L02: Descriptive statistics

“ I don't even know what I'm doing here.”

Chrom, Tron

Things we cover in this session

- Describing and visualizing data by boxplots and simple statistics

Things you need for this session

W01-2: Read the field data

Things to take home from this session

At the end of this session you should be able to

- Calculate simple descriptive statitistics of your dataset
- Create boxplots
- Interpret results based on boxplots

Descriptive statistics: min/max/mean/median/sd

To get an overview of your data descriptive statistics are an auxiliary tool. They help you to describe and understand the features of your dataset as the give you a short summary about the sample and measures of your data.

Comparing the mean and median values of your data can give you a good overview of your parameters. The mean value is calculated by summing up all values of the dataset of interest and divide it by the number of observations. Though the mean value is widely used to characterize datasets, it has the major disadvantage of being highly affected by outliers. The median, in contrast is the value which is located in the middle of an ordered dataset. Thus it is robust to outliers.

The standard deviation (sd) describes the spread of the data. It is the average deviation from each value to the mean value of the distribution.

Descriptive statistics: Do it in R

Luckily, as a R user you don't have to calculate these measures by hand. The functions

```
mean()
max()
```

```
min()
median()
sd()
```

will do it for you!

Boxplots

A boxplot is a useful visualization of the measures shown in the section above. It is therefore often used to depict the differences of distributions eg. between predicted and observed values.

<html> </html> *(Chen-Pan Liao [CC_BY_SA] via wikimedia.org) A Boxplot shows several components: - The **box** includes the distribution of the values located in the second and third quartil, thus of the 50% of values which are closest to the mean value. - The **median** is depicted by the line in the box. The whiskers and representation of outliers represent the spread of the values. - The **Whiskers** mark the remaining values which don't fall into the second and third quartile. The length of the whiskers is not standardizized. Often they are expanded to 1.5*the interquartile range (IQR). - The **interquartile range** is the range between the lowest value falling into the second quartile and the highest value falling into the third quartile. - All values which are higher than 1.5*IQR are considered as **outliers** and are usually marked by points over or under the whiskers, respectively. ===== Time for practice ===== W02-1: Explore the field data*