R wie Rotz

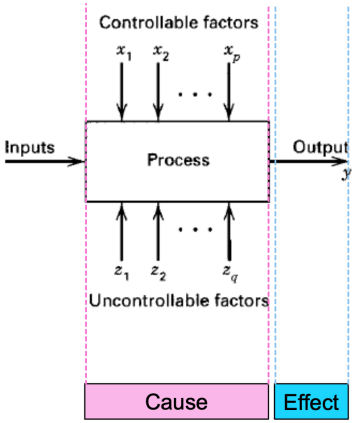
by dcamenisch

1 Introduction

This document is a summary of the 2022 edition of the lecture *Applied Analysis of Variance and Experimental Design* at ETH Zurich. I do not guarantee correctness or completeness, nor is this document endorsed by the lecturers. If you spot any mistakes or find other improvements, feel free to open a pull request on `https://github.com/DannyCamenisch/anova-summary`. This work is published as CC BY-NC-SA.

2 Learning from Data

From an abstract point of view, we are in the situation where we have a system or a process with many input variables (**predictors**) and an output (**response**). We want to find **cause-effect relationships**, meaning that when we actively change one of the inputs (intervention), this will cause the output to change. This is what we do in **experimental studies**. If we can just observe a system under different settings (observational studies), it is much harder to make a statement about causal effects. With observational data, we can typically just make a statement about an association between two variables. One potential danger is the existence of **confounders** (a common cause for two variables).



2.1 Experimental Studies

Before designing an experimental study, we must have a precise research question that is actually testable, i.e., that we can do the appropriate interventions and that we can measure the right response.

An experimental study consists of:

- **Treatments / Predictors**: the different interventions on the system
- **Experimental units**: the actual objects on which we apply the treatments
- A method that assigns experimental units to treatments, typically **randomization**
- **Response(s)**: the output that we measure

2.1.1 Treatments or Predictors

We distinguish between the following types of predictors:

- Predictors that are of primary interest and that can (ideally) be varied according to our wishes
- Predictors that are systematically recorded such that potential effects can later be eliminated in our calculations (**covariates**)
- Predictors that can be kept constant and whose effects are therefore eliminated
- Predictors that we can neither record nor keep constant

2.1.2 Randomization

Randomization ensures that the only systematic difference between the groups is the treatment. This protects us from confounders and is the reason why a properly randomized experiment allows us to make a statement about a cause-effect relationship between treatment and response. Typically, we then do a randomization within homogeneous blocks. This restricted version of randomization is called blocking. A block is a subset of experimental units that is more homogeneous than the entire set.

2.1.3 Experimental and Measurement Units

An **experimental unit** is defined as the object on which we apply the treatments by randomization. On the other hand, a **measurement unit** is the object on which the response is being measured.

2.1.4 Experimental Error

Different experimental units will give different responses to the same treatment (**experimental error**). Therefore we need multiple replicates receiving the same treatment. If the difference between the treatments is much larger than the experimental error, we can conclude that there is a treatment effect.

2.1.5 Blinding

Blinding means that those who measure the response do not know which treatment is given. With humans it is common to use **double-blinding** where in addition the patients do not know the assignment either. Blinding protects us from (unintentional) bias due to expectations.
A **control treatment** is typically a standard treatment with which we want to compare. It can also be no treatment at all.

3 Completely Randomized Design

We assume for the moment that the experimental units are homogeneous. We know how to compare two independent groups using the two-sample t-test. If we have more than two groups, this is not applicable anymore.

3.1 One-Way Analysis of Variance

On an abstract level we want to compare $g \geq 2$ treatments, having N experimental units, that we assign randomly to the different treatment groups having n_i observations each. This is what we call **completely randomized design**, it is the most elementary experimental design. If all the treatment groups have the same number of experimental units, we call the design **balanced**. Such random assignments can be done as follows:

```
sample(treat.ord)
```

3.1.1 Cell Means Model

Let y_{ij} be the observed response from the j-th experimental unit in treatment group i. In the **cells mean model** we allow each treatment group (cell) to have its own expected value. This means that y_{ij} is the realised value of the random variable:

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{or} \quad Y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

As for the standard two-sample t-test, the variance is assumed to be equal for all groups. We say that Y is the response and the treatment allocation is a categorical predictor. A categorical predictor is also called a factor. We sometimes distinguish between unordered (or nominal) and ordered (or ordinal) factors. We can rewrite the equation as:

$$\mu_i = \mu + \alpha_i$$

Where α_i is called the **treatment effect**. This will later help us to untangle the influence of multiple treatment factors on the response. Through this rewrite we have secretly introduced an additional parameter, to remove it again we need a side constraint. Possible constraints could be:

- weighted sum-to-zero: $\sum_{i=1}^{g} n_i \alpha_i = 0$
- sum-to-zero: $\sum_{i=0}^{g} a_i = 0$
- reference group: $\alpha_1 = 0$

For all of the choices it holds that μ determines some sort of "global level" of the data and α_i contains information about differences between the group means μ_i from that "global level". If we know $g-1$ of the α_i, we automatically know the remaining α_i, we also say that the treatment effect has $g-1$ degrees of freedom.

3.1.2 Parameter Estimation

We estimate the parameters using the least squares criterion:

$$\hat{\mu}, \hat{\alpha}_i = \underset{\mu, \alpha_i}{\operatorname{argmin}} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

Some notation:

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{i.} = \frac{1}{n_i} y_{i.}$$

$$y_{..} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{i.} = \frac{1}{N} y_{..}$$

As we can independently estimate the values of μ_i, one can show that $\hat{\mu}_i = \bar{y}_{i.}$. From $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$ we can get all the other parameters needed (they still depend on the side constraint).
The estimate of the error variance is also called mean squared error MS_E:

$$\hat{\sigma}^2 = MS_E = \frac{1}{N-g} SS_E$$

Where SS_E is the error or residual sum of square:

$$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2$$

3.1.3 Tests

With the two-sample t-test, we could test whether two samples share the same mean. We will now extend this for $g > 2$. Saying that all groups share the same mean is equivalent to saying:

$$Y_{ij} = \mu + \epsilon_{ij},\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

This is the so-called **single mean model**, a special case of the cell means model. We have the global null hypothesis

$$H_0 : \mu_1 = ... = \mu_g$$

vs. the alternative hypothesis

$$H_A : \mu_k \neq \mu_l \text{ for at least one pair } k \neq l$$

The idea is to check whether the variation between the different treatment groups (the "signal") is larger than the variation within the groups (the "noise"). We can decompose the total variation as follows:

$$\underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_{..})^2}_{SS_T} = \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}_{SS_{Trt}} + \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2}_{SS_E}$$

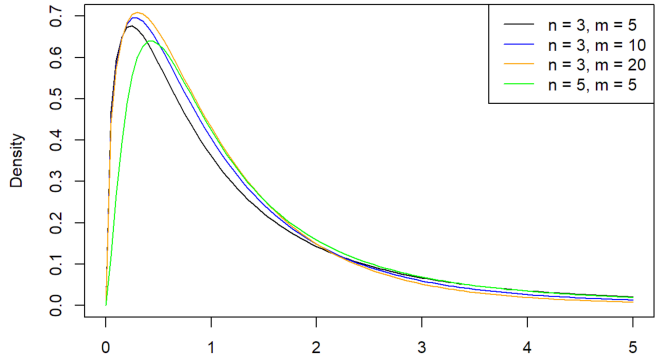
All this information can be summarized in a so-called **ANOVA** table.

Source	df	Sum of Squares	Mean Squares	F-ratio
Treatment	$g-1$	SS_{Trt}	$MS_{\text{Trt}} = \frac{SS_{\text{Trt}}}{g-1}$	$\frac{MS_{\text{Trt}}}{MS_E}$
Error	$N-g$	SS_E	$MS_E = \frac{SS_E}{N-g}$	

The MS and SS are normalized with the corresponding degrees of freedom. This is a so-called one-way ANOVA, because there is only one factor involved. If all groups share the same expected value, the treatment sum of squares is typically small. We introduce the so called F-ratio.

$$F - \text{ratio} = \frac{MS_{Trt}}{MS_E} \sim F_{g-1, N-g}$$

If the variation between groups is substantially larger than the variation within groups (higher F-ratio), we have evidence against the null hypothesis. The F-distribution looks as follows:



As with any other statistical test, we reject the null hypothesis if the observed value of the F-ratio, our test statistics, lies in an extreme region of the corresponding F-distribution. As this test is based on the F-ratio we call it an **F-test**.

3.2 Checking Model Assumptions

Statistical inference is only valid if all model assumptions are fulfilled. So far this means:

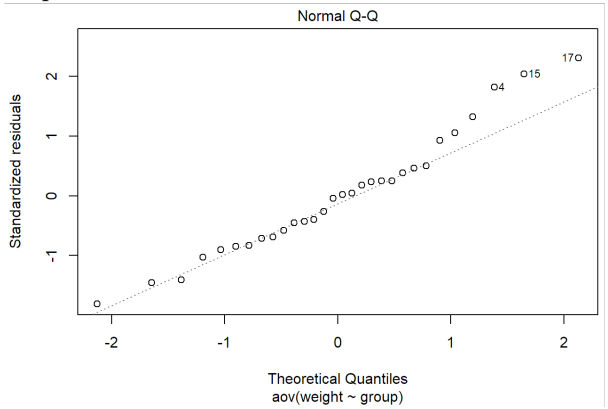
- The errors are independent
- The errors are normally distributed

- The error variance is constant
- The errors have mean zero

We now introduce different plots to check these assumptions. This means that we use graphical tools to perform qualitative checks.

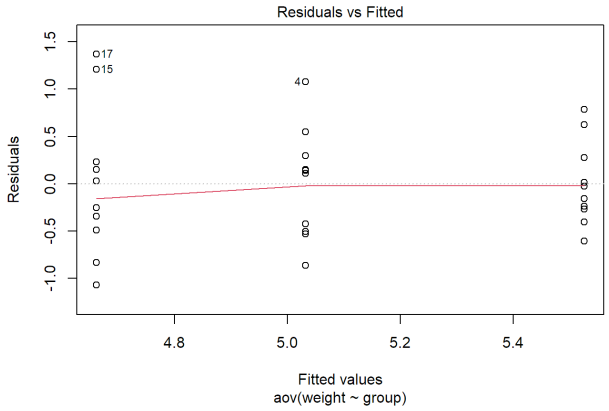
3.2.1 QQ-Plot

In a QQ-plot we plot the empirical quantiles of the residuals or "what we see in the data" vs. the theoretical quantiles or "what we expect from the model". The plot should show a more or less straight line if the normality assumption is correct.



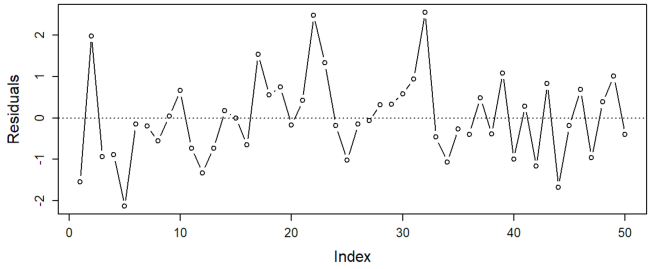
3.2.2 Tukey-Anscombe Plot

The Tukey-Anscombe plot (TA-plot) plots the residuals r_{ij} vs. the fitted values $\hat{\mu}_i$ (estimated cell means). It allows us to check whether the residuals have constant variance.



3.2.3 Index Plot

If the data has some serial structure, i.e., if observations were recorded in a certain time order, we typically want to check whether residuals close in time are more similar than residuals far apart. For this we use the index plot where we plot the residuals against time. For positively dependent residuals, we would see time periods where most residuals have the same sign, while for negatively dependent residuals, the residuals would jump too often from positive to negative compared to independent residuals.



4 Contrast and Multiple Testing

4.1 Contrast

The F-test is rather unspecific and gives us basically a yes/no answer. Often we have a more specific question than the global null hypothesis we want to answer. Such kind of questions can be formulated as so-called **contrasts**. As hypothesis we choose:

$$H_0 : \sum_{i=1}^{g} c_i \mu_i = 0 \text{ and } H_A : \sum_{i=1}^{g} c_i \mu_i \neq 0$$

Typically we have the side constraint that $\sum_{i=1}^{g} c_i = 0$. The contrast is about the differences between treatments and not about the overall response.
We estimate the value of $\sum_{i=1}^{g} c_i \mu_i$ with:

$$\sum_{i=1}^{g} c_i \hat{\mu}_i = \sum_{i=1}^{g} c_i \bar{y}_{i\cdot}$$

In addition, we could derive its accuracy (standard error), construct confidence intervals and do tests.

(3.1.2 Some Technical Details are left out on purpose)

4.2 Multiple Testing