

# V. Applied Statistics in R (presentation)

16 June, 2022

## Structure of a biostatistical analysis in R

For questions and comments - please use the padlet here: <https://ucph.padlet.org/tugcekaraderi/kyk81qgc1l6qtng>

The very basic structure of an R script doing a classical statistical analysis is as follows:

- Load packages that you will be using.
- Read the dataset to be analyzed. Possibly also do some data cleaning and manipulation.
- Visualize the dataset by graphics and other descriptive statistics.
- Fit and validate a statistical model.
- Hypothesis testing. Possibly also post hoc testing.

Of course there are variants of this set-up, and in practice there will often be some iterations of the steps. In this manuscript we will exemplify the proposed steps in the analysis of a simple dataset:

- In our current scenario, you are a researcher investigating psoriasis, an inflammatory skin disease. You have data on the expression of a number of genes that are suspected to have something to do with the disease, but you cannot be sure until you perform some formal statistical analysis.
- This is a great example where R skills would come very handy!
- You will start with your gene of special interest IGFL4. IGFL4 belongs to the insulin-like growth factor family of signaling molecules that play critical roles in cellular energy metabolism as well as in growth and development.
- You decide that your analysis approach will be one-way ANOVA of the expression of the IGFL4 gene against the skin type in psoriasis patients.

**Load packages** We will use **ggplot2** to make plots, and to be prepared for data manipulations, we simply load this together with the rest of the **tidyverse**.

The psoriasis data are provided in an Excel sheet, so we also load **readxl**. Finally, we will use the package **emmeans** to make post hoc tests.

Remember that you should install the wanted packages before they can be used (but you only need to install the packages once!).

Thus,

```
#install.packages("tidyverse")
#install.packages("readxl")
#install.packages("emmeans")
```

```
library(tidyverse)
library(readxl)
library(emmeans)
```

Now, we are done preparing for our analyses. Next, we will look specifically at the possible association between IGFL4 gene expression and psoriasis. Finally, we conclude with a brief outlook on other statistical models in R.

Please refer to the ‘STATS CHEAT SHEET’ provided in the slides for hints as well as other cheat sheets provided in other sessions where necessary.

## Example: Analysis of variance

**Step 1: Data** Psoriasis is an immune-mediated disease that affects the skin. You, as a researcher, carried out a micro-array experiment with skin from 37 people in order to examine a potential association between the disease and a certain gene (IGFL4). For each of the 37 samples the gene expression was measured. Fifteen skin samples were from psoriasis patients and from a part of the body affected by the disease (**psor**); 15 samples were from psoriasis patients but from a part of the body not affected by the disease (**psne**); and 7 skin samples were from healthy people (**control**).

The data are saved in the file **psoriasis.xlsx**. At first the variable named **type** (i.e. skin sample type) is stored as a character variable, we change it to a factor (and check that indeed there are 15, 15 and 7 skin samples in the three groups).

```
# Read in the data from Excel file and call it psoriasisData
# Start with command: psoriasisData <- read_excel()

# View the top rows of the dataset
# Use code line: head(psoriasisData)

# Extract the data of interest containing the IGFL4 expression levels and skin sample type from the data
# Use code line: psorData <- select(psoriasisData, type, IGFL4)

# View the top rows of the dataset psorData

# Change variable named 'type' to factor so that we can use in our analysis in the next steps
# First let's check if it is character
# Use code line: is.character(psorData$type)

# Now change to factor
# Use code line: psorData <- mutate(psorData, type = factor(type))

# Again, let's check if it is factor now

# Check that there are 15, 15 and 7 skin samples in the three groups. Hint: count()
```

QUESTION 1: Are there 15, 15 and 7 skin samples in the three groups? Please share your answer here: [menti.com 5103 9109](https://www.menti.com/51039109)

**Step 2: Descriptive plots and statistics** To get an impression of the data, we make two plots, and compute group-wise means and standard deviations. Make (1) a plot showing three groups of skin samples

(healthy, psne, psor) and IGFL4 expression levels from each skin sample so that we can visually compare each group of samples with regards to IGFL4 expression levels; (2) a boxplot for the same group-wise comparison. And finally, obtain the group-wise (healthy, psne, psor) descriptive statistics (mean, median and standard deviation). Hint: Consider using ggplot and tidyverse!

```
# (1) a plot showing three groups of skin samples (healthy, psne, psor) and IGFL4 expression levels from
# Use code line: ggplot(psorData, aes(x=type, y=IGFL4)) + geom_point() + labs(x="Skin type", y="IGFL4")

# (2) a boxplot for the same group-wise comparison. Hint: geom_boxplot()

# And finally, obtain the group-wise (healthy, psne, psor) descriptive statistics (mean, median and sta
```

QUESTION 2: Please rank the three groups by their mean/average IGFL4 expression (descending order). Please share your answer here: [menti.com 5103 9109](https://www.menti.com/51039109)

**Step 3: Fit of oneway ANOVA, model validation** The scientific question is whether the gene expression level of IGFL4 differs between the three types/groups. Thus, the natural type of analysis is a oneway analysis of variance (ANOVA). The oneway ANOVA is fitted with the `lm()` function in R. It is a good approach to assign a name (below *oneway*) to the object with the fitted model. This object contains all relevant information and may be used for subsequent analysis. Note that we need to logarithmic transform the response (i.e. IGFL4 expression levels) as intensities are often on a multiplicative scale.

```
# oneway analysis of variance (ANOVA)
# Use code line: oneway <- lm(log(IGFL4) ~ type, data=psorData)

# View the model coefficients
```

QUESTION 3: Are you done? Please share how you feel here: [menti.com 5806 6079](https://www.menti.com/58066079)

**Step 4: Hypothesis test + Post hoc tests** It is standard to carry out an  $F$ -test for the overall effect of the explanatory (i.e. independent) variable. To be precise, the hypothesis is that the expected values are the same in *all* groups. The most easy way to do this test is to use `drop1`. The option `test="F"` is needed to get the  $F$ -test using that function:

```
# Carry out an $F$-test for the overall effect of the explanatory (i.e. independent) variable
```

QUESTION 4: What is your conclusion from the  $F$ -test result? Please share your answer here: [menti.com 3783 8151](https://www.menti.com/37838151)

It might be that the gene expression in two of the three groups may or may not be significantly different. To investigate that we do post hoc testing. This is nicely done within the framework of *estimated marginal means* using the **emmeans** package. Here **emmeans** makes the estimated marginal means (that is the predicted gene expression IGFL4 on the log scale), and the `pairs()` command provide post hoc pairwise comparisons (package automatically adjusts for multiple comparisons using the default tukey method):

```
# Run emmeans
# Use code line: emmeans(oneway,~type)

# Perform post hoc pairwise comparisons
```

QUESTION 5: What is your conclusion from the pairwise comparisons? Please share your answer here: [menti.com 3783 8151](https://www.menti.com/37838151)