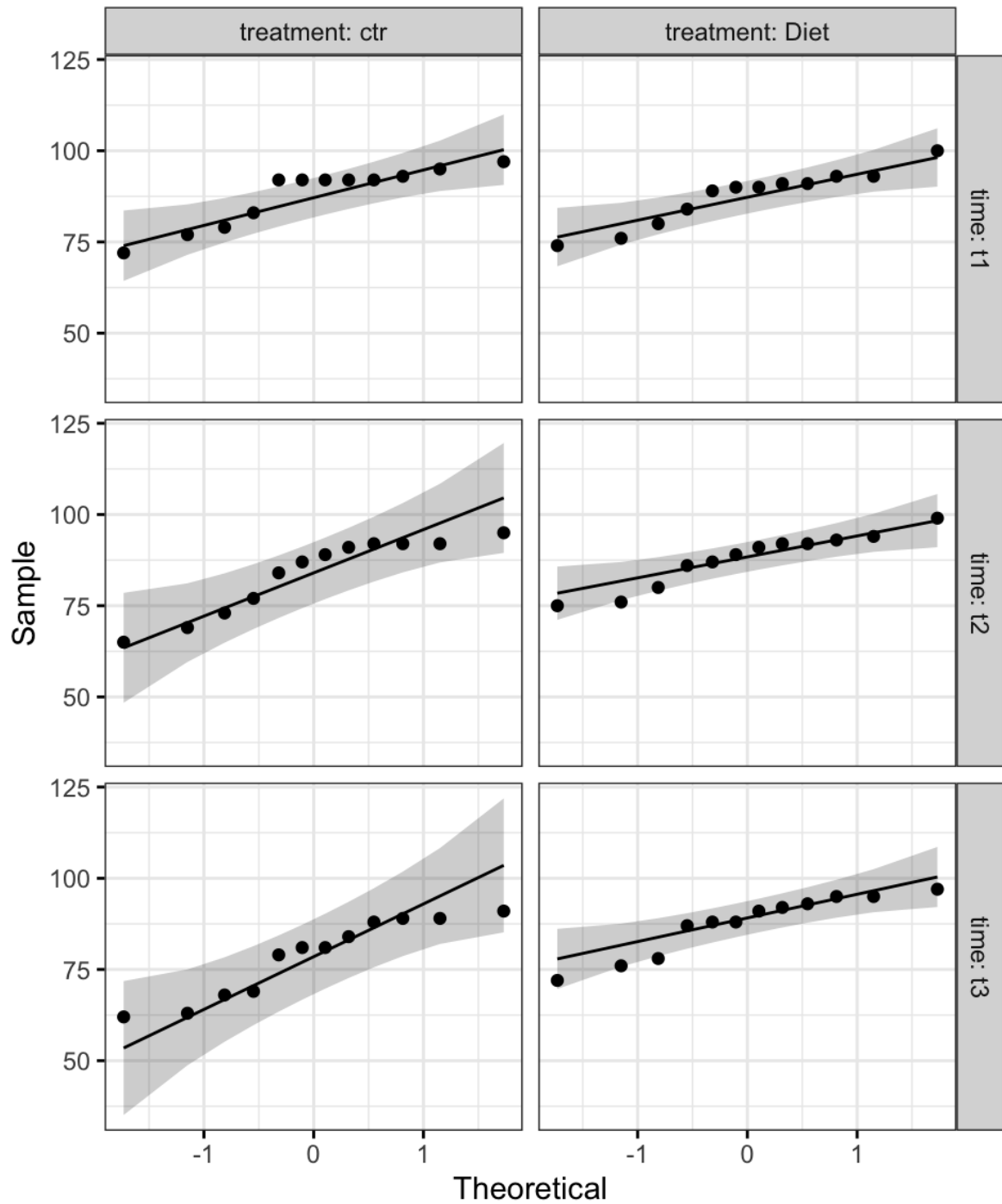1  selfesteem2 %>%
2    group_by(treatment, time) %>%
3    shapiro_test(score)
4  ## # A tibble: 6 x 5
5  ##   treatment time  variable statistic      p
6  ##   <fct>     <fct> <chr>        <dbl>  <dbl>
7  ## 1 ctr       t1    score        0.828 0.0200
8  ## 2 ctr       t2    score        0.868 0.0618
9  ## 3 ctr       t3    score        0.887 0.107
10 ## 4 Diet      t1    score        0.919 0.279
11 ## 5 Diet      t2    score        0.923 0.316
12 ## 6 Diet      t3    score        0.886 0.104
```

The self-esteem score was normally distributed at each time point (p > 0.05), except for ctr treatment at t1, as assessed by Shapiro-Wilk's test.

Create QQ plot for each cell of design:

```
1  ggqqplot(selfesteem2, "score", ggtheme = theme_bw()) +
2    facet_grid(time ~ treatment, labeller = "label_both")
```

From the plot above, as all the points fall approximately along the reference line, we can assume normality.

Computation

```
1   res.aov <- anova_test(
2     data = selfesteem2, dv = score, wid = id,
3     within = c(treatment, time)
4     )
5   get_anova_table(res.aov)
6   ## ANOVA Table (type III tests)
7   ##
8   ##            Effect  DFn   DFd     F        p p<.05    ges
9   ## 1       treatment 1.00 11.0  15.5 2.00e-03     * 0.059
10  ## 2            time 1.31 14.4  27.4 5.03e-05     * 0.049
11  ## 3 treatment:time 2.00 22.0  30.4 4.63e-07     * 0.050
```

There is a statistically significant two-way interactions between treatment and time, F(2, 22) = 30.4, p < 0.0001.

Post-hoc tests

A **significant two-way interaction** indicates that the impact that one factor (e.g., treatment) has on the outcome variable (e.g., self-esteem score) depends on the level of the other factor (e.g., time) (and vice versa). So, you can decompose a significant two-way interaction into:

- **Simple main effect**: run one-way model of the first variable (factor A) at each level of the second variable (factor B),
- **Simple pairwise comparisons**: if the simple main effect is significant, run multiple pairwise comparisons to determine which groups are different.

For a **non-significant two-way interaction**, you need to determine whether you have any statistically significant **main effects** from the ANOVA output.

Procedure for a significant two-way interaction

Effect of treatment. In our example, we'll analyze the effect of `treatment` on self-esteem `score` at every `time` point.

Note that, the `treatment` factor variable has only two levels ("ctr" and "Diet"); thus, ANOVA test and paired t-test will give the same p-values.

```
1   # Effect of treatment at each time point
2   one.way <- selfesteem2 %>%
3     group_by(time) %>%
4     anova_test(dv = score, wid = id, within = treatment) %>%
5     get_anova_table() %>%
6     adjust_pvalue(method = "bonferroni")
7   one.way
8   ## # A tibble: 3 x 9
9   ##   time Effect      DFn   DFd     F      p `p<.05`      ges    p.adj
10  ##   <fct> <chr>     <dbl> <dbl> <dbl>  <dbl> <chr>      <dbl>    <dbl>
11  ## 1 t1    treatment     1    11 0.376 0.552 ""      0.000767 1
12  ## 2 t2    treatment     1    11 9.03  0.012 *       0.052    0.036
13  ## 3 t3    treatment     1    11 30.9  0.00017 *     0.199    0.00051
14  # Pairwise comparisons between treatment groups
15  pwc <- selfesteem2 %>%
16    group_by(time) %>%
17    pairwise_t_test(
18      score ~ treatment, paired = TRUE,
```

```
19      p.adjust.method = "bonferroni"
20      )
21  pwc
22  ## # A tibble: 3 x 11
23  ##   time  .y.   group1 group2   n1    n2 statistic   df       p   p.adj
    p.adj.signif
24  ## * <fct> <chr> <chr>  <chr>  <int> <int>      <dbl> <dbl>    <dbl>    <dbl>
    <chr>
25  ## 1 t1    score ctr    Diet     12    12      0.613   11 0.552    0.552
    ns
26  ## 2 t2    score ctr    Diet     12    12     -3.00    11 0.012    0.012     *

27  ## 3 t3    score ctr    Diet     12    12     -5.56    11 0.00017 0.00017
    ***
```

Considering the Bonferroni adjusted p-value (p.adj), it can be seen that the simple main effect of treatment was not significant at the time point t1 (p = 1). It becomes significant at t2 (p = 0.036) and t3 (p = 0.00051).

Pairwise comparisons show that the mean self-esteem score was significantly different between ctr and Diet group at t2 (p = 0.12) and t3 (p = 0.00017) but not at t1 (p = 0.55).

Effect of time. Note that, it's also possible to perform the same analysis for the `time` variable at each level of `treatment`. You don't necessarily need to do this analysis.

The R code:

```
1   # Effect of time at each level of treatment
2   one.way2 <- selfesteem2 %>%
3     group_by(treatment) %>%
4     anova_test(dv = score, wid = id, within = time) %>%
5     get_anova_table() %>%
6     adjust_pvalue(method = "bonferroni")
7   # Pairwise comparisons between time points
8   pwc2 <- selfesteem2 %>%
9     group_by(treatment) %>%
10    pairwise_t_test(
11      score ~ time, paired = TRUE,
12      p.adjust.method = "bonferroni"
13      )
14  pwc2
```

After executing the R code above, you can see that the effect of `time` is significant only for the control trial, F(2, 22) = 39.7, p < 0.0001. Pairwise comparisons show that all comparisons between time points were statistically significant for control trial.

Procedure for non-significant two-way interaction

If the interaction is not significant, you need to interpret the main effects for each of the two variables: `treatment` and `time`. A significant main effect can be followed up with pairwise comparisons.

In our example (see ANOVA table in `res.aov`), there was a statistically significant main effects of treatment (F(1, 11) = 15.5, p = 0.002) and time (F(2, 22) = 27.4, p < 0.0001) on the self-esteem score.

Pairwise paired t-test comparisons:

```
1   # comparisons for treatment variable
2   selfesteem2 %>%
3     pairwise_t_test(
4       score ~ treatment, paired = TRUE,
5       p.adjust.method = "bonferroni"
6       )
7   # comparisons for time variable
8   selfesteem2 %>%
9     pairwise_t_test(
10      score ~ time, paired = TRUE,
11      p.adjust.method = "bonferroni"
12      )
```

All pairwise comparisons are significant.

Report

We could report the result as follow:

A two-way repeated measures ANOVA was performed to evaluate the effect of different diet treatments over time on self-esteem score.

There was a statistically significant interaction between `treatment` and `time` on self-esteem score, F(2, 22) = 30.4, p < 0.0001. Therefore, the effect of `treatment` variable was analyzed at each `time` point. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of treatment was significant at t2 (p = 0.036) and t3 (p = 0.00051) but not at the time point t1 (p = 1).

Pairwise comparisons, using paired t-test, show that the mean self-esteem score was significantly different between ctr and Diet trial at time points t2 (p = 0.012) and t3 (p = 0.00017) but not at t1 (p = 0.55).

```
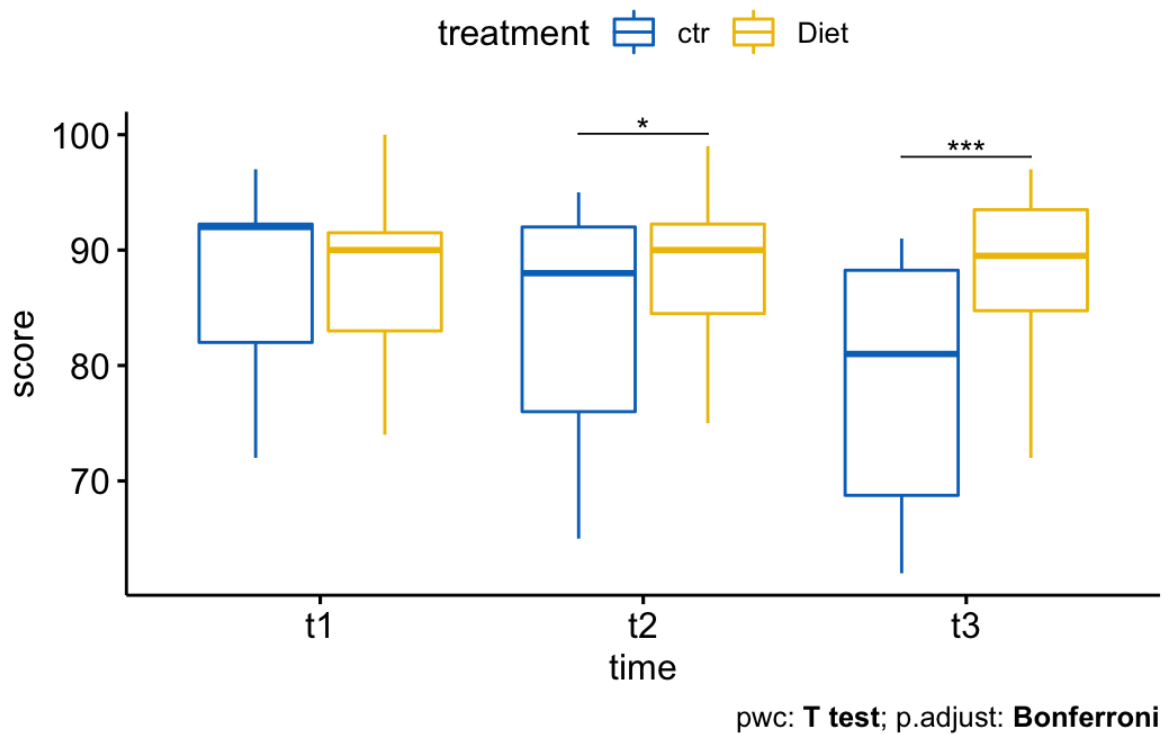1   # Visualization: box plots with p-values
2   pwc <- pwc %>% add_xy_position(x = "time")
3   bxp +
4     stat_pvalue_manual(pwc, tip.length = 0, hide.ns = TRUE) +
5     labs(
6       subtitle = get_test_label(res.aov, detailed = TRUE),
7       caption = get_pwc_label(pwc)
8     )
```

$$\text{Anova, } F(2,22) = 30.42, p = <0.0001, \eta_g^2 = 0.05$$

Three-way repeated measures ANOVA

Data preparation

We'll use the `weightloss` dataset [datarium package]. In this study, a researcher wanted to assess the effects of Diet and Exercises on weight loss in 10 sedentary individuals.

The participants were enrolled in four trials: (1) no diet and no exercises; (2) diet only; (3) exercises only; and (4) diet and exercises combined.

Each participant performed all four trials. The order of the trials was counterbalanced and sufficient time was allowed between trials to allow any effects of previous trials to have dissipated.

Each trial lasted nine weeks and the weight loss score was measured at the beginning (t1), at the midpoint (t2) and at the end (t3) of each trial.

Three-way repeated measures ANOVA can be performed in order to determine whether there is a significant interaction between diet, exercises and time on the weight loss score.

Load the data and show some random rows by groups:

```
1   # wide format
2   set.seed(123)
3   data("weightloss", package = "datarium")
4   weightloss %>% sample_n_by(diet, exercises, size = 1)
5   ## # A tibble: 4 x 6
6   ##   id    diet  exercises    t1    t2    t3
7   ##   <fct> <fct> <fct>     <dbl> <dbl> <dbl>
8   ## 1 4     no    no         11.1   9.5  11.1
9   ## 2 10    no    yes        10.2  11.8  17.4
10  ## 3 5     yes   no         11.6  13.4  13.9
```

```
## 4 11    yes   yes         12.7  12.7  15.1
# Gather the columns t1, t2 and t3 into long format.
# Convert id and time into factor variables
weightloss <- weightloss %>%
  gather(key = "time", value = "score", t1, t2, t3) %>%
  convert_as_factor(id, time)
# Inspect some random rows of the data by groups
set.seed(123)
weightloss %>% sample_n_by(diet, exercises, time, size = 1)
## # A tibble: 12 x 5
##    id    diet  exercises time  score
##    <fct> <fct> <fct>     <fct> <dbl>
## 1 4     no    no        t1    11.1
## 2 10    no    no        t2    10.7
## 3 5     no    no        t3    12.3
## 4 11    no    yes       t1    10.2
## 5 12    no    yes       t2    13.2
## 6 1     no    yes       t3    15.8
## # … with 6 more rows
```

In this example, the effect of the "time" is our **focal variable**, that is our primary concern.

It is thought that the effect of "time" on the weight loss score will depend on two other factors, "diet" and "exercises", which are called **moderator variables**.

Summary statistics

Group the data by `diet`, `exercises` and `time`, and then compute some summary statistics of the `score` variable: mean and sd (standard deviation)

```
weightloss %>%
  group_by(diet, exercises, time) %>%
  get_summary_stats(score, type = "mean_sd")
## # A tibble: 12 x 7
##    diet  exercises time  variable     n  mean    sd
##    <fct> <fct>     <fct> <chr>    <dbl> <dbl> <dbl>
## 1 no    no        t1    score       12  10.9 0.868
## 2 no    no        t2    score       12  11.6 1.30
## 3 no    no        t3    score       12  11.4 0.935
## 4 no    yes       t1    score       12  10.8 1.27
## 5 no    yes       t2    score       12  13.4 1.01
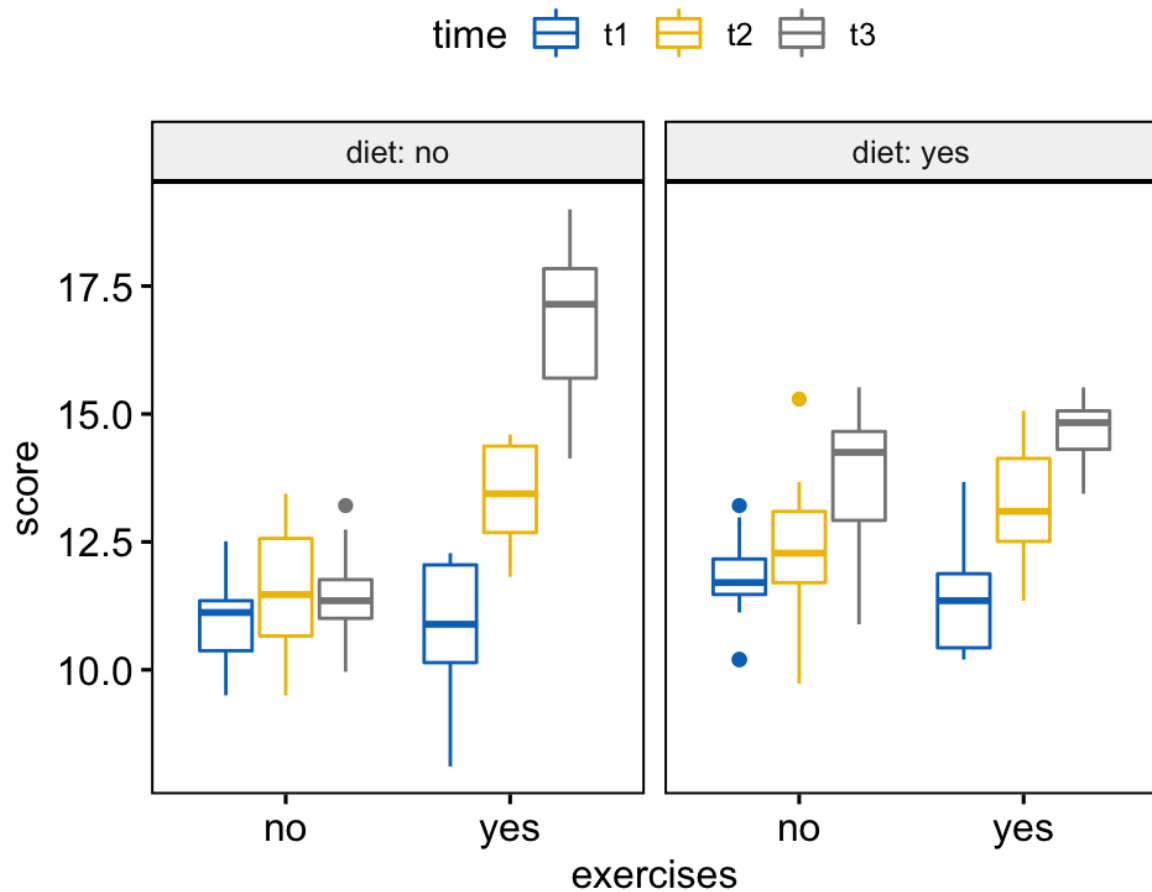## 6 no    yes       t3    score       12  16.8 1.53
## # … with 6 more rows
```

Visualization

Create box plots:

```
bxp <- ggboxplot(
  weightloss, x = "exercises", y = "score",
  color = "time", palette = "jco",
  facet.by = "diet", short.panel.labs = FALSE
  )
bxp
```

Check assumptions

Outliers

```
 1  weightloss %>%
 2    group_by(diet, exercises, time) %>%
 3    identify_outliers(score)
 4  ## # A tibble: 5 x 7
 5  ##   diet  exercises time  id    score is.outlier is.extreme
 6  ##   <fct> <fct>     <fct> <fct> <dbl> <lgl>      <lgl>
 7  ## 1 no    no        t3    2      13.2 TRUE       FALSE
 8  ## 2 yes   no        t1    1      10.2 TRUE       FALSE
 9  ## 3 yes   no        t1    3      13.2 TRUE       FALSE
10  ## 4 yes   no        t1    4      10.2 TRUE       FALSE
11  ## 5 yes   no        t2    10     15.3 TRUE       FALSE
```

There were no extreme outliers.

Normality assumption

Compute Shapiro-Wilk test for each combinations of factor levels:

```
1   weightloss %>%
2     group_by(diet, exercises, time) %>%
3     shapiro_test(score)
4   ## # A tibble: 12 x 6
5   ##    diet  exercises time  variable statistic     p
6   ##    <fct> <fct>     <fct> <chr>        <dbl> <dbl>
7   ## 1 no    no        t1    score        0.917 0.264
8   ## 2 no    no        t2    score        0.957 0.743
9   ## 3 no    no        t3    score        0.965 0.851
10  ## 4 no    yes       t1    score        0.922 0.306
11  ## 5 no    yes       t2    score        0.912 0.229
12  ## 6 no    yes       t3    score        0.953 0.674
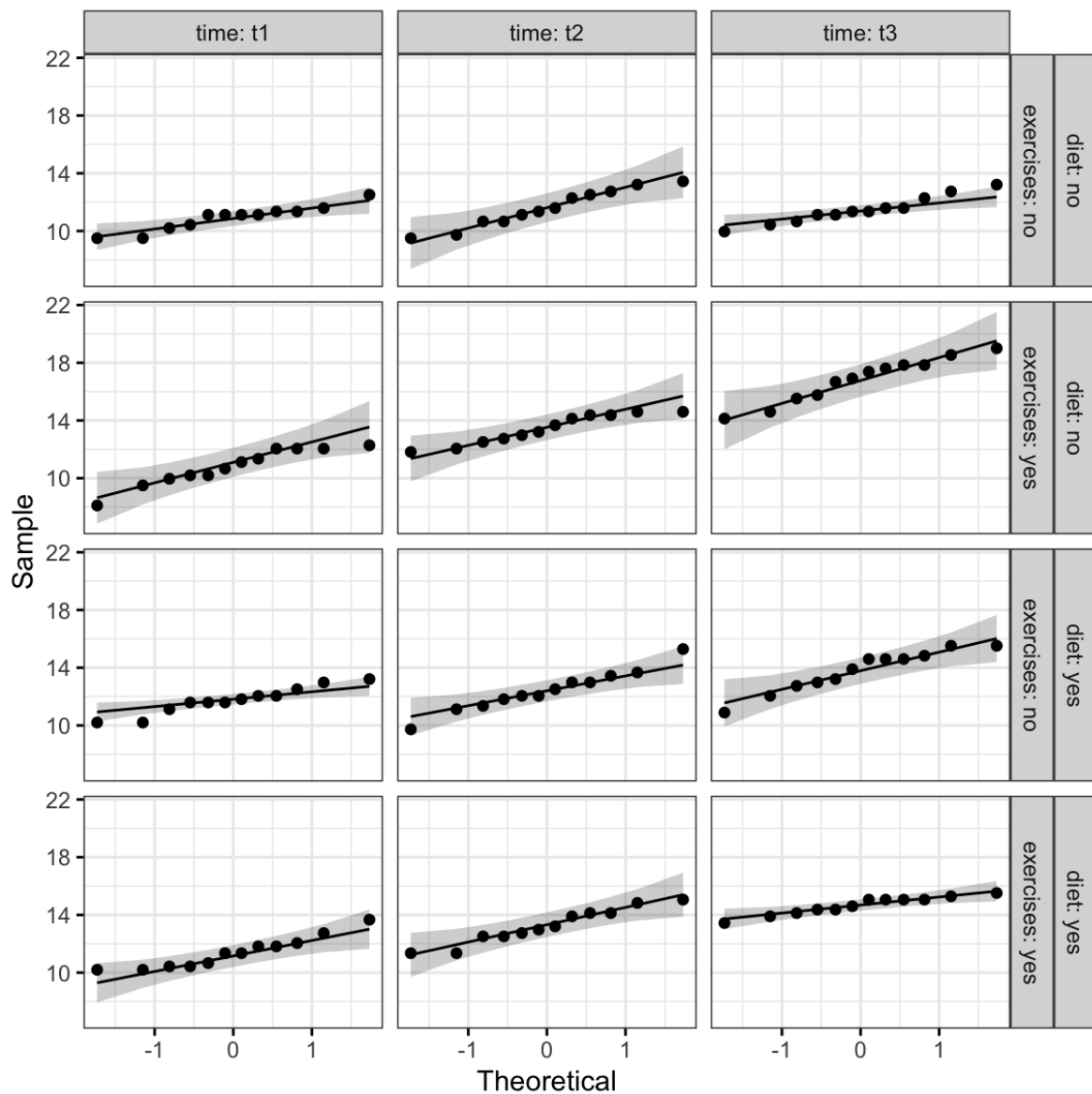13  ## # … with 6 more rows
```

The weight loss score was normally distributed, as assessed by Shapiro-Wilk's test of normality ($p > .05$).

Create QQ plot for each cell of design:

```
1   ggqqplot(weightloss, "score", ggtheme = theme_bw()) +
2     facet_grid(diet + exercises ~ time, labeller = "label_both")
```

From the plot above, as all the points fall approximately along the reference line, we can assume normality.

Computation

```
 1  res.aov <- anova_test(
 2    data = weightloss, dv = score, wid = id,
 3    within = c(diet, exercises, time)
 4    )
 5  get_anova_table(res.aov)
 6  ## ANOVA Table (type III tests)
 7  ##
 8  ##                    Effect  DFn  DFd        F        p p<.05    ges
 9  ## 1                    diet 1.00 11.0    6.021 3.20e-02     * 0.028
10  ## 2                exercises 1.00 11.0   58.928 9.65e-06     * 0.284
11  ## 3                    time 2.00 22.0  110.942 3.22e-12     * 0.541
12  ## 4          diet:exercises 1.00 11.0   75.356 2.98e-06     * 0.157
13  ## 5               diet:time 1.38 15.2    0.603 5.01e-01       0.013
14  ## 6          exercises:time 2.00 22.0   20.826 8.41e-06     * 0.274
15  ## 7 diet:exercises:time 2.00 22.0   14.246 1.07e-04     * 0.147
```

From the output above, it can be seen that there is a statistically significant three-way interactions between diet, exercises and time, $F(2, 22) = 14.24$, $p = 0.00011$.

Note that, if the three-way interaction is not statistically significant, you need to consult the two-way interactions in the output.

In our example, there was a statistically significant two-way diet:exercises interaction ($p < 0.0001$), and two-way exercises:time ($p < 0.0001$). The two-way diet:time interaction was not statistically significant ($p = 0.5$).

Post-hoc tests

If there is a significant three-way interaction effect, you can decompose it into:

- **Simple two-way interaction**: run two-way interaction at each level of third variable,
- **Simple simple main effect**: run one-way model at each level of second variable, and
- **simple simple pairwise comparisons**: run pairwise or other post-hoc comparisons if necessary.

Compute simple two-way interaction

You are free to decide which two variables will form the simple two-way interactions and which variable will act as the third (moderator) variable. In the following R code, we have considered the simple two-way interaction of `exercises*time` at each level of `diet`.

Group the data by `diet` and analyze the simple two-way interaction between `exercises` and `time`:

```
 1  # Two-way ANOVA at each diet level
 2  two.way <- weightloss %>%
 3    group_by(diet) %>%
 4    anova_test(dv = score, wid = id, within = c(exercises, time))
 5  two.way
 6  ## # A tibble: 2 x 2
```

```
 7   ##   diet  anova
 8   ##   <fct> <list>
 9   ## 1 no    <anov_tst>
10   ## 2 yes   <anov_tst>
11   # Extract anova table
12   get_anova_table(two.way)
13   ## # A tibble: 6 x 8
14   ##   diet  Effect          DFn   DFd    F        p `p<.05`   ges
15   ##   <fct> <chr>         <dbl> <dbl> <dbl>    <dbl> <chr>    <dbl>
16   ## 1 no    exercises         1    11 72.8  3.53e- 6 *        0.526
17   ## 2 no    time              2    22 71.7  2.32e-10 *        0.587
18   ## 3 no    exercises:time    2    22 28.9  6.92e- 7 *        0.504
19   ## 4 yes   exercises         1    11 13.4  4.00e- 3 *        0.038
20   ## 5 yes   time              2    22 20.5  9.30e- 6 *        0.49
21   ## 6 yes   exercises:time    2    22  2.57 9.90e- 2 ""       0.06
```

There was a statistically significant simple two-way interaction between exercises and time for "diet no" trial, F(2, 22) = 28.9, p < 0.0001, but not for "diet yes" trial, F(2, 22) = 2.6, p = 0.099.

Note that, it's recommended to adjust the p-value for multiple testing. One common approach is to apply a Bonferroni adjustment to downgrade the level at which you declare statistical significance.

This can be done by dividing the current level you declare statistical significance at (i.e., p < 0.05) by the number of simple two-way interaction you are computing (i.e., 2).

Thus, you only declare a two-way interaction as statistically significant when p < 0.025 (i.e., p < 0.05/2). Applying this to our current example, we would still make the same conclusions.

Compute simple simple main effect

A statistically significant simple two-way interaction can be followed up with **simple simple main effects**.

In our example, you could therefore investigate the effect of `time` on the weight loss `score` at every level of `exercises` or investigate the effect of `exercises` at every level of `time`.

You will only need to consider the result of the simple simple main effect analyses for the "diet no" trial as this was the only simple two-way interaction that was statistically significant (see previous section).

Group the data by `diet` and `exercises`, and analyze the simple main effect of `time`. The Bonferroni adjustment will be considered leading to statistical significance being accepted at the p < 0.025 level (that is 0.05 divided by the number of tests (here 2) considered for "diet:no" trial.

```
 1   # Effect of time at each diet X exercises cells
 2   time.effect <- weightloss %>%
 3     group_by(diet, exercises) %>%
 4     anova_test(dv = score, wid = id, within = time)
 5   time.effect
 6   ## # A tibble: 4 x 3
 7   ##   diet  exercises anova
 8   ##   <fct> <fct>     <list>
 9   ## 1 no    no        <anov_tst>
10   ## 2 no    yes       <anov_tst>
11   ## 3 yes   no        <anov_tst>
```

```
12  ## 4 yes    yes         <anov_tst>
13  # Extract anova table
14  get_anova_table(time.effect) %>%
15    filter(diet == "no")
16  ## # A tibble: 2 x 9
17  ##   diet  exercises Effect   DFn   DFd     F         p `p<.05`    ges
18  ##   <fct> <fct>     <chr>  <dbl> <dbl> <dbl>     <dbl> <chr>    <dbl>
19  ## 1 no    no        time       2    22  1.32 2.86e- 1 ""       0.075
20  ## 2 no    yes       time       2    22 78.8  9.30e-11 *        0.801
```

There was a statistically significant simple simple main effect of time on weight loss score for "diet:no,exercises:yes" group (p < 0.0001), but not for when neither diet nor exercises was performed (p = 0.286).

Compute simple simple comparisons

A statistically significant simple simple main effect can be followed up by **multiple pairwise comparisons** to determine which group means are different.

Group the data by `diet` and `exercises`, and perform pairwise comparisons between `time` points with Bonferroni adjustment:

```
1   # Pairwise comparisons
2   pwc <- weightloss %>%
3     group_by(diet, exercises) %>%
4     pairwise_t_test(score ~ time, paired = TRUE, p.adjust.method =
    "bonferroni") %>%
5     select(-df, -statistic) # Remove details
6   # Show comparison results for "diet:no,exercises:yes" groups
7   pwc %>% filter(diet == "no", exercises == "yes") %>%
8     select(-p)     # remove p columns
9   ## # A tibble: 3 x 9
10  ##   diet  exercises .y.   group1 group2    n1    n2        p.adj
    p.adj.signif
11  ##   <fct> <fct>     <chr> <chr>  <chr>  <int> <int>        <dbl> <chr>

12  ## 1 no    yes       score t1     t2        12    12 0.000741       ***

13  ## 2 no    yes       score t1     t3        12    12 0.0000000121 ****

14  ## 3 no    yes       score t2     t3        12    12 0.000257       ***
```

In the pairwise comparisons table above, we are interested only in the simple simple comparisons for "diet:no,exercises:yes" groups. In our example, there are three possible combinations of group differences. We could report the pairwise comparison results as follow.

All simple simple pairwise comparisons were run between the different time points for "diet:no,exercises:yes" trial. The Bonferroni adjustment was applied. The mean weight loss score was significantly different in all time point comparisons when exercises are performed (p < 0.05).

Report

A three-way repeated measures ANOVA was performed to evaluate the effects of diet, exercises and time on weight loss. There was a statistically significant three-way interaction between diet, exercises and time, $F(2, 22) = 14.2$, $p = 0.00011$.

For the simple two-way interactions and simple simple main effects analyses, a Bonferroni adjustment was applied leading to statistical significance being accepted at the $p < 0.025$ level.

There was a statistically significant simple two-way interaction between exercises and time for "diet no" trial, $F(2, 22) = 28.9$, $p < 0.0001$, but not for "diet yes"" trial, $F(2, 22) = 2.6$, $p = 0.099$.

There was a statistically significant simple simple main effect of time on weight loss score for "diet:no,exercises:yes" trial ($p < 0.0001$), but not for when neither diet nor exercises was performed ($p = 0.286$).

All simple simple pairwise comparisons were run between the different time points for "diet:no,exercises:yes" trial with a Bonferroni adjustment applied. The mean weight loss score was significantly different in all time point comparisons when exercises are performed ($p < 0.05$).

```
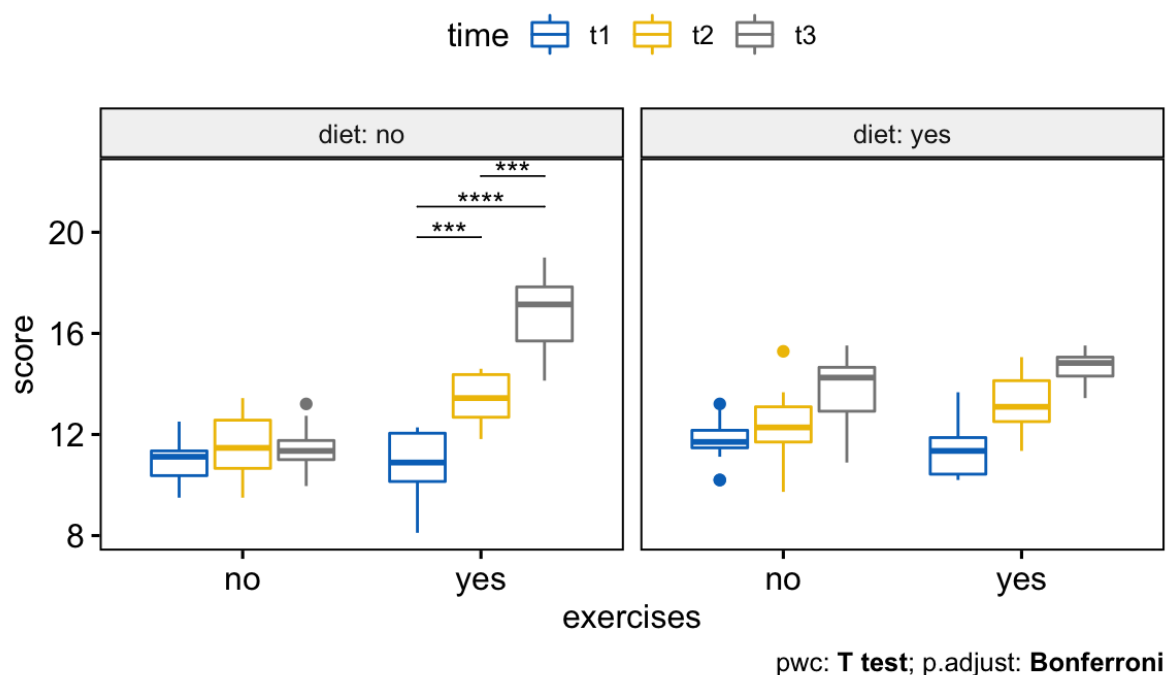1   # Visualization: box plots with p-values
2   pwc <- pwc %>% add_xy_position(x = "exercises")
3   pwc.filtered <- pwc %>%
4     filter(diet == "no", exercises == "yes")
5   bxp +
6     stat_pvalue_manual(pwc.filtered, tip.length = 0, hide.ns = TRUE) +
7     labs(
8       subtitle = get_test_label(res.aov, detailed = TRUE),
9       caption = get_pwc_label(pwc)
10    )
```

Anova, $F(2,22) = 14.25$, $p = 0.00011$, $\eta_g^2 = 0.15$

pwc: **T test**; p.adjust: **Bonferroni**

Summary

This chapter describes how to compute, interpret and report repeated measures ANOVA in R. We also explain the assumptions made by repeated measures ANOVA tests and provide practical examples of R codes to check whether the test assumptions are met.