A ONE-WAY ANOVA TUTORIAL

A ONE-WAY ANOVA TUTORIAL

Dirk Ostwald
Institute of Psychology
Otto-von-Guericke-Universität Magdeburg

August 23, 2025

1 Introduction

Analysis of variance (ANOVA) is an umbrella term for a large variety of interwoven experimental designs and data-analytical methods that aim to quantify the effects of interventions. As the name implies, ANOVA in its traditional sense is concerned with analyzing data variability, seeking to partition it into variation attributable to experimental manipulations and variation arising from unsystematic, non-interventional sources. ANOVA was popularized a century ago by Ronald A. Fisher as an inference method to assess the evidence for an effect of experimental intervention when comparing a number of treatment groups (Fisher (1925)). Since then, experimental designs have diversified tremendously and along with them the data-analytical approaches under the ANOVA label. The form of ANOVA we are concerned with in the current tutorial is typically referred to as *one-way ANOVA* and applies in the context of a single experimental factor, such as different and unrelated treatment groups. In the following, we use the term ANOVA to exclusively refer to the one-way ANOVA scenario. In Section 2, we discuss the traditional computation-oriented approach to ANOVA.

Over the course of the 20th century, ANOVA has become regarded as a special linear model that describes how data in an ANOVA experimental design is generated. This perspective enables experimental effects to be expressed through model parameters while also providing a framework for quantifying the uncertainty associated with the resulting parameter inferences. While a comprehensive introduction of this (general) linear model approach to statistical inference is beyond the scope of this tutorial, we discuss some of its aspects in Section 3.

Finally, over the course of the last forty years, the digital revolution has allowed the development of computerized algorithms that evaluate parameter estimates numerically within seconds. This in turn has allowed the development of ANOVA models that are more flexibly adapted to specific experimental contexts and can directly capture structural context information of no interest to the evaluation of experimental effects. Broadly speaking, these methods come under the label of *linear mixed models* and dominate the applied use of ANOVA approaches in today's research. In Section 4 we aim to provide some intuition about these approaches in the context of randomized block and nested subsampling designs.

Throughout, we assume familiarity with the sum symbol (for a brief refresher, see Section 6.1). We eschew introducing the matrix formulation of linear models, which in fact provides the formal language in which such models are most precisely expressed. For the essential probabilistic concepts, we mainly appeal to intuition. However, please note that a proper understanding of modern ANOVA can only be achieved based on firm foundations in linear algebra and probability theory.

1.1 An example scenario

As a first concrete data analysis scenario, we consider the effect of four different *treatments* that are assessed based on 10 *observational units* each. The observational units that are exposed to one of the treatments will also be referred to as a *(treatment) group*. For the example, we assume $p = 4$ treatment groups of $m = 10$ observational units per group and thus $n = p \cdot m = 4 \cdot 10 = 40$ observational units with associated univariate measurements in total. In the language of (factorial) experimental designs the treatments form what is called an *experimental factor* and the different variations that the treatment can take are referred to as *levels* of the group factor. We label the treatments generically as `T1`, `T2`, `T3`, `T4`.

For a biological example, the observational units could be fibroblasts from a given cell culture strain, say `S1` and the measurements could be the amount of collagen fiber produced by each fibroblast within 24 hours measured by some optical method. The treatments could correspond to different cytokines that may or may not increase the production of collagen fiber production. Table 1 shows a set of exemplary measurements.

Table 1: Example data set 1

T1	T2	T3	T4
4.6	10.5	5.7	2.8
4.0	11.4	6.8	6.4
3.4	9.0	5.1	2.6
2.0	10.8	6.5	4.2
4.0	5.8	1.0	3.2
3.7	10.0	8.1	2.9
4.5	13.4	3.2	1.0
6.9	12.8	4.3	5.8
5.6	7.9	6.7	3.3
6.3	13.6	5.7	1.4

The data set of Table 1 is summarized in terms of its treatment group means and treatment group standard deviations and visualized in Figure 1 using the **R** code below. Note that all individual data points are shown on the figure. The visualization uses the Base R Graphics functionality and does not require additional packages for full control of all figure properties. The different functions can be studied using `?functioname` in **R** or with the help of one's favorite general purpose large language model, such as ChatGPT.

```
# data loading and group mean and standard deviation evaluation
D       = read.csv("./_data/anova-data-set-1.csv")                    # dataframe in long format
D       = data.frame(split(D$Collagen, D$Treatment))                  # wide format organization
gmeans  = colMeans(D)                                                 # group means
gstds   = apply(D,2,sd)                                               # group standard deviations

# figure formatting
pdf(                                                                  # figure file format
file      = "./_figures/anova-data-set-1.pdf",                        # figure file path and name
width     = 7,                                                        # figure width in inches
height    = 5)                                                        # figure height in inches
par(                                                                  # figure parameter
mfcol     = c(1,1),                                                   # 1 x 1 subplot layout
family    = "sans",                                                   # fontfamily (sans serif)
pty       = "m",                                                      # plot area (maximal not square)
bty       = "l",                                                      # axes in L form
mgp       = c(2,1,0),                                                 # margin spacing
font.main = 1,                                                        # title font
lwd       = 1,                                                        # linewidth
las       = 1)                                                        # orientiation of x-ticks

# bars
```

```
x         = barplot(                                                    # barplot and barlocations
gmeans,                                                                 # group mean bars
ylim      = c(0,16),                                                    # y-axis limits
col       = "gray90",                                                   # bar color
ylab      = "Collagen",                                                 # y-axis label
xlab      = "Treatment",                                                # x-axis label
main      = "Data Set 1")                                               # title

# standard deviation error bars
arrows(                                                                 # errorbars
x0        = x,                                                          # x-ordinate
y0        = gmeans - gstds,                                             # y-ordinate, lower end
x1        = x,                                                          # x-ordinate
y1        = gmeans + gstds,                                             # y-ordinate upper end
code      = 3,                                                          # caps on both ends
angle     = 90,                                                         # cap
length    = 0.05)                                                       # cap length

# data points
for(i in 1:ncol(D)){                                                    # data point iterations
points(                                                                 # point plot
jitter(rep(x[i], length(D[,i])), amount = .3),                          # left/right jitter for overlap avoidance
D[,i],                                                                  # data points of interest
pch     = 21,                                                           # point type
col     = "White",                                                      # outline color
bg      = "Black")}                                                     # fill color
dev.off()                                                               # pdf closing
```

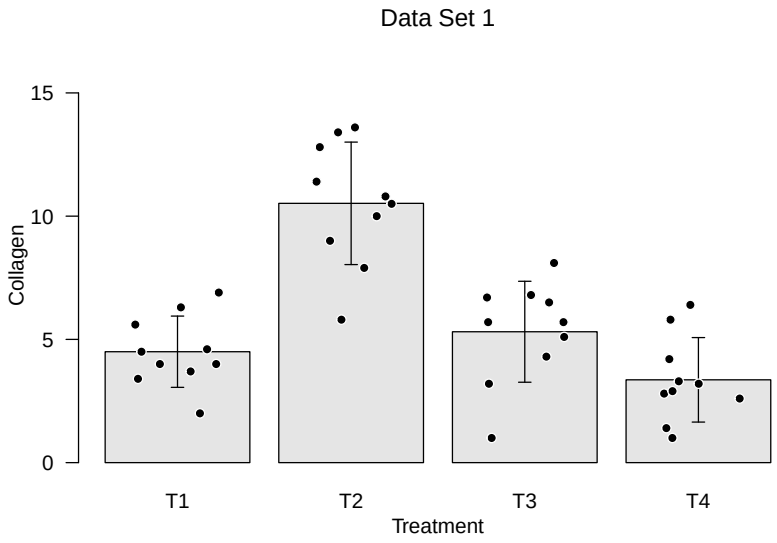

Figure 1: Barplot visualization of data set 1. Each dot represents the collagen production of a single fibroblast undergoing one of the treatment levels

As is evident from Figure 1, Treatment 1 (T1) and Treatment 3 (T3) induce similar average collagen production per fibroblast. Treatment 2 (T2) induces a marked increase collagen production and Treatment 4 (T4) induces less collagen production.

For comparison purposes, we also consider a second example data set generated using fibroblasts from a second cell culture strain S2 but otherwise identical to the first example data set. We tabulate this data set in Table 2 and visualize it in Figure 2. Obviously, in this data set, the treatment levels do not induce differences in the average collagen production per fibroblast.

Table 2: Example data set 2

T1	T2	T3	T4
4.6	10.5	5.7	2.8
4.0	11.4	6.8	6.4
3.4	9.0	5.1	2.6
2.0	10.8	6.5	4.2
4.0	5.8	1.0	3.2
3.7	10.0	8.1	2.9
4.5	13.4	3.2	1.0
6.9	12.8	4.3	5.8
5.6	7.9	6.7	3.3
6.3	13.6	5.7	1.4

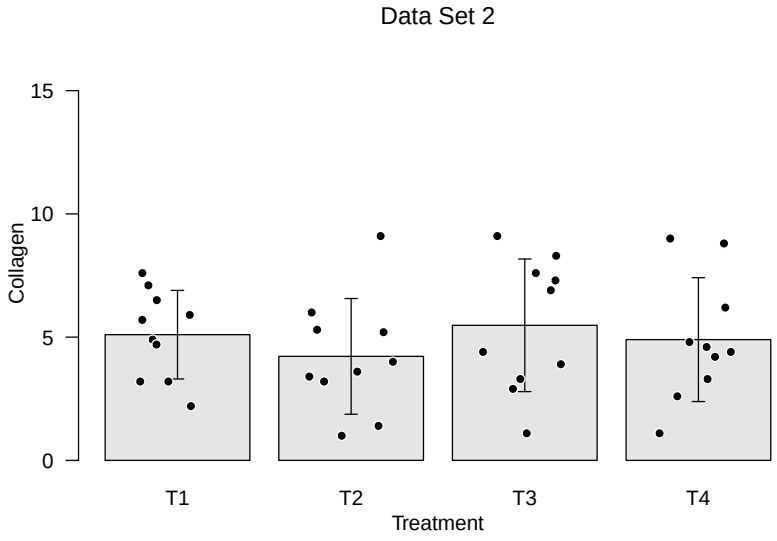

Figure 2: Barplot visualization of data set 2. Each dot represents the collagen production of a single fibroblast undergoing one of the treatment levels

2 ANOVA as a computational procedure

Empirical variance. A simple measure of the variability of n scalar data points $x_1, ..., x_n$ is their *empirical variance*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2, \tag{1}$$

where $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$ denotes the *mean* of the data points. Conceptually, s^2 in Equation 1 is computed by first computing the mean $\bar{x}$ of all data points as a measure of their central tendency. Next, the difference $x_i - \bar{x}$ between each data point x_i and the the mean $\bar{x}$ of all data points is evaluated. If the data point is larger than the mean, this yields a positive number. If the data point is smaller than the mean, this yields a negative number. If the data point and the mean are virtually identical, this yields a number close to zero. To summarize the n deviations from the mean into a single number, the differences $x_i - \bar{x}$ are added up. However, simply summing the differences $x_i - \bar{x}$ for

$i = 1, ..., n$ will have the effect that positive and negative deviations from the mean cancel each other out. To prevent this, each difference is squared. The square of any number, positive or negative, is always non-negative. Squaring the differences, i.e., evaluating $(x_i - \bar{x})^2$ for $i = 1, ...n$ thus has the effect of making all terms in the sum of Equation 1 positive. It addition, large deviations between a data point and the mean contribute a lot to the sum, whereas small deviations contribute little. For example, a diffence $x_i - \bar{x} = -5$, after squaring, contributes 25, whereas a difference $x_i - \bar{x} = 0.5$, after squaring contributes only 0.25. Evaluation of the sum of squared deviations then yields a single number that represents the variability of the data points about their mean. In Equation 1, this number is finally standardized by division by $n-1$, i.e., the variability is expressed roughly as an average per data point. It would be the exact average, if instead of dividing by $n - 1$, the division would be by n. However, for a large number of data points, this difference is virtually negligible, as e.g., for $n = 100$, $\frac{1}{100} = 0.01$ and $\frac{1}{100-1} = 0.0099 \approx 0.01$. The motivation of the denominator $n - 1$ is best understood in the context of designing *unbiased estimators*, a concept which is beyond the scope of this tutorial. In general, expressions such as

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

in Equation 1 are called *sum-of-squares*, as they obviously represent sums of squared values. Similarly, expressions such as

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

are called *mean sum-of-squares* as they represent averages of sum of squares.

Between-sum-of-squares and within-sum-of-squares. The fundamental idea of ANOVA as a computational procedure is to use sum-of-squares to quantify the variability observed in an ANOVA data set. To this end, the data variability is partitioned in a way that allows for inferring how much of it is likely due to the experimental treatment factor and how much of it is due to unsystematic variation. The approach is rather simplistic. On the one hand, ANOVA evaluates the variability of the individual group means around the grand mean across all observational units in all groups as a measure of potential treatment-induced effects. On the other hand, ANOVA evaluates the average variability of individual data points about their group mean across groups as a measure of unsystematic variation. Note that the latter variability cannot be treatment-induced, as each data point in a given group was exposed to exactly the same treatment level (if the experiment was performed adequately). Both measures of variability are single numbers and are referred to as *between-sum-of-squares* (variability between groups) and *within-sum-of-quares* (variability within groups), respectively. Finally, asking how much bigger or smaller the mean between-sum-of-squares is than the mean within-sum-of-squares by dividing the former by the latter yields a single number that indicates the size of the experimental treatment-induced variability in units of unsystematic variability. This number is referred to as the F*-value* or F*-value* in commemoration of Ronald A. Fisher, who famously described this computational procedure in his book *Statistical Methods for Research Workers* a century ago (Fisher (1925)).

To represent this logic formally, it is first necessary to introduce some general notation for the data of a one-way ANOVA design. We here use the convention, that y_{ij} denotes the jth of n_i data points in the ith of p groups. Using this convention, the example data sets of Table 1 and Table 2 can be denoted as shown in Table 3.

Table 3: One-way ANOVA notation. In this scenario $p = 4$, $n_1 = n_2 = n_3 = n_4 = 10$ and $n = \sum_{i=1}^{p} n_i = 40$

T1	T2	T3	T4
y_{11}	y_{21}	y_{31}	y_{41}
y_{12}	y_{22}	y_{32}	y_{42}
y_{13}	y_{23}	y_{33}	y_{43}
y_{14}	y_{24}	y_{34}	y_{44}
y_{15}	y_{25}	y_{35}	y_{45}
y_{16}	y_{26}	y_{36}	y_{46}
y_{17}	y_{27}	y_{37}	y_{47}
y_{18}	y_{28}	y_{38}	y_{48}
y_{19}	y_{29}	y_{39}	y_{49}
y_{110}	y_{210}	y_{310}	y_{410}

The *between-sum-of-square* is then defined as

$$\text{SQB} := \sum_{i=1}^{p} n_i \left(\bar{y}_i - \bar{y}\right)^2, \tag{2}$$

where

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij} \tag{3}$$

denotes the grand mean (mean across all data points in all groups) and

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{4}$$

denotes the mean of the data points in the ith group. Large values of SQB imply a strong dependence of the group sample means on the respective treatment level i, while small values of SQB imply a weak dependence on the respective treatment level. In this sense, SQB represents the data variability explained by considering the respective treatment level. Note that SQB formula weighs the squared deviation between the ith group mean and the grand mean by the number of data points n_i in the ith group. Deviations in larger groups thus contribute more to the SQB than deviations in smaller groups. In *balanced designs*, for which by definition all group sizes are identical, all groups contribute equally.

The *within-sum-of-squares* is defined as

$$\text{SQW} := \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2.$$

Here, the sum of squared deviations of all data points from their group-specific mean within a given group $\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$ is summed across all p groups. SQW thus represents the data variability summed across all treatment levels that remains after the variability in the ith group has been explained by its respective group mean. Note that in the SQB the factor n_i effectively scales the squared deviation between a group mean and the grand mean to account for the missing summation over $j = 1, ..., n_i$ in SQW.

Mean-sum-of-squares and F-value. Dividing the SQB and the SQW by the number of contributing data points to obtain an average squared deviation per data point yields the so-called *mean between-sum-of-squares* and *mean-*

between-sum-of-squares

$$\text{MSB} := \frac{\text{SQB}}{p-1} \text{ and } \text{MSW} := \frac{\text{SQW}}{n-p}. \tag{5}$$

The denominators in Equation 5 can be understood by realizing, that approximately p "data points", i.e. group means contribute to the MSB and approximately n data points contribute to the SQW. The exact numbers in the denonminators are referred to as *degrees of freedom*, such that the *between-group degrees of freedom (DFB)* are given by $p-1$ and the *within-group degrees of freedom (DFW)* are given by $n-p$.

The F*-value* is defined as the ratio of MSB and MSW,

$$F = \frac{\text{MSB}}{\text{MSW}}. \tag{6}$$

F-values larger than 1 thus indicate that the MSB is larger than the MSW. In this sense, F-values larger than 1 thus indicate that the averaged treatment-induced group mean variability is larger than the data point variability at a constant treatment level. The following **R** code evaluates the necessary quantities for computing the F-value for the example data sets of Table 1 and Figure 1.

```
D       = read.csv("./_data/anova-data-set-1.csv")                          # dataframe in long format
Y       = as.matrix(data.frame(split(D$Collagen, D$Treatment)))             # wide format as simple matrix
y_bar   = mean(Y)                                                           # grand mean
y_bar_i = colMeans(Y)                                                       # group means
DFB     = ncol(Y) - 1                                                       # p -1
DFW     = length(Y) - ncol(Y)                                               # n - p
SQB_i   = (y_bar_i - y_bar)^2                                               # SQB terms
SQB     = sum(SQB_i * nrow(Y))                                              # SQB
SQW_i   = colSums((Y-matrix(y_bar_i,nrow(Y),length(y_bar_i), byrow = TRUE))^2) # SQW terms
SQW     = sum(SQW_i)                                                        # SQW
MSB     = SQB/DFB                                                           # MSB
MSW     = SQW/DFW                                                           # MSW
Eff     = MSB/MSW                                                           # F-value
```

The results are as follows.

```
y_bar   : 5.923
y_bar_i : 4.5 10.52 5.31 3.36
DFB     : 3
DFW     : 36
SQB_i   : 2.024 21.137 0.375 6.566
SQB     : 301.021
SQW_i   : 18.82 55.556 37.749 26.444
SQW     : 138.569
MSB     : 100.34
MSW     : 3.849
F       : 26.068
```

Obviously, the group means deviate significantly from the grand mean, while the deviation of each data point from its respective group mean is not too large. Together, this results in an F-value that is a lot larger than 1.

For the example data set of Table 2 and Figure 2, the results are as follows.

```
y_bar   : 4.93
y_bar_i : 5.09 4.32 5.4 4.91
DFB     : 3
DFW     : 36
SQB_i   : 0.026 0.372 0.221 0
SQB     : 6.19
SQW_i   : 22.069 36.996 49.58 42.669
SQW     : 151.314
MSB     : 2.063
MSW     : 4.203
F       : 0.491
```

For this data set, the individual group means do not vary much about the grand mean, such that the SQB is not large in comparison to the SQW and the F-value thus smaller than 1.

In **R**, the in-built function `aov()` can be used to obtain the same results. In the following code, the `aov(D$Collagen ~ D$Treatment, data = D)` syntax means that the data in the dataframe field `Collagen` should be grouped according to the information in the dataframe field `Treatment`. The generic `summary()` function than prints results of the computation to the command window. Here `D$Treament` corresponds to the *between* quantities and `Residuals` corresponds to the *within* quantities. The meaning of `Pr(>F)` and the associated Significance codes will become clearer after in Section 3.

```
D       = read.csv("./_data/anova-data-set-1.csv")                    # data set 1
res.aov = aov(D$Collagen ~ D$Treatment, data = D)                     # ANOVA as a computational procedure
summary(res.aov)                                                      # results
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
D$Treatment  3  301.0  100.34   26.07 3.86e-09 ***
Residuals   36  138.6    3.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
D       = read.csv("./_data/anova-data-set-2.csv")                    # data set 2
res.aov = aov(D$Collagen ~ D$Treatment, data = D)                     # ANOVA as a computational procedure
summary(res.aov)                                                      # results
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
D$Treatment  3   6.19   2.063   0.491  0.691
Residuals   36 151.31   4.203
```

Partitioning of the Total-Sum-Squares. The SQB and SQW are mutually dependent quantities, in the sense that their sum corresponds to the *total-sum-squares* of a given data set. The total-sum-of-squares corresponds to the squared deviations of all data points regardless of group membership from the grand mean,

$$\text{SQT} := \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2. \tag{7}$$

As shown in Section 6.2, it holds that

$$\text{SQT} = \text{SQB} + \text{SQW}. \tag{8}$$

If one understands the SQB as a measure of the treatment-induced variability and the SQW as a measure of the data set's inherent unsystematic and random variability, Equation 8 can intuitively be read as

$$\text{Total data variability} = \text{Treatment-induced data variability} + \text{Unsystematic data variability}. \tag{9}$$

Equation 9 then forms the basis for understanding ANOVA from a linear model perspective as discussed next.

3 ANOVA as a linear model

3.1 Model formulation

A deeper understanding of the ANOVA approach is afforded by viewing it as an instance of a *linear modelling approach* to data analysis. In contrast to the computational viewpoint described above, modelling approaches do not start with a data set at hand and merely constitute a set of formulas to summarize such data set, but first and foremost provide hypotheses about the generative processes that give rise to observed data. As such, modelling approaches subserve the theoretical and deductive side of science, as well as its empirical and inductive aspects. Linear models can be shown to underlie a wide variety of methods from classical statistics, such as z- and t-tests, one-factorial and multifactorial ANOVA, simple linear regression, multiple linear regression and the analysis of covariance (ANCOVA). They also form the backbone for more generalized approaches to data analysis, such as multivariate, latent variable or nonlinear data modelling. A modern comprehensive introduction to linear models cannot eshew a detour through probability theory and some basics of linear algebra. However, for the purposes of this tutorial, attention is restricted to the structural features of the linear model in the context of ANOVA. The probabilistic components are addressed primarily through intuitive explanations, and references to matrix-based computations are deliberately omitted.

In essence, the linear modelling approach to ANOVA conceives each observed data point as the sum of a *deterministic component* induced by the treatment at each level and an unsystematic *random component* induced by additional processes such as the particularities of the observational unit under study or inadvertent variations in the experimental application of the respective treatment level:

$$\text{Data} = \text{Deterministic treatment-induced component} + \text{Random unsystematic component} \tag{10}$$

Note the similarity of this idea to the decomposition of the total data variability in Equation 9. More formally, the basic model for jth data point in the ith treatment group of an ANOVA design takes the form

$$y_{ij} = \mu_i + \varepsilon_{ij}. \tag{11}$$

Here, μ_i is a fixed, deterministic unknown number that depends on the treatment level $i = 1, ..., p$. ε_{ij} is a random variable that is specific to the jth data point in the ith treatment group, hence the double index. ε_{ij} is variably referred to as an *error variable*, a *measurement error*, or a *random fluctuation*. As a random variable, ε_{ij} takes on values at random with prespecified probabilities. A standard assumption for the error terms in linear models is that they follow a normal distribution (cf. Section 6.3), that is, a Gaussian bell-shaped curve centered at zero with variance parameter σ^2. This means that if one generates multiple realizations of a random variable like ε_{ij}, these realizations have a higher probability to be located close to zero than being located far away from zero. Because under this assumption most of the time ε_{ij} in Equation 11 takes on values close to zero, y_{ij} most of the time takes on values close to μ_i, rendering μ_i its so-called *expected value*. We will go a bit deeper into the significance of the ε_{ij} component for the applied ANOVA procedure in Section 3.3. For the moment we focus on the deterministic component μ_i that we have identified as the expected value of the data point y_{ij}.

While in principle using a single number μ_i for each treatment-level to reflect the deterministic data component of a data point y_{ij} is perfectly fine, most statistical software solutions use a different representation that has better generalisation properties to more complex data acquisition scenarios and fits the matrix computations underlying the evaluation of linear models in software better. This parameterization is commonly referred to as *reference group coding*. The idea of this approach is to encode the expected value of the data points in a specific group, referred to as the *reference group*, by a value μ_0 and to only encode the *differences* in expected values between the treatment groups and the reference group as additional parameters. Which group is chosen as the reference group is not of primary importance, when using statistical software, its usually the first group in a data table. However, if there exists a baseline group in which e.g., no treatment or a sham-treatment is applied, it is sensible to use this as the reference group for the remaining parameters to be readily interpretable. In the reference group coding approach to ANOVA, the linear model for the data points y_{ij} takes the form

$$y_{1j} = \mu_0 + \varepsilon_{1j} \text{ and } y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij} \text{ for } i = 2, ..., p. \tag{12}$$

The values $\alpha_2, ..., \alpha_p$ encode the differences in expected values between the ith group and the reference group. A fundamental aim of ANOVA as an inference approach is to infer whether the true values of $\alpha_2, ..., \alpha_p$ are different from zero or not.

3.2 Parameter estimation

Equation 12 constitutes a model for a data set acquired in an ANOVA design. As most models of quantitative phenomena, Equation 12 has parameters that are *estimated* based on the observed data y_{ij}. To understand this, it is sensible to reflect again on the data generative process that is encoded in Equation 12. The idea is that the data values $y_{1j}, j = 1, ..., n_1$ in the reference group are generated based on a fixed unknown value for the measured property in the reference group, which is termed μ_0. However, this value of μ_0 is not directly observed, but only after addition of the observational unit-specific random term ε_{1j}. Likewise, the values of μ_0 and α_i for the remaining groups are only observed after addition of their respective random contributions ε_{ij}. In this setting, the values of μ_0 and α_i are

referred to as *true, but unknown, values*. In fact, even after an ANOVA is computed, the actual values of μ_0 and α_i will remain unknown. However, during the execution of an ANOVA procedure *reasonable guesses* or *estimates* for μ_0 and the α_i's are evaluated.

The ANOVA parameter estimates can be motivated by noting that the random contributions are symmetrically distributed around zero under the assumption of their normality. If we consider for example the reference group with $n_1 = 5$ and assume that $\mu_0 = 5$ and $\varepsilon_{11} = 0, \varepsilon_{12} = 1, \varepsilon_{13} = -1, \varepsilon_{14} = 2\, \varepsilon_{15} = -1$, we obtain the data points

$$\begin{aligned} y_{11} &= \mu_0 + \varepsilon_{11} = 5 + 0 = 5 \\ y_{12} &= \mu_0 + \varepsilon_{12} = 5 + 1 = 6 \\ y_{13} &= \mu_0 + \varepsilon_{13} = 5 - 1 = 4 \\ y_{14} &= \mu_0 + \varepsilon_{14} = 5 + 2 = 7 \\ y_{15} &= \mu_0 + \varepsilon_{15} = 5 - 1 = 4 \end{aligned}$$

As positive and negative contributions to μ_0 cancel each other out when computing an average across these y_{1i}'s, a reasonable procedure to estimate μ_0 is to use the mean of the y_{1i} as an estimate for μ_0. Estimates for true, but unknown, parameters are usually denoted with a little hat. Thus, the usual estimate of μ_0 in this scenario is

$$\hat{\mu}_0 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j},$$

which for the data shown above evaluates to $\hat{\mu}_0 = 5.2$. This value is obviously not identical with the true, but unknown value of μ_0, but also not too far off. The statistical theory of linear models indeed assures that in the one-way ANOVA scenario with reference group coding, the formation of group averages is a reasonable approach to estimate the parameters $\mu_0, \alpha_2, \ldots, \alpha_p$. Specifically, it can be shown that a reasonable estimator for μ_0 is indeed the mean of the reference group and the differences between the respective group means and the mean of the reference group are reasonable estimators for $\alpha_2, \ldots, \alpha_p$, formally,

$$\hat{\mu}_0 = \bar{y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \text{ and } \hat{\alpha}_i = \bar{y}_i - \bar{y}_1 = \frac{1}{m} \sum_{j=1}^{m} y_{ij} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \text{ for } i = 1, ...p. \tag{13}$$

Note that the estimation of α_i as the difference between the ith group mean and the reference group mean is well in line with the interpretation of α_i as encoding the difference in expected value between the ith group values and the reference group values.

In **R** parameter estimates such as $\hat{\mu}_0$ and $\hat{\alpha}_2, \ldots, \hat{\alpha}_p$ are routinely evaluated when using the linear modelling approach to ANOVA. The generic function to apply all sorts of linear models to data in **R** is `lm()`. To execute an ANOVA as described above, the syntax `lm(D$Collagen ~ D$Treatment, data = D)` here again means that the data in the dataframe field `Collagen` should be grouped according to the information in the dataframe field `Treatment`. When calling `lm()` in this manner, the respective linear model is formulated and, based on the data at hand, the parameter estimates are evaluated. These can be inspected in the `coefficients` field of the resulting **R** object. We demonstrate this for the data set in Table 1. Note that the `(Intercept)` term $\hat{\mu}_0$ corresponds to the mean of the first group (`T1`), which is used as reference group, and the `D$TreatmentT2`, `D$TreatmentT3`, `D$TreatmentT4` terms for $\hat{\alpha}_2$, $\hat{\alpha}_3$, and $\hat{\alpha}_4$ correspond to the differences of the respective group means and the reference group mean.

```
# ANOVA as a linear model analysis
D         = read.csv("./_data/anova-data-set-1.csv")                    # data set 1
anova     = lm(D$Collagen ~ D$Treatment, data = D)                      # ANOVA as a computational procedure
print(anova$coefficients)                                               # parameter estimates
```

```
  (Intercept) D$TreatmentT2 D$TreatmentT3 D$TreatmentT4
         4.50          6.02          0.81         -1.14
```

```
# comparison with group means
Y         = as.matrix(data.frame(split(D$Collagen, D$Treatment)))              # wide format as simple matrix
y_bar_i   = colMeans(Y)                                                        # group means
mean_ref  = y_bar_i[1]                                                         # reference group mean
mean_ref                                                                       # print
```

```
 T1
4.5
```

```
mean_diff = y_bar_i[2:4] - y_bar_i[1]                                          # group mean differences
mean_diff                                                                       # print
```

```
   T2    T3    T4
 6.02  0.81 -1.14
```

3.3 Inference

To understand the logic of *p-values* and related concepts such as *Frequentist estimator properties*, *confidence intervals* and *error control in hypothesis testing*, it is crucial to remind oneself of the *Frequentist probability interpretation* used standard applications of linear models. We have already noted that in the ANOVA model Equation 12 the parameters μ_0 and $\alpha_2, ..., \alpha_p$ are conceived as fixed unknown values that govern the data generating process of the ANOVA scenario. For $i = 1, ..., p$ and $j = 1, ..., m$ the ε_{ij}'s, on the other hand, are conceived as random variables. Random variables take on values with prespecified probabilities. For example, we can talk about the probability that the random variable ε_{23} takes on a value in an interval such as $[-0.10, 0.10]$. The probability for the error variable ε_{23} to take on specific values also determines the probability of y_{23} to take on specific values, given fixed values of μ_0 and α_2. For example, if $\mu_0 = 1$ and $\alpha_2 = 2$ and ε_{23} has a high probability to take on values in the interval $[-0.10, 0.10]$, then y_{23} has a high probability to take on values in the interval $[1 + 2 - 0.10, 1 + 2 + 0.10] = [2.90, 3.10]$. The important thing here is to realize that given the fixed nature of the parameters μ_0 and $\alpha_i, i = 2, ..., p$ and the random nature of the ε_{ij}'s also the data variables are conceived as random variables.

Now the fundamental assumption of the Frequentist approach to statistics is to assume that a given data set, such as the data presented in Table 1, reflects a single realization of each of the ε_{ij}'s and in turn the y_{ij}'s for true, but unknown, fixed parameters μ_0 and $\alpha_i, i = 2, ..., p$. From this viewpoint, the collection of the y_{ij}'s in Table 1 thus represents a single joint realization of the random variables constituting the respective ANOVA design. Against the background of the assumed random nature of the ε_{ij}'s involved, however, it is easily conceivable to imagine a second joint realization of the random variables constituting the ANOVA design. In basic introductions to statistics this is sometimes presented as "repeating the same study under identical circumstances and collecting data that, by an large, up to random fluctation, resembles that of the initial study". Yet, it is crucial to realize that the Frequentist assumption underlying classical inference in linear models is just this, an assumption. In essence it constitutes a cognitive model that is superimposed onto observed data and is no more "real" or "unreal" than any other form of human imagination. Nevertheless superimposing the Frequentist assumption on a given experimental and data-analytical scenario allows for asking how observed data, or numbers computed from these data, such as means, sum-of-squares, mean sum-of-squares or F-values are distributed, assuming known-properties of the random variables involved in the data generating process, given here by the ε_{ij}'s.

A straighforward way to gain intuition about the Frequentist distributional properties of data-analytical quantities is to simulate the repeated execution of a given study with identical true, but unknown, parameters on a computer. As this approach makes abundant use of random number generators, it is sometimes referred to as a *Monte Carlo approach* in reference to the famous casino in Monte Carlo. We prefer to refer to it as a *sampling approach*, as random variables are repeatedly sampled. For example, the following **R** code generates 10 samples from a single normal random variable with expectation parameter 2 and variance parameter 1.

```
n = 1                                                                          # number of random variables
s = 10                                                                         # number of samples taken
m = 2                                                                          # expectation parameter
v = 1                                                                          # variance parameter
```

```
for(i in 1:s){                                  # repeated sampling
  print(rnorm(n,m,v))}                          # sampling a normally distributed random variable
```

```
[1] 2.79408
[1] 1.142718
[1] 2.161665
[1] 3.216463
[1] 3.066284
[1] 1.657042
[1] 0.5682848
[1] 3.682252
[1] 2.886641
[1] 1.15286
```

To understand the notion of an F-*distribution* and the associated p-*value* of an observed F-*value*, it is crucial to realize that in essence, an F-value is a fairly condensed representation of a data set as presented in Table 1. Everytime a new data set is realized (in simulation or reality), new grand means, new group means, new sum-of-squares, new mean-sum-of-squares and a new F-value are realized. The fundamental logic of p-*values* is then to ask, how probable a random F-value at least as extreme as an actual observed F-value under the assumption that, in the ANOVA design, all true, but unknown, group differences with respect to the reference group $\alpha_2, ..., \alpha_p$ are zero. This latter assumption

$$\alpha_2 = \cdots = \alpha_p = 0$$

is referred to as the *(global) null hypothesis* in an ANOVA design, as it represents the assumption of no treatment-induced effects. Note that the null hypothesis is formulated with respect to the true, but unknown, values not their empirical counterparts, such as the $\hat{\alpha}_2$,..., $\hat{\alpha}_p$. The logic of Frequentist inference is then to conclude that if the probability p of observing a random F-value at least as extreme, i.e., equal to or larger than, the actual observed F-value is small, the null hypothesis should be rejected. This is the essence of the highly controversial, but very popular *null-hypothesis significance testing* approach of classical Frequentist statistics. In the following, we trace how the distribution of the F-value is constructed based on the normal assumption about the error variables in the ANOVA linear model to illustrate the impact of the probabilistic assumptions in Equation 12 for the practice of reporting p-values.

Construction of the F-distribution

The Frequentist distribution of the F-value under the null hypothesis of the ANOVA model derives from sequential transformations of the error variables ε_{ij} for all $i = 1, ..., p$ and $j = 1, ..., n_i$. As discussed above, the error variables are assumed to be *identically* distributed according to a normal distribution with expectation parameter 0 and variance parameter σ^2. Furthermore, the error variables are assumed to be *independently* distributed. In the context of normal distributions, this also implies that they are uncorrelated. Formally, this means that the distribution of a single ε_{ij} can be described by the probability density function of the univariate normal distribution,

$$p\left(\varepsilon_{ij}\right) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\varepsilon_{ij} - \mu\right)^2\right)$$

for $\mu := 0$. Intuitively, this means, that if a certain observational unit realizes by chance for example a high value for its error variable, the probability for none of the other observational units to also realize a very high (or, for negative covariations, a very low) error variable value is affected in any way. The essence of the Frequentist distribution framework for the ANOVA model leading up to the F-value distribution is then to mathematically derive the formulas for the distributions of the transformed independent and identically normally distributed error variables through the various stages of the ANOVA model and the evaluation of the various statistics cumulating in the F-distribution. Tracing this process mathematically is beyond the scope of the current tutorial. We nevertheless list the core steps below and provide a sampling-based validation in Figure 3.

(1) The assumption of identically and independently normally distributed error variables, exemplarily shown for ε_{11} in Figure 3A, results in the fact that under the ANOVA model, the data variables y_{ij} are independently normally distributed with the same variance as the error variables, examplarily shown in Figure 3B, as adding constant numbers to normally distributed random variables results in another normally distributed random variable.

(2) The normal distribution of the data variables y_{ij} in turn implies the normal distribution of the group means, exemplarily shown for $\bar{y}_1$ in Figure 3C and the grand mean, as shown in Figure 3D, as summing together normally distributed random variables and multiplying them by a constant renders the result normally distributed.

(3) Subtracting normally distributed random variables from each other yields yet other normally distributed random variables, such that the deviations in the SQB and the SQW are themselves normally distributed, as shown exemplarily for $y_{11} - \bar{y}_1$ and $\bar{y}_1 - \bar{y}$ in Figure 3E and Figure 3F, respectively.

(4) Squaring normally distributed random variables, in the current case the deviations between individual data variables and their group mean or the group means and the grand mean, does not result in normally distributed random variables. This is readily appreciated by considering a random variable centered around zero and considering the square of its realizations - all negative realizations will become positive, values between 0 and 1 will move closer to zero and values above 1 will move further away from 1. Squaring thus breaks the symmetry of normally distributed values. Broadly speaking, the square of an independent normally distributed random variable, as well as the sum of such squares, follows a χ^2 distribution. Unlike the normal distribution, the corresponding probability density function has a different — yet analytically available — form. In the context of the F-distribution, the variance-scaled versions of SQB and SQW are independently χ^2-distributed with parameters $p-1$ and $n-p$, respectively, as shown in Figure 3G and Figure 3H.

(5) Finally, forming the ratio of two independent χ^2-distributed random variables divided by their respective parameters leads to a distribution with probability density function referred to as the F-distribution, as shown in Figure 3I.

The p-value

Because the F-value takes on larger values, if the true, but unknown, parameters $\alpha_2, ..., \alpha_p$ take on values different from zero, asking how probable a given observed value of the F-value or a more extreme value is, gives a probability that becomes small for evidence against the null hypothesis. This probability is known as the *p-value* and can be visualized as the area under the curve of the F-distribution to the right of an observed F-value (Figure 4). If this probability is low, for example lower than 0.05, the evidence in a data set is commonly interpreted as considerable evidence against the null hypothesis. Note however, that the p-value itself has no implications for the veracity or non-veracity of the null hypothesis: F-values at least as large as an observed F-value are just not very likely in the case that the null hypothesis is true, but they are by no means impossible. In fact, their associated probability is just the p-value. Also note that the existence of a p-value is crucially dependent on the normal distribution assumption of the error terms, as its value derives from the F-distribution, which is constructed based on exactly this distribution. In the **R** ANOVA table, the p-value is listed as `Pr(>F)`. With the help of the cumulative distribution function of the F-distribution, which encodes the probability $\mathbb{P}(F \leq f)$ for all observable F-values f, the p-value can be readily evaluated based on the following **R** code. Figure 4 visualizes this process.

```
p       = 4                                   # number of groups
m       = 10                                  # number of observational units per group
n       = p*m                                 # total number of data points
F_value = 0.5                                 # observed $F$-value
p_value = 1-pf(F_value, p-1, n-p)             # associated p-value
cat("p-value: ", round(p_value,3))            # print
```

```
p-value:  0.685
```

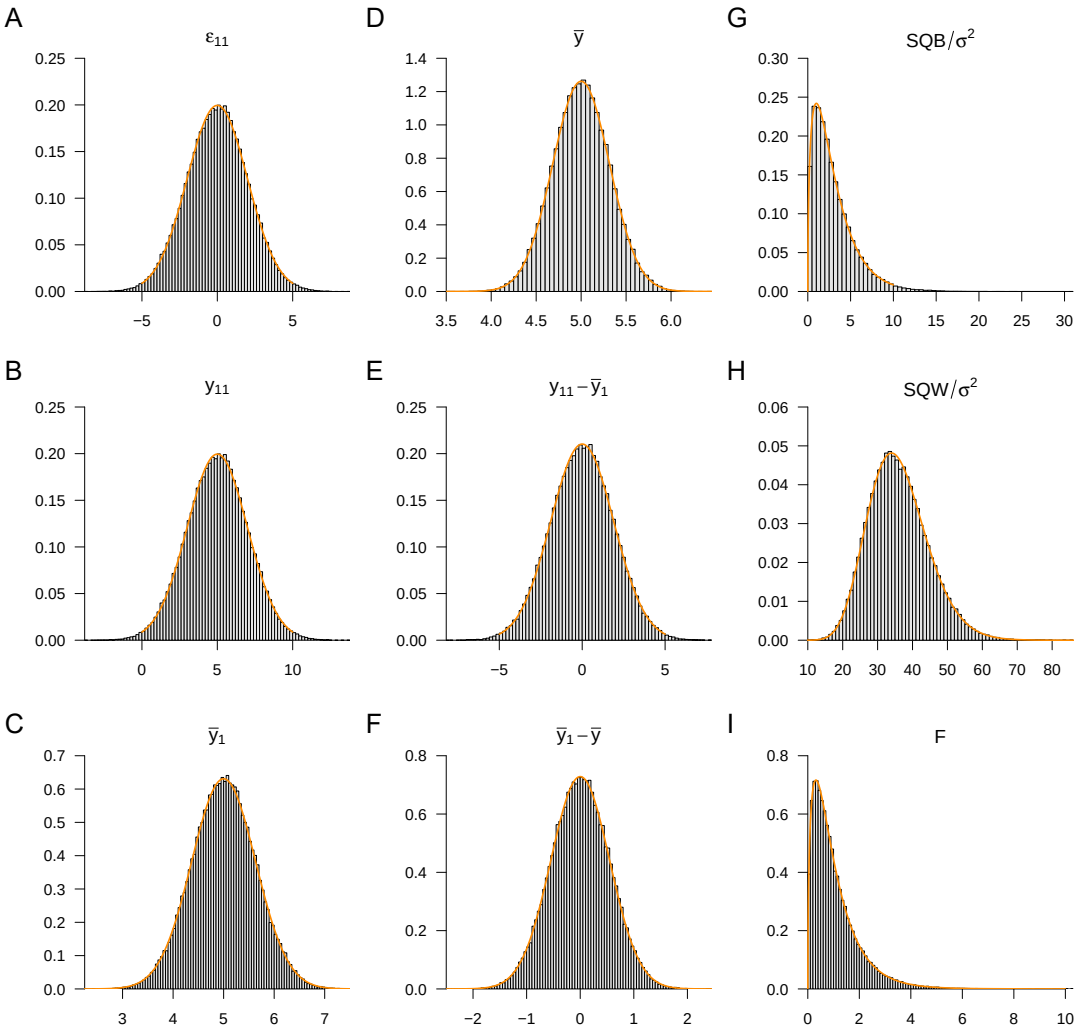

Figure 3: Construction of the F-distribution.

Finally, we note that the distribution of the F-value cannot only be evaluated under the null hypothesis $\alpha_2 = \cdots = \alpha_p = 0$, but equally well under any other hypothesis about the true, but unknown, parameters. This leads to the concept of the non-central F-distribution, an important tool in the context of ANOVA power analyses and optimal sample size calculations in the Frequentist setting.

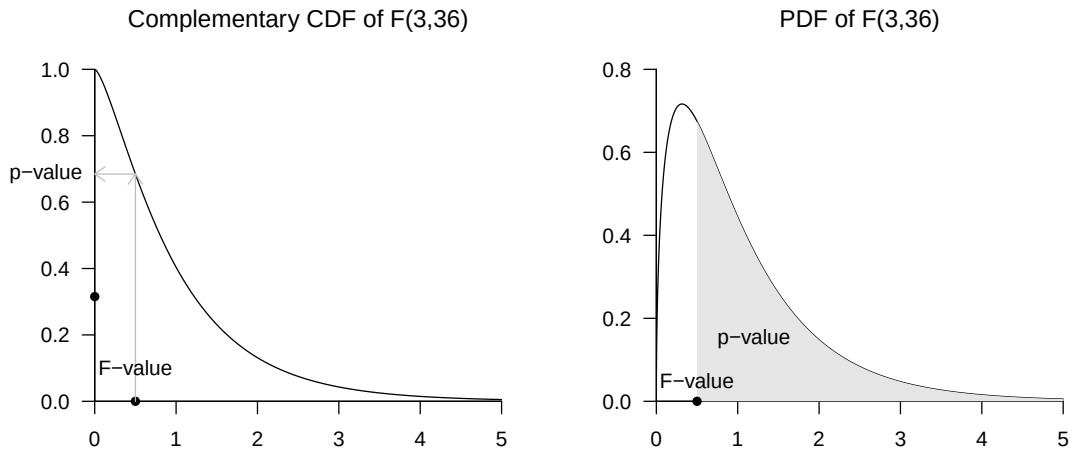

Figure 4: Evaluation of the ANOVA p-value for the example data set 2 with observed F-value 0.5.

4 ANOVA as a linear mixed model

Thus far, we have assumed that the observational units aside from being assigned to different treatment groups do not exhibit any predetermined inherent structure. In fact, in Table 1 and Table 2, we have separated the inherent structure of observational units, i.e., fibroblasts from one cell culture strain vs. fibroblasts from another cell culture strain, into two independent experiments and found evidence for treatment effects in one strain, but not the other. If there exist inherent structure in the observational units that is in one way or another superimposed onto the structure of the experimental manipulation, it often pays to account for this structure in the modelling of the experimental data set. Today, such structure modelling is often achieved using *linear mixed models*, which in addition to (fixed) experimental treatment effects and (random) unsystematic effects also account for (random) observational unit effects that reflect some form of structure. In fact, not accounting for structured variation of experimental data that is induced by the observational units, but treating all observational units as if they were independent, can have effects on the results of ANOVA analyses. The reason for this is that the analytical form of the F-distribution is built on the assumption of independent and identically distributed error variables for all observational units. If this assumption is violated because certain error correlations exist, the actual distribution of F-values as computed by the procedures discussed in Section 2 may differ from the analytically available form as discussed in Section 3. While in the past, these instances were usually dealt with by adapting the analytical form of the F-distribution appropriately, e.g., by evaluating modified degrees of freedom parameters for it as in the *Huyhn-Feldt* or *Greenhouse-Geisser corrections*, these methods have by and large been superseeded by linear mixed model approaches. To motivate and explore these approaches, we consider two scenarios of inherent observational unit structure that, if analysed using the conventional ANOVA approach assuming independent observational unit error variables, lead to an increase of false negative results (loss of sensitivity) or an increase of false positive results. These scenarios are referred to as *randomized block designs* and *nested subsampling designs*.

Randomized block designs and the closely related *repeated-measures designs* are characterized by the fact that all experimental conditions are measured within a given experimental unit. In the cell biological example, if three different cell culture strains were used and each of the four experimental treatments was applied within each strain, each to a set of fibroblasts from the particular strain, this setup would be referred to as a randomized block design. Another common example are cognitive psychology experiments, in which behavioral responses of human participants are typically measured in response to all experimental treatments of interest. Figure 5A depicts the typical design of an

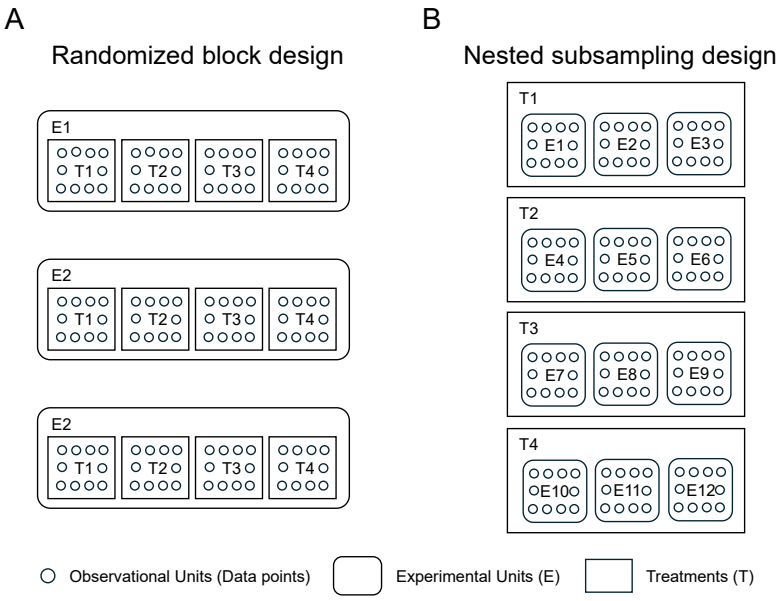

Figure 5: Exemplary randomized block and nested subsampling designs. Both example designs comprise four experimental treatments. In the randomized block design, the observational units within an experimental unit (block) are subjected to all treatments. Differences between experimental units thus cancel out, if differences between the treatments are assessed. However, the spread of measurements in one condition may be larger than in a one-way ANOVA design with one experimental unit due to the varying experimental unit effects. In the nested subsampling design, observational units of different experimental units are used across the different treatments. If, by chance, the observational units of experimental units E10 to E12 have high values due to the nature of the experimental units E10 to E12, this may be mistaken as an effect of T4 if not accounted for.

ANOVA randomized block design. Randomized block designs have the benefit, that, at least in theory, data variation due to the experimental unit is equally expressed across all experimental treatments, such that observed differences between experimental treatments are not affected by the experimental units. They have the disadvantage that the effect of the experimental units may increase the data spread about each treatment group mean.

A basic linear mixed model for a randomized block design takes the form

$$y_{kij} = \mu_0 + \alpha_i + \mu_k + \varepsilon_{kij}, \tag{14}$$

where for $k = 1, ..., n_k$, $i = 1, ..., p$ and $j = 1, ..., n_{ki}$

- y_{kij} is the data of the jth observational unit in the ith treatment group in the kth experimental unit
- μ_0 is the average reference group effect across experimental units,
- α_i is the average difference between treatment level i and the average reference group effect with $\alpha_1 := 0$
- $\mu_k \sim N(0, \sigma_k^2)$ is the random effect of the kth experimental unit,
- $\varepsilon_{kij} \sim N(0, \sigma_\varepsilon^2)$ is the error of the jth observational unit in the ith treatment in the kth experimental unit.

Note that in contrast to Equation 12, Equation 14 includes explicit terms that reference the experimental unit, in particular the normally distributed random experimental unit effect $\mu_k \sim N(0, \sigma_k^2)$. Parameter estimation for linear mixed models such as Equation 14 is complicated by the fact that in contrast to Equation 12, linear mixed models include multiple latent random variables such as μ_k and ε_{kij} and multiple parameters such as σ_k^2 and σ_ε^2.

Figure 6 shows simulated data from a randomized block design with $k = 3$ experimental units `E1`, `E2`, `E3`, $p = 4$ treatments `T1`,`T2`,`T3` and `T4` and $n_{ki} = 10$ observational units per treatment group in each experimental unit. The data

is generated based on the null hypothesis parameter setting $\alpha_2 = \alpha_3 = 0$. Note that when averaging for example the data points of `T1` across all experimental units, and assessing the remaining differences between the data points of `T1` across all experimental units, this component of the within-sum-of-squares will be larger than compared to a scenario in which only data from one of the experimental units is assessed. On the other hand, the observable differences across all treatment groups across experimental units cancel out, if differences between treatment means are evaluated in a per experimenal unit manner.

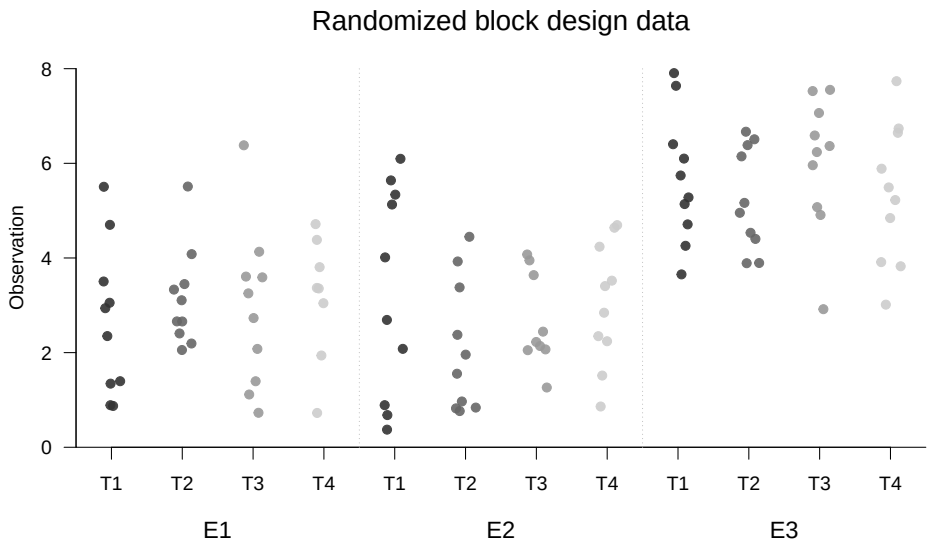

Figure 6: Randomized block design data example.

Formally, it can be shown that the presence of random terms such as μ_k induces correlations in the marginal error and data covariance matrices. In the case of randomized block designs, these correlations induce an increase of the within-sum-of-squares such that, when not accounted for, on average lower F-values are observed than predicted based on independent observational unit error terms. Figure 7B shows this shift of observed F-values under the null hypothesis when computed based on samples from Equation 14 as opposed to computed based on samples from Equation 12 as shown in Figure 7A.

The popularity of linear mixed models in the analysis of modern ANOVA designs owes much to the availability of software packages that make data analysis with mathematically fairly complex models very easy in research applications. The most popular **R** packages for linear mixed model analysis are `nlme` (Pinheiro and Bates (2000)) and `lme4` (Bates and DebRoy (2004)). We here demonstrate a linear mixed model analysis using the `lme()` function of `nlme` which capitalizes on a *restricted maximum-likelihood* approach for covariance component estimation. In addition to specifying which data should be modeled with respect to which experimental factor, linear mixed model implementation typically required additional information about the observational unit structure (grouping). For the data shown in Figure 6, this information is encoded in the `EU` (experimental unit) column of the data set table. The following code snippet shows the analysis of this data set using the linear mixed model of Equation 14.

```
library(nlme)                                                    # (non)linear mixed model package
D       = read.csv("./_data/anova-randomized-block-design-data-set.csv")   # dataframe in long
anova(lme(y ~ TRM, data = D, random = ~ 1|EU))                   # LMM treatment effect evaluation
```

```
            numDF denDF   F-value p-value
(Intercept)     1   114 14.995844  0.0002
TRM             3   114  0.332327  0.8020
```

Using this approach in the Frequentist simulations of Figure 7C shows that an analytical F-distribution is available that conforms to the observed F-value distribution of the linear mixed model. Using a linear mixed model that properly accounts for the induced error correlations thus renders an ANOVA analysis of data from randomized block design more sensitive.

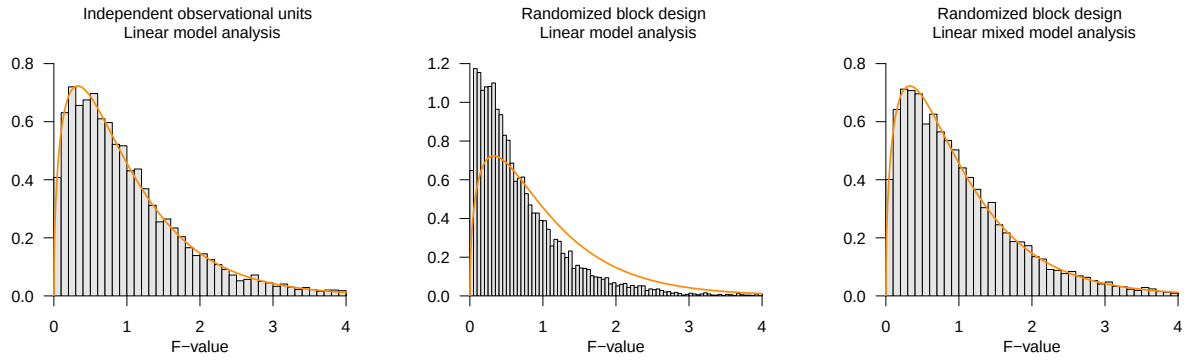

Figure 7: Loss of sensitivity in a randomized block design when analyzed with a linear model ANOVA, but not a linear mixed model ANOVA.

Nested subsampling designs are experimental designs sometimes encountered in biological studies (cf. Yu et al. (2022)). Here, the basic idea is that each experimental unit provides a set of observational units and treatments are applied to sets of experimental units. For example, suppose four different cytokine treatments are tested. For each treatment, fibroblasts are harvested from three different animals, and collagen production is then measured in these fibroblasts. Because the fibroblasts are not independent samples but instead are nested within animals (which are themselves grouped by treatment), this setup is considered a nested subsampling design. Intuitively, nested subsampling designs have the disadvantage that they can easily conflate experimental unit and treatment effects. In Figure 5B, assume that by chance the fibroblasts of experimental units E10 to E12 produce high levels of Collagen, irrespective of any treatment. If observed under the nested subsampling design, this random experimental unit effect can readily be mistaken as an effect of the fourth treatment level. Nested subsampling designs are thus prone to produce false positive effects, if their inherent observational unit structure is not accounted for.

A basic linear mixed model for a nested subsampling design takes the following form.

$$y_{ikj} = \mu_0 + \alpha_i + \mu_{ik} + \varepsilon_{ikj}, \tag{15}$$

where for $k = 1, ..., n_k$, $i = 1, ..., p$ and $j = 1, ..., n_{ik}$

- y_{ikj} is the data of the jth observational unit of the kth experimental unit in the ith treatment group,
- μ_0 is the average reference group effect across experimental units,
- α_i is the average difference between treatment level i and the average reference group effect with $\alpha_1 := 0$
- $\mu_{ik} \sim N(0, \sigma_k^2)$ is the random effect of the kth experimental unit in the ith treatment group,
- $\varepsilon_{ikj} \sim N(0, \sigma_\varepsilon^2)$ is the error of the jth observational unit in the kth experimental unit in the ith treatment.

Note that in contrast to Equation 14, the random effects μ_{ik} are experimental unit *and* treatment-specific.

Figure 8 shows simulated data from a nested subsamplig design with $n_k = 12$ experimental units `E1` to `E12`, $p = 4$ experimental treatments and $n_{ik} = 10$ observational units per experimental unit. Note that the elevated observations in `T4` may well be induced by the experimental unit effects of `E10` ,`E11`, and `E12`. In general it is easily possible, that experimental unit effects inflate the mean between-sum-of-squares in a linear model ANOVA of nested subsampling

design data. In analogy to the the analysis of the randomized block design data, the following code snippet shows the analysis of this data set using the linear mixed model of Equation 15.

```
library(nlme)                                                    # (non)linear mixed model package
D       = read.csv("./_data/anova-nested-subsampling-design-data-set.csv")   # dataframe in long format
anova(lme(y ~ TRM, data = D, random = ~ 1|EU))                   # LMM treatment effect evaluation
```

```
            numDF denDF  F-value p-value
(Intercept)     1   108 777.8877  <.0001
TRM             3     8  22.6174   3e-04
```

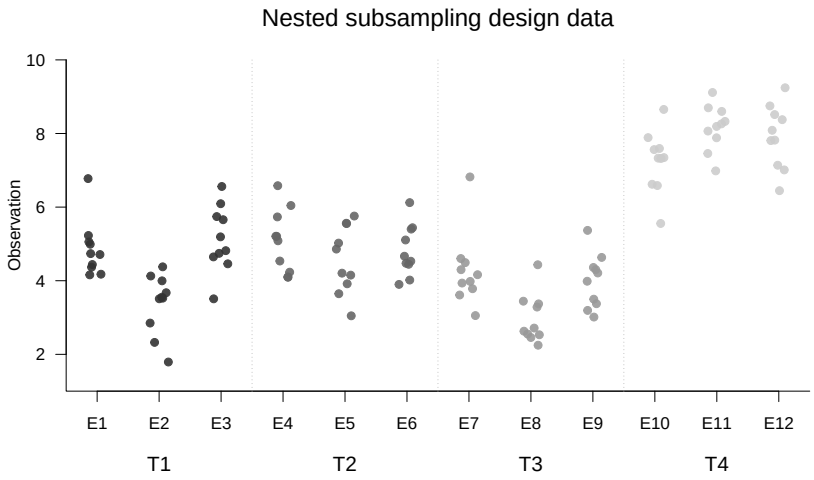

Figure 8: Nested subsampling design data example.

As for the randomized block design, it can be shown that the presence of random terms such as the μ_{ik}s induces correlations in the marginal error and data covariance matrices. In the case of nested subsampling designs, these correlations induce an increase of the between sum of squares such that, when not accounted for, on average higher F-values are observed than predicted based on independent observational unit error terms. Figure 9B shows this shift of observed F-values under the null hypothesis when computed based on samples from Equation 15 as opposed to computed based on samples from Equation 12 as shown in Figure 9A. Again, as shown in Figure 9C, using a linear mixed model that properly accounts for the induced error correlations renders the analysis less prone to false positive results.

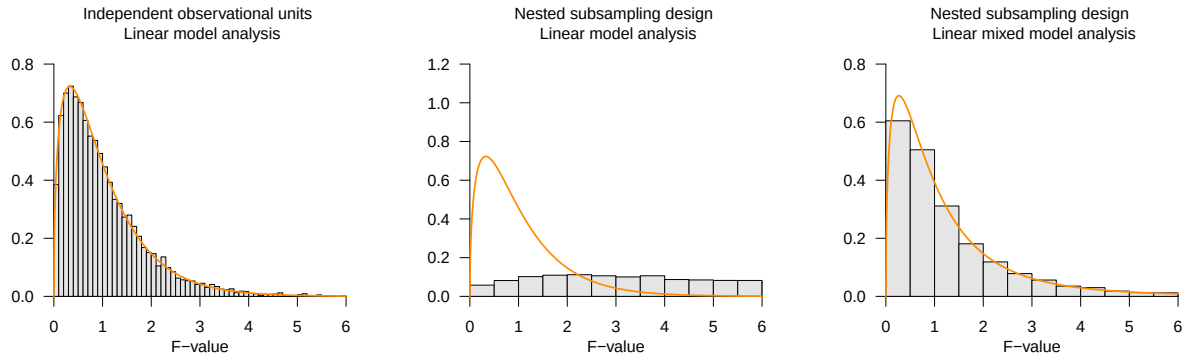

Figure 9: Increase of false positives in a nested subsampling design when analyzed using a linear model ANOVA, but not a linear mixed model ANOVA.

5 Bibliographic remarks

Fisher (1925) and Scheffé (1959) provide classical treatments of ANOVA as a computational procedure. Winer (1971) and Maxwell, Delaney, and Kelley (2018) provide excellent, largely matrix calculus-free introductions to ANOVA as a linear model. Hays (1994) is a good text for newcomers to statistics. Snijders and Bosker (2012) provide a general non-technical introduction to linear mixed models and Yu et al. (2022) provides an introduction to linear mixed models in the context of cell and animal research.

6 Supplementary Material

6.1 The sum symbol

The sum symbol is defined as

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n. \tag{16}$$

Here:

- Σ denotes the Greek letter Sigma, mnemonic for Sum,
- the subscript $i = 1$ specifies the running index and the starting index,
- the superscript n specifies the ending index, and
- $x_1, x_2, \dots, x_n$ are the summands.

For meaningful use of the summation symbol, it is essential to specify the lower and upper limits of summation via the subscript and superscript.The particular choice of symbol for the running index is irrelevant to the value of the sum; indeed,

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j. \tag{17}$$

Sometimes the running index is given as an element of an index set. For example, if the index set $I := 1, 5, 7$ is defined, then

$$\sum_{i \in I} x_i := x_1 + x_5 + x_7. \tag{18}$$

In the following, we present several examples illustrating the use of the summation symbol.

Summation of predefined summands. Let $x_1 := 2$, $x_2 := 10$, and $x_3 := -4$. Then

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 2 + 10 - 4 = 8. \tag{19}$$

Summation of weighted predefined summands. Let $x_1 := 2$, $x_2 := 10$, and $x_3 := -4$. Furthermore, let the weighting coefficients be $a_1 := \frac{1}{2}$, $a_2 := \frac{1}{5}$, and $a_3 := 2$. Then

$$\sum_{i=1}^{3} a_i x_i = a_1 x_1 + a_2 x_2 + a_3 x_3 = \frac{1}{2} \cdot 2 + \frac{1}{5} \cdot 10 + 2 \cdot (-4) = 1 + 2 - 8 = -5. \tag{20}$$

Expressions of the form $\sum_{i=1}^{n} a_i x_i$ are referred to as *linear combinations* of $x_1, \dots, x_n$ with coefficients or weights $a_1, \dots, a_n$.

Summation of natural numbers.

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15. \tag{21}$$

Summation of even natural numbers.

$$\sum_{i=1}^{5} 2i = 2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 = 2 + 4 + 6 + 8 + 10 = 30. \tag{22}$$

Summation of odd natural numbers.

$$\sum_{i=1}^{5}(2i-1) = 2\cdot 1-1+2\cdot 2-1+2\cdot 3-1+2\cdot 4-1+2\cdot 5-1 = 1+3+5+7+9 = 25. \tag{23}$$

Nested sums The computational procedure for evaluating nested sums, i.e. expressions involving multiple sequential sum signs is best understood by expanding it from the inside outward, noting that during each iteration of the inner sum, the index of the outer sum remains fixed, The following example illustrates this:

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3}(x_i+y_j) &= \sum_{i=1}^{2}(x_i+y_1+x_i+y_2+x_i+y_3)\\ &= (x_1+y_1+x_1+y_2+x_1+y_3)+(x_2+y_1+x_2+y_2+x_2+y_3).\end{aligned} \tag{24}$$

6.2 Total sum-of-squares partition

By definition of the SQT, SQB and SQW, we have

$$\begin{aligned}\text{SQT} &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2\\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y}\right)^2\\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\left(y_{ij}-\bar{y}_i\right)+\left(\bar{y}_i-\bar{y}\right)\right)^2\\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\left(y_{ij}-\bar{y}_i\right)^2+2\left(y_{ij}-\bar{y}_i\right)\left(\bar{y}_i-\bar{y}\right)+\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{j=1}^{n_i}2\left(y_{ij}-\bar{y}_i\right)\left(\bar{y}_i-\bar{y}\right)+\sum_{j=1}^{n_i}\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+2\left(\bar{y}_i-\bar{y}\right)\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+2\left(\bar{y}_i-\bar{y}\right)\sum_{j=1}^{n_i}\left(y_{ij}-\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}\right)+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+2\left(\bar{y}_i-\bar{y}\right)\left(\sum_{j=1}^{n_i}y_{ij}-\sum_{j=1}^{n_i}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}\right)\right)+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+2\left(\bar{y}_i-\bar{y}\right)\left(\sum_{j=1}^{n_i}y_{ij}-\frac{n_i}{n_i}\sum_{j=1}^{n_i}y_{ij}\right)+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+2\left(\bar{y}_i-\bar{y}\right)\left(\sum_{j=1}^{n_i}y_{ij}-\sum_{j=1}^{n_i}y_{ij}\right)+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+n_i\left(\bar{y}_i-\bar{y}\right)^2\right)\\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{p}n_i\left(\bar{y}_i-\bar{y}\right)^2\\ &= \text{SQW}+\text{SQB}\end{aligned} \tag{25}$$

6.3 The normal distribution

The normal (or Gaussian) distribution is one of the most commonly encountered mathematical models for random phenomena, such as measurement errors in scientific experiments. A key reason for its importance is the *Central Limit Theorem*, which states

that the sum of many independent random variables - regardless of their individual distributions - tends to converge to a normal distribution. Thus, the idea that, beyond the explanations offered by a deterministic scientific theory, numerous unknown factors influence any given observation provides strong motivation for modelling error variables with the normal distribution. Formally, the normal distribution is described by the probability density function.

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{26}$$

The expectation parameter μ determines the location of the function on the x axis, the variance parameter σ^2 the width of the curve. In general, a probability density function defined on the real numbers represents probabilities through densities. To determine the probability that a normally distributed random variable falls within a given interval, one integrates its probability density function over that interval in the Riemannian sense. Since interval length is measured uniformly along the real line, regions of high probability density - such as those near the expectation parameter μ - correspond to higher probabilities of the random variable taking values there. The situation is analogous to the concept of densities in physics: only multiplying a non-zero volume (interval) by the density (probability) of some material yields a non-zero mass (probybility mass). Repeated sampling with replaced of a random variable (or simultaneous sampling of many identical copies of a single random variable) produces data that, when analyzed statistically, reflect the characteristic distribution of the underlying mathematical model.

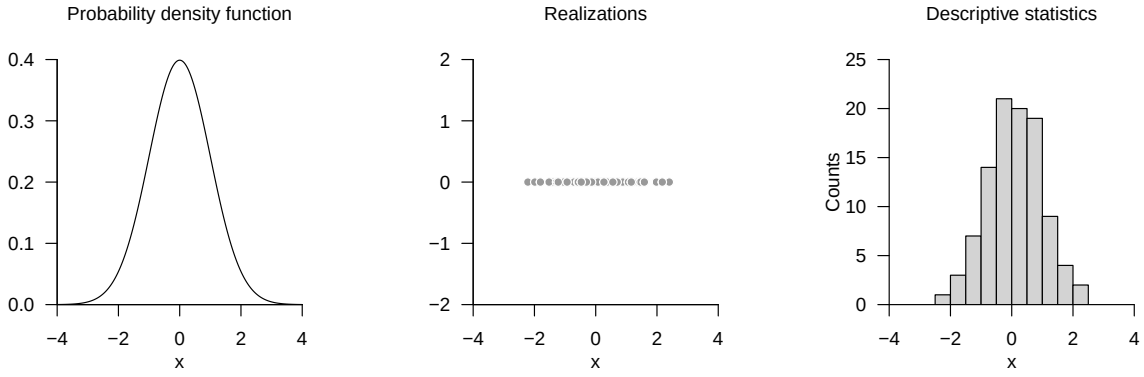

Figure 10: The normal distribution from a probability theory-, data-, and descriptive statistics- perspective.

References

Bates, Douglas M, and Saikat DebRoy. 2004. "Linear Mixed Models and Penalized Least Squares." *Journal of Multivariate Analysis* 91 (1): 1–17. https://doi.org/10.1016/j.jmva.2004.04.013.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Oliver & Boyd.

Hays, William. 1994. *Statistics*. 5th ed. Harcourt Brace College Publishers.

Maxwell, Scott E., Harold D. Delaney, and Ken Kelley. 2018. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Third edition. New York London: Routledge, Taylor & Francis Group.

Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. New York: Springer.

Scheffé, Henry. 1959. *The Analysis of Variance*. Wiley classics library ed. A Wiley Publication in Mathematical Statistics. New York: Wiley-Interscience Publication.

Snijders, T. A. B., and R. J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Los Angeles: Sage.

Winer, B J. 1971. *Statistical Principles in Experimental Design*.

Yu, Zhaoxia, Michele Guindani, Steven F. Grieco, Lujia Chen, Todd C. Holmes, and Xiangmin Xu. 2022. "Beyond t Test and ANOVA: Applications of Mixed-Effects Models for More Rigorous Statistical Analysis in Neuroscience Research." *Neuron* 110 (1): 21–35. https://doi.org/10.1016/j.neuron.2021.10.030.