1. Philosophical foundations of empirical science and statistics

You are all students of science. But have you ever thought of what actually is the aim of science? Probably, we can all agree that the aim of science is to increase human knowledge. But how this is done? We may think that adding new pieces of knowledge to what is already known is actually the process of science. These new pieces come in the form of *universal statements (laws; theories)* describing natural processes. Some scientific disciplines, including biology, use data or experience to increase current knowledge and are thus called empirical science. Intuitively, we may assume that the new pieces of knowledge are first collected as the newly gathered experience or data (singular observations, statements) from which the theories and hypotheses (universal statements) are built. *Statistics* should then be the language of empirical science to summarize the data and make the inference of universal statements from the singular ones. This approach to empirical science would be called *induction*. Despite intuitive, it is ***not*** the approach we use in modern science to increase knowledge.

We may also agree that only *true* universal statements or theories represent a real addition to knowledge and may be used to infer correct *causal explanations*. So we should aim at truth, which should be an essential aspect of our scientific work. But how does science and scientists recognize the truth of their theories? This is not an easy task. Truth can be defined as a correspondence of statements with the facts[1]. But the question is how to measure such correspondence. There are two apparent ways: 1. We can believe authorities who may issue a judgement on this. The authorities may be of various kind: experienced scientists, distinguished professors, priests or books written by them (note that this is well compatible with the accumulative process of science described above) or 2. We can believe that truth is manifest – that truth is revealed by reason and everybody (who is not ignorant) can see it. The first way was largely applied in the Middle Age with the church, priests and the Bible as the authorities and ultimate source of truth. This led to a long-term stagnation of science and a few guys burnt at the stake. The second approach stems from the Renaissance thinking revolving against the dogmatic doctrines of the church. It was a foundation of many great discoveries made since the Renaissance time. Unfortunately, there is also devil hidden in this approach to truth. It lies in the fact that if truth is manifest, then those who cannot see it are either ignorant, or worse, pursue some evil intentions. Declaring itself as the only science-based approach to the society and politics, the Marxist-Leninist doctrine largely relies on the belief that its truth is obvious, which also provided justification for the ubiquitous cruel handling of its opponents whenever possible.[2,3]

---

[1] Facts (i.e. for instance measurements) are (usually) considered true. There is always sort of measurement error, but that is mostly negligible. Reporting false facts is unacceptable. It is basically cheating, which, if occurs, has a great negative effect on knowledge, because challenging published facts is something, which is rarely done.

[2] Note here that if the conflict between the Renaissance thinkers such as Galileo Galilei or Giordano Bruno and the church is viewed as a fight between the two views on truth both of which may lead to evil ends, you may reconsider the outright negative view on the representatives of the inquisition. Nevertheless, burning your opponents at the stake is not an acceptable means of discussion in any case.

[3] A strange mix of both approaches to truth is still largely applied in secondary education in some countries (e.g. Czechia). Textbooks and the teachers' knowledge may be used here as the ultimate authority for truth. At the same time, students are punished for making mistakes (by low grades) because truth is manifest. If they cannot see it, they are considered ignorant and as such deserve the punishment.

It seems that we have a problem with truth and need to find the way out of it. The solution of the problem was summarized the philosopher of science Karl R. Popper (1902-1994). Popper states that although truth exists and we should pursue it, we can never be sure that our theories are true. This is because we are prone to make mistakes with the interpretation of what our senses tell us. This view is not that novel as K.R. Popper himself refers to ancient Greek philosophers some of whom have identified this paradox of truth. One illustrative account of this is the story of prisoners in cave contained in Plato's Republic. This is the story about prisoners who are kept in a cave from the very beginning of their life and have their heads fixed to look at a wall. Fire is located far behind them and persons and objects pass between the fire and the prisoners' back casting shadows on the wall, which the prisoners can see. Then, as Plato says (by the speech of Socrates): "To them, I said, the truth would be literally nothing but the shadows of the images.". In this writing, Plato also declares ourselves to be like these prisoners. This may seem strange as we tend to believe that what we see is real but consider e.g. the recent observation of gravitational waves. We observe them by super-complicated and ultra-sensitive devices and can only see shadows of them (nobody can see them directly).

Box 1. Misleading empirical experience

1. Ancient Greek philosopher Anaximandros (*c*. 610 – *c*. 546 BC) was the first who identified the Earth as an individual celestial body and presented the first cosmology. This was a great achievement of human reason. However, he supposed the Earth to be of barrel shape because he only could see flat world around him – as we actually do.

Life of Anaximandros on barrel Earth

2. Jean-Baptiste Lamarck (1744-1829) formulated the first comprehensive evolutionary theory based on his naturalist experience with adaptations of organisms to their environment. He asserted that organisms adapt to their environment by adjustments of their bodies, which changes are inherited by the offspring. This is very intuitive but demonstrated to be false by a long series of experimental testing.

Although we can only see shadows of reality, these shadows still contain some information. We can actually use this information to make *estimates* about the reality and more importantly to demonstrate our universal statements ***false***. The ability to demonstrate some theories and hypotheses *false* is the principal strength of empirical science. This leads to rejection of theories demonstrated not to be true while those, for which the falsifying evidence is not available (yet) are retained. If a theory is rejected on the basis of falsifying evidence, a new one can be suggested to replace the false theory, but note, that this

new theory is never produced by an "objective" process based on the data. Instead, it is produced by subjective human reasoning (which aims to formulate the theory not to be in conflict with objective facts though).

In summary, experience can tell us that a theory is wrong but no experience can prove truth of a theory (note here, that we actually do not use the word "proof" in terminology of empirical science). Consider e.g. the universal statement "All plants are green". It is not important how many green plants you observe to prove it true. Instead, observation of e.g. single non-green parasitic *Orobanche* (Fig. 1.1) is enough to demonstrate that it is false. Our approach of doing science is thus *not* based on induction. Instead it is *hypothetical-deductive* as we formulate hypotheses and from them deduce how world should look like if the hypotheses were true. If such predictions can be *quantified*, their (dis)agreement with the reality can be measured by statistics. The use of statistics is however not limited to hypothesis testing. We also use statistics for *data exploration* and for *parameter estimates*.

Fig. 1.1. Non-green parasitic plant *Orobanche lutea*.

Finally, you may wonder how *Biostatistics* differs from *Statistics* in general. Well, there no fundamental theoretical difference, Biostatistics refers to application of statistical tools in biological disciplines. Biostatistics generally acknowledges, that biologists mostly fear maths so the mathematical roots of statistics are not discussed in details and also e.g. complicated formulae are avoided wherever possible.

Literature

Plato: Republic (Book VII)
Popper KR: Conjectures and Refutations
Popper KR: Logic of Scientific Discovery
https://en.wikipedia.org/wiki/Anaximander
https://en.wikipedia.org/wiki/Jean-Baptiste_Lamarck

2. Data exploration and data types

If you have some data, say a variable describing observations of 100 objects (e.g. tail length of 100 rats), you may wish to explore these values to be able to say something about these data. That is, you may wish to describe the data using *descriptive statistics*.
The data are here:

```
  [1] 4.57 5.69 4.49 6.09 5.46 6.28 4.90 5.80 4.39 4.32 4.85 4.05 6.36 3.10 5.30 3.74 5.45 4.08
 [19] 4.97 3.31 4.71 5.49 6.37 5.32 5.31 5.20 2.29 3.91 4.09 5.59 6.85 3.56 6.13 3.73 6.41 4.01
 [37] 4.77 5.84 6.37 6.49 5.27 5.26 5.92 5.27 4.17 7.00 4.73 5.26 5.17 3.76 7.03 6.79 5.94 7.42
 [55] 5.87 5.61 5.25 4.45 4.41 7.27 5.53 5.69 3.59 5.47 5.69 3.63 2.03 5.65 3.36 3.60 5.39 3.90
 [73] 5.82 3.17 3.73 4.81 4.70 4.71 5.02 5.61 2.99 3.96 3.28 4.99 5.30 5.23 6.06 6.31 5.60 5.85
 [91] 5.15 4.62 5.79 5.36 3.89 4.35 5.26 3.76 4.68 5.77
```

First, we need to know the *size* of the data, i.e. number of observations (*n*).

Here *n* = 100.

Second, we are interested is the central tendency, i.e. certain middle value around which, the data are located. This is provided by the ***median***. Which is the middle value[4] of the ordered data dataset from the lowest to the highest value. Here *med* = 5.24

Third, we need to know the spread of the data. A simple characteristic is ***range*** (***minimum*** and ***maximum***. Here *min* = 2.03 and *max* = 7.42. However, the minima and maxima may be affected by ***outliers*** and ***extremes***. While, it is useful to know them, we may also prefer some more robust characteristics. This comes with ***quartiles***. Quartiles are 25% and 75% quantiles. ***XX%-quantile*** refers to a value compared to which XX% of other observations are lower. In our case the first quartile (25%) = 4.15 and the third quartile (75%) = 5.71. The second (50%) quartile is the median.

These descriptive statistics can be summarized graphically in the form of ***boxplot***. That is very useful for comparisons between different datasets (e.g. comparison of mouse tail length with a similar dataset on rats):

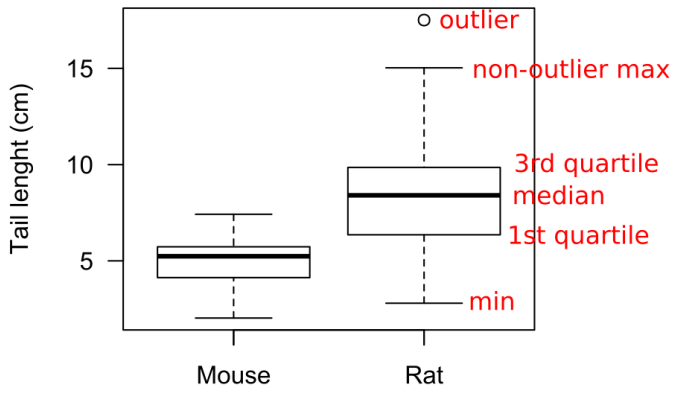

Fig. 2.1. Boxplot displaying tail length of mice and rats. The bold lines in boxes represent medians, boxes represent quartiles (i.e. 25 and 75% quantiles) and the lines extending from the box boundaries (whiskers) represent the range or non-outlier range of values, whichever is smaller. The non-outlier range is defined as the interval between (25% quantile) 1.5 × interquartile range) and (75% quantile + 1.5 × interquartile range). Any point outside this interval is considered an outlier and is depicted separately.[5]

Another useful type of plot is the ***histogram***. Histogram is very useful for displaying data distributions (but less so for comparisons between different datasets). To plot a histogram,

[4] Note here, that if *n* is even and the two values close to the middle are not equal, median is computed as their arithmetic mean.

[5] This is a very detailed description of a boxplot. Usually it can be briefer. Still, I was forced to make it this detailed by the editor of one paper I published.

values of the variable are assigned into intervals (called also bins). Numbers of observation (frequency) within each bin is then plotted on in the graph.

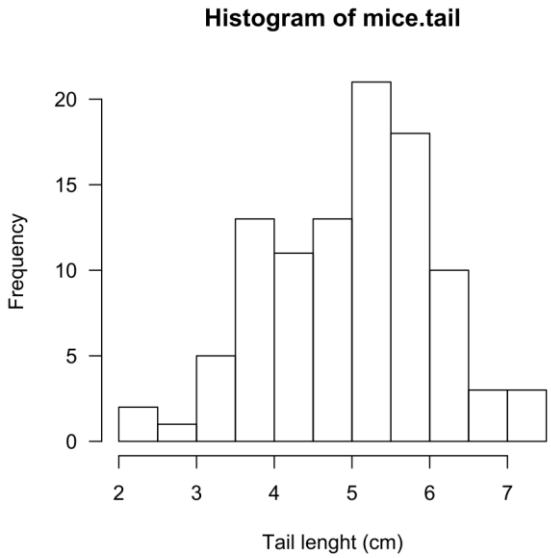

Fig. 2.2. Histogram of mouse tail length.

Types of data

The data on mouse tail length we have explored are called ***data on ratio scale***. Several other types of data can be defined on the basis of their properties. These are summarized in Table 2.1. in ratio-scale and interval data, further distinction can be made between continuous and discrete data but that makes little difference for practical computation.

Table 2.1. Summary of data types definition and properties.

Data type	Criteria	Possible math. operations	Examples	Object class in R
Ratio scale data	constant intervals between values, meaningful zero	+,-,×,/	length, mass, temperature in K	numeric
Interval scale data	constant intervals between values, zero not meaningful	+,-	temperature in °C	numeric
Ordinal data (also called semi-quantitative)	variable intervals between values	comparison of values	exam grades, Braun-Blanquet cover	numeric (but may require conversion)
Categorical data	non-numeric values	none	colors, sex, species identity	character, factor

Categorical variables cannot be explored by the methods described above. Instead, frequencies of individual categories can be summarized in a table, or a ***barplot*** can be used to illustrate the data graphically.

Consider e.g. 163 bean plant individual with flowers of three colors: white, red, purple.

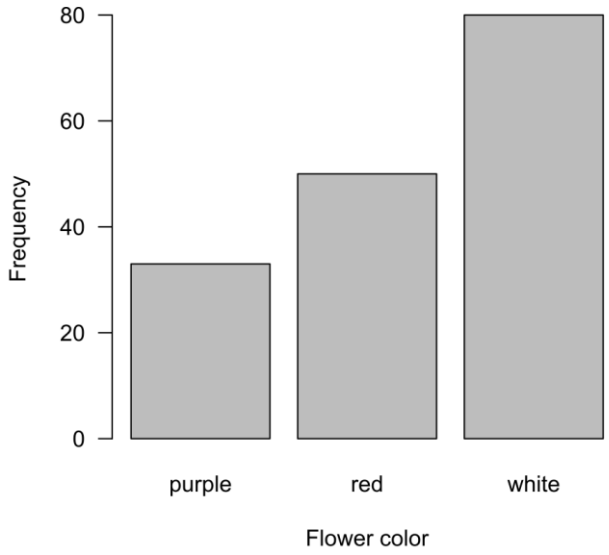

Fig 2.3. Barplot of frequencies of flower colors in the bean dataset.

How to do in R
Obtaining R: download from https://cloud.r-project.org/ and install
Obtaining R Studio: download from https://rstudio.com/products/rstudio/download/#download and download

Installing an R package
install.package("package")

Loading an R package for use
library(package)

Importing data:

1. Using clipboard: read.delim("clipboard") with **decimal point** format or **read.delim2("clipboard")** with **decimal comma** format
2. Directly from excel: install and load the package readxl, then **read_xlsx("file")**

Size of data: function **length**
Median: function **median**
Range: function **range**
Minimum: function **min**
Maximum: function **max**
Quartiles: function **quantile** with default settings produces 5 values: min, lower quartile, median, upper quartile, maximum

`Boxplot: function` **`boxplot`** supports the **formula notation**, i.e. response variable ~ classifying variable)
`Histogram: function` **`hist`**
`Barplot: function` **`barplot`** requires frequencies to be provided e.g. by `table` or `tapply`