FLORIAN HARTIG

ESSENTIAL STATISTICS

Course material for the MSc model Research Skills, University of Freiburg (website here)

Comments / questions to:

Florian Hartig
University of Freiburg
Germany

Contents

1
Introduction

1.1 Purpose and intended audience

This document provides a short introduction to the field of statistics, as well as to statistical analyses most likely to be encountered in elementary experimental and observational situations. It is intended as a primer and will not replace a more thorough lecture or textbook. Regarding the latter, I provide recommendations in the final section 7.2.

1.2 Topics of statistics and data science

Statistics, or more broadly data science, deals with the visualization, summary and interpretation of data. This primer provides an introduction the four most important pillars of statistical methods for a quantitative researcher:

Descriptive statistics: Descriptive statistics include summary statistics such as mean and median as well as the various options to visualize data.

Descriptive statistics = plots, summary statistics

Inferential statistics: Inferential statistics deals with testing hypothesis and estimating parameters. Inference is typically based on assumptions that are summarized in a statistical model. A statistical model is often also called the “data-generating process”, because it describes the assumptions about the processes that lead to variation in the data (systematic and stochastic).

Inferential statistics = parameter estimates, p-values, tests

A statistical model describes how the data was generated = data-generating process

Predictive statistics and machine learning: Predictive statistics and machine learning deals with deriving predictions from data, in particular from ”big data”. The main difference to inferential statistics is that the focus is on developing methods to make good predictions, without the necessity to describe, infer or test assumptions about the data-generating process.

Machine learning = predictive models. Big data = large datasets, e.g. from consumers on Amazon

Experimental design: Experimental design covers all aspects of generating data, in particular questions such as “which variables need

Experimental design or study design = how to obtain and create data to be measured?", "how many replicates do we need?", "how should we optimally vary the variables of interest in an experiment"?

1.3 The R environment for statistical computing

For the better or worse, the times when statistics was done with pen, paper and a calculator are over. Statistical analysis nowadays happens on the computer, and a number of software environments exist to do so. In the ecological sciences, R has become the de-facto standard for statistical analysis. R is open-source, free, and has a very larger user base, specially in the environmental sciences, that contribute packages for specialised ecological and environmental analysis.

R is a script-based language, which means that you communicate with the computer not by clicking on buttons, but by writing commands either directly in the R console, or first in a text file and then sending it to the R console, which is then evaluating your commands. If you aren't used to this kind of approach yet, it may take a short while to get used to this, but once you are used to it, you will notice how advantageous it is to have all steps of your analysis listed in a text file, being able to repeat everything at any moment.

In this primer, all examples will be calculated with R; however, the focus will not be to introduce R itself. If you need one, follow this link for an introduction, including help on how to install the software. I highly recommend to using R together with RStudio. Follow the link above to get further information

2
Data, sample and population

The following four chapters will be devoted to the four types of statistical analysis described in the introduction: descriptive statistics, inferential statistics, predictive statistics and experimental design. Before we go into these topics, however, a few short comments on how data arises, and how we represent it.

2.1 Sample, population, and the data-generating process

The very reason for doing statistics is that the data that we observe is somehow random. But how does this randomness arise?

Imagine that we are interested in the average growth rate of trees in Germany during two consecutive years. Ideally, we would measure them all and be done, without having to do statistics. In practice, however, we hardly ever have the resources to do so. We therefore have to make a selection of trees, and infer the growth rate of all trees from that. The statistical term for all the trees is the "population", and the term fro the trees that you have observed is the "sample". Hence, we want to infer properties of the population from a sample.

The population is the set of all observations that you could have made. The sample is the observations that you have actually made

The population as such is fixed and does not change, but time we observe a random selection (sample) of the population, we may get elements with slightly different properties. As a concrete example: imagine we have the resources to sample 1000 trees across Germamy. Therefore, every time we take a random selection of 1000 trees out of the population, we will get a slightly different average growth rate.

Sampling creates randomness

The process of sampling from the population does explains how randomness arises in our data. However, a slight issue with this concept is that it does not match very well with more complex random processes. Imagine, for example, that data arises from a person going to randomly selected plots to measure radiation (which varies within minutes due to cloud cover), using a measurement instrument that measures with some random error. Does it really make sense to think of the data arising from sampling from a "population" of possible observations?

However, not all randomness comes from sampling from a population

A more modern and general concept to describe how data is created is the the concept of the "data-generating process", which is pretty self-explenatory: the data-generating process describes how

A more modern concept that replaces the "population" is the "data-generating process". The data-generating process describes how the observations from a random sample arise, including systematic and stochastic processes the observations from a random sample arise, including systematic and stochastic processes. It therefore includes the properties of what would classically be called ”sampling from a population”, but it is broader and includes all other processes that would create systematic and random patterns in our data. In this picture, instead of infer properties of the population from a sample, we would say we want to infer the properties of the data-generating process from a sample of observations created by this process.

Whether you think in populations or data-generating processes: the important point to remember from this section is that there are two objects that we have to distinguish well: on the one hand, there is our sample. We may describe it in terms of it's properties (mean, minimum, maximum), but the sample is not the final goal. Ultimately, we want to infer the properties of the population / data-generating process from the sample. We will explain how to do this in the next sections, in particular in the section on inferential statistics.

2.2 Representation and classes of data

Before we come to that, however, let us talk in a bit more detail about the representation of the sample, i.e. the data that we observe. A typical dataset consists of multiple observations of number of variables (e.g. temperature, precipitation, growth). You can think of this situation as a spreadsheet where the columns are the variables and the rows are the observations. Of course, there are other data structures, but this is the most common one.

Usually, this data will contain one variable that is our focus, meaning that we want to understand how this variable is influenced by other variables. We call this variable the “response variable” (sometimes also the dependent variable or outcome variable), because we are interested if and how this variable of interest varies (responds, depends) when something else changes. The variables that affect the response could be an environmental factors (e.g. temperature), it could be experimental treatments (fertilized vs. non fertilized), or anything else. Those other variables that affect our response are called “predictor variables” (synonymous terms are explanatory variables, covariates or independent variables).

The response variable is the variable for which we try to understand how it responds to other factors.

The predictor variables are those that affect the response.

The most common case is that the response variable is a single variable (e.g. a single number or a categorical outcome), and we will concentrate on this case. However, there are cases when the response has more than one dimension, or when we are interested in the change of several response variables at a time. The analysis of such data is known as multivariate statistics. We will not cover this methods here, but if you need it some further links are here.

Multivariate statistics deals with response variables that have several dimensions, such as species composition

Another important distinction is the type of each variables Independent of whether we are speaking about the response or the predictor, we distinguish:

Variables can be continous, discrete or categorical. Categorical variables can be ordered, unordered, or binary.

- Continuous numeric variables (ordered and continous / real), e.g.

temperature

- Integer numeric variables (ordered, integer), e.g. count data
- Categorical variables (e.g. a fixed set of options such as red, green blue), which can further be divided into
  - Unordered categorical variables (Nominal) such as red, green, blue
  - Binary (Dichotomous) variables (dead / survived)
  - Ordered categorical variables (small, medium, large)

It is important that you record the variables according to their "nature". And if you use a statistics software, you have to make sure that the type is properly recognized after reading in the data, because many methods treat a variable differently if it is numeric of categorical.

Check that your variables have the right type after reading them in in your statistic software

Experience shows that there is certain tendency of beginners to use categorical variables for things that are actually continuous, e.g. by coding body weight of animals into light, medium, heavy. The justification stated is often that this avoids the measurement uncertainty. In short: it doesn't, it just creates more problems. Don't use categorical variables for things that can also be recorded numerically!

Don't use categorical variables for things that can also be recorded numerically!

In R:

To represent data, R has a basic data structure, the data.frame. A data frame is like a spread sheet, with columns, and each column can have a different type. Possibilities are

- integer - what it says
- numeric - continuos number (float)
- boolean - true / false
- factor - normally unordered, i.e. red, green, blue. Can also be made ordered (small, medium, large), although it is then often better to code this as an integer

Also, although not really a data type, a common case is the absence of an observation for a particular variable. This is typicall code by ”NA”. Similar, but not the same is ”NaN” (not a number), which occurs as the result of a calculation that cannot be performed.

To read in data as a data.frame in R-studio, go to the upper right part, ”Import Dataset”. For more explanation, see http://biometry.github.io/APES/R/R20-DataStructures.html

It is very important that you check that your columns have the right type after reading in the data. To do so, type in str(TheNameOfMyData), and check the type of the columns

3
Descriptive statstics and visualization

Descriptive statistics deals with the summary and visualization of (sampled) data

We will speak about how to create data in the later chapter 6 on experimental design.

3.1 Summary statistics

Summary statistics are numerical calculations that summarize a dataset. They are used to display the properties of a dataset in a more compact way.

Summary statistics summarize data

Summarizing a single continous variable

A common situation in which we want to use summary statistics are repeated observations of a continous variable. If it helps you, imagine we have measured and remeasured 2000 trees, and we have now a distribution of observed diameter increments (see Fig.3.1).

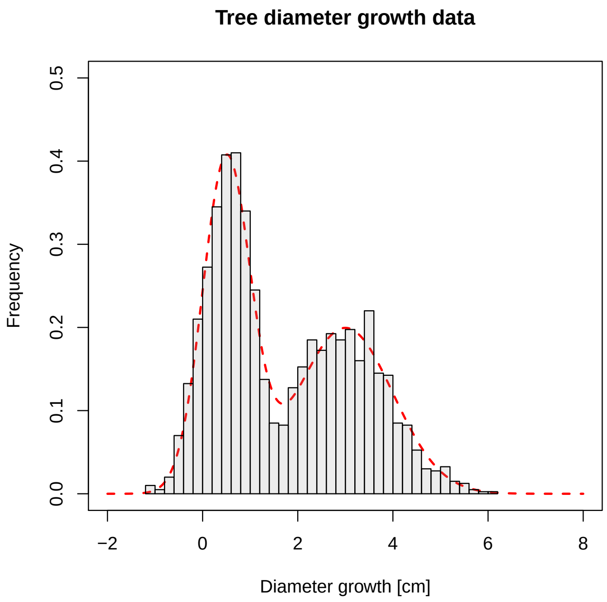


Figure 3.1: A distribution of observed values diameter increment values (gray bars). We assume (in this case I know it) that these values come from some true distribution (population or data-generating process) that I plotted here in dashed red color. If we would draw more and more data, the gray bars would approach the true distribution

How can we summarize the properties of this observed sample? Some basic properties would be the minimum and the maximum value, the mean, or the mode (the maximum of the distribution, i.e. value with the highest density of observations). There are two further central summaries that are much used: moments and quantiles.

The moments may not sound familiar to you, but you have probably already used the first and the second moment of the distribution, which are known as the mean and the standard deviation. In general, the n-th moment is defined as the average across the distribution, taking each observed value to the power of n.

$$< x^2 >_d \tag{3.1}$$

where <> denotes the average, and d the distribution. The first four moments are called the mean (average value), the standard deviation (measure of spread), the skewness (measure of asymmetry in the distribution) and the kurtosis.

Quantiles are the second central class of summary statistics for describing continous distributions. If we have a distribution such as the figure above, we can ask ourself: which is the value of the variable that divides the data in half, so that half of the observed data are lower, and half are higher than this value? This point is called the median, and also the 0.5 quantile. More general, the 0.x quantile is the value at which a fraction of 0.x of the data is smaller.

Half of the data are lower, and half of the data are higher than the 0.5 quantile, which is called the median

Correlation - summarizing the dependence between continous variables

A second important case for summary statistics is correlation. Correlation methods measure the dependence between continous variables. Unfortunately, there are quite a number of measures of correlation, and it is imporant to distinguish between them. The two most important are:

Linear coefficients: Linear coefficients, most notably the widely used "Pearson's correlation coefficient", measure the linear dependence between two variables. Pearson's correlation coefficient is widely used because it computes fast and is easily interpretable. However, it can be misleading if variables are not in a linear dependence. This effect is displayed in Fig. 3.2.

Rank correlation coefficients: Rank correlation coefficients, such as “Spearman's rank correlation coefficient” and “Kendall tau rank correlation coefficient” measure how well the variables match in their tendency to increase or decrease, without considering the extend or linearity of this increase. They are preferable if you think variables could be in a nonlinear relationship.

Strong correlation != important effect: Also visible in Fig. 3.2 is an often misunderstood property of correlations and dependence - a high correlation coefficient does not mean that a variable has a strong reaction to another variable. All that is needed to obtain a high correlation coefficient is little spread around the line (see middle row - the effects are different, but the correlation is the same).

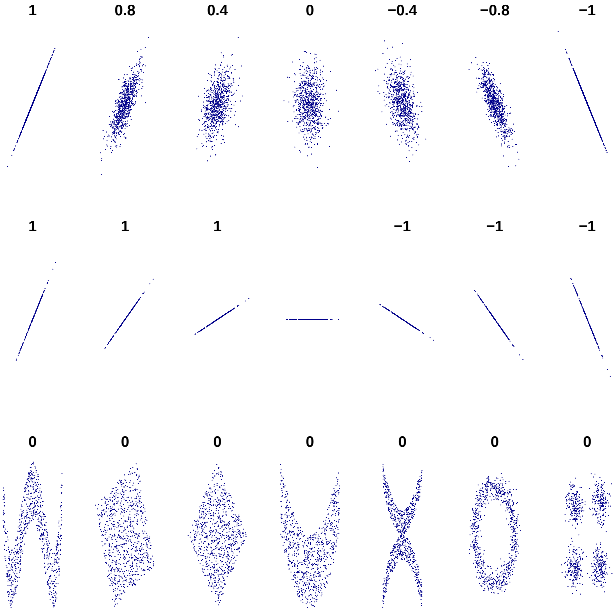


Figure 3.2: Demonstration of possible correlations with the Pearson's correlation coefficients. Note that many datasets that show a clear dependence between variables are assigned a Pearson's correlation coefficient of 0 because the dependence is not linear.

Contingency tables - summarizing disrecte outcomes of several variables

Finally, a classic concept for summarizing binary or categorical data are contingency tables. Here an example from a classical dataset available in R on aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex. I show only the first department.

This dataset from Berkeley is a famous example for the Simpson's paradox. Read up on wikipedia about this important statistical trap.

```
> UCBAdmissions[,,1]
```

```
          Gender
Admit       Male Female
  Admitted   512     89
  Rejected   313     19
```

In R:

For the various options to calculate descriptive statistics in R, see here

3.2 *Visualization*

Summary statistics are useful, but also dangerous. A famous example is Anscombe's Quartet, a hypothetical dataset of four observations that are identical in classical summary statistics such as mean, variance, correlation, regression line, etc. (Anscombe, 1973).

Type ?anscombe in R to see the code to create these plots and to calculate the statistical properties of the datastes

It is therefore very useful to get a graphical overview of your data, additionally to the summary statistics that you may calculate.

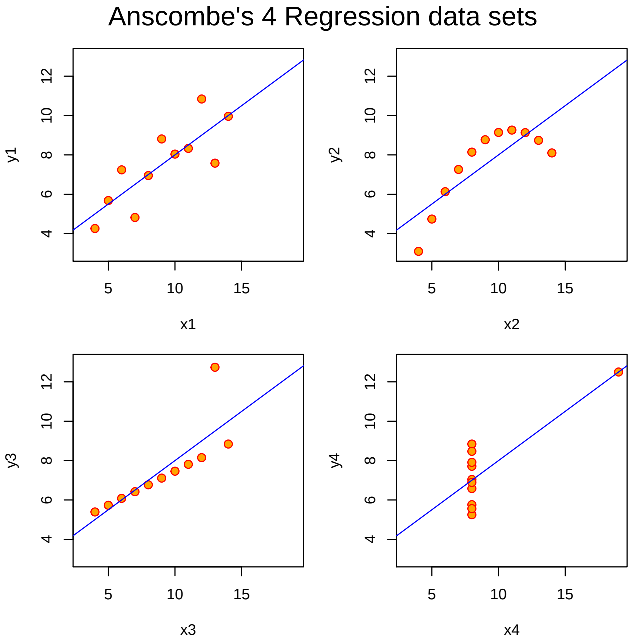


Figure 3.3: Anscombe's Quartet

Principles of visualization

Examples of distorting graphics here

The principle of graphics and visualization is to represent the data as accessible and truthful as possible. The reader should get the best possible overview about the data in the shortest possible time. And, of course, the graphics should look nice as well. First of all, some general hints that may help:

- Simple is better than complicated
- Avoid excessive color. Graphics should be b/w readable if possible (use a color gradient that is a gradient in intensity at the same time, use dashing of lines additional to colors). If your graphic relies on color, try to choose colors that can be read by color blinds (avoid red/green).
- Truthfulness: avoid distortions. Use quadratic figures unless there are particular reasons. Axis should start at 0 unless you have good reasons not to. If presenting several graphics, use the same scale unless there are good reasons for it.
- Don't manipulate graphics by hand
- Output in a vector format (pdf, eps, svg)

Graph types

There is a large, nearly infinte number of possible visual representation of data. I provide here four very common graph types.

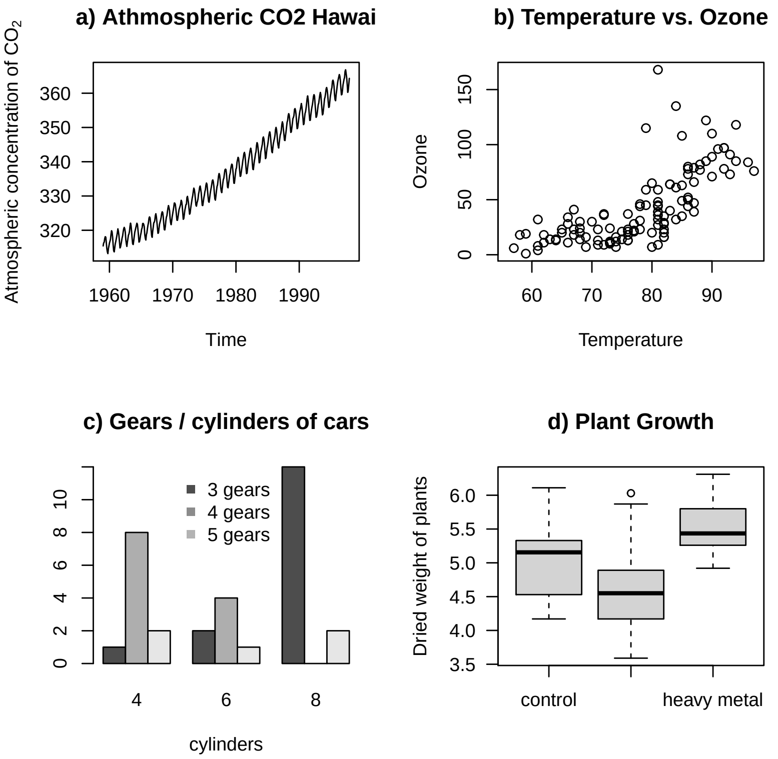


Figure 3.4: Four typical plot types, from top left to bottom right: a) A line plot to represent continous measurements of one variable; b) a scatter plot to represent the relationship between two continous variables; c) a bar plot to represent measurements in discrete groups / variables; d) a box plot to represent repeated continous measurements in discrete groups.

Line plots: Line plots are used to visualize continous, ordered measurements. Typical examples would be time series, continous variations of a parameter, or a mathematical function. Example in Fig. 3.4a.

Scatter plots: Scatter plots show two continous continous variables that are measured in pairs. Typical example is when you have repeated measurements of several variables, and you want to see if they correlate. Example in Fig. 3.4b.

Bar plots: Bar plots show an information (counts, or a continous variable) for discrete groups. Example in Fig. 3.4c.

Box plots: Box plots are very common to show the distribution of a continous variable across several discrete groups. They typically consist of a box, whiskers (lines), and potentially points around the whiskers. What those means depends on the software used to create the plots, but the typical interpretation is that the box covers the central 50%, with the median in indicated in the middle. The whiskers aim at providing an estimate of the range of your data, except for outliers. Of course, what is counted as an outlier depends on your assumptions. If you must know, the technical definition is that the whiskers are at the most distant observation less than or equal to the upper quartile plus 1.5 the length of the interquartile range. Example in Fig. 3.4d.

In R:

There are number of excellent introductions to graphics with R, so that I won't bother to provide details here. As a start, I recommend looking at exercise on graphics with R that accompanies this primer, as well as at

- QuickR
- RGraphCompendium
- rexrepos

4
Inferential statistics

Inferential statistics is concerned with inference, i.e. drawing conclusions from observations. Inference is not always, but in most cases, connected to the idea of a data-generating model.

Statistical inference is drawing conclusions from observations using statistical methods

4.1 Obtaining a data-generating model

Why do we need the idea of a data-generating model to draw conclusions from data? Imagine we want to know whether plant growth can be affected by music. We might take two pots, each with a plant, and expose one to classical music and the other to heavy metal. Inevitably, one of them will grow taller, but this could be by chance, as there is always some variation in the growth rates.

In statistics, one often uses the word treatment to describe manipulations to experimental units. Here, the two types of music would be called treatments. The particular treatment of ”doing nothing” is called the control.

Hence, we need more repetitions. Let’s say we take a few more pots, 30 in the hypothetical case I show below

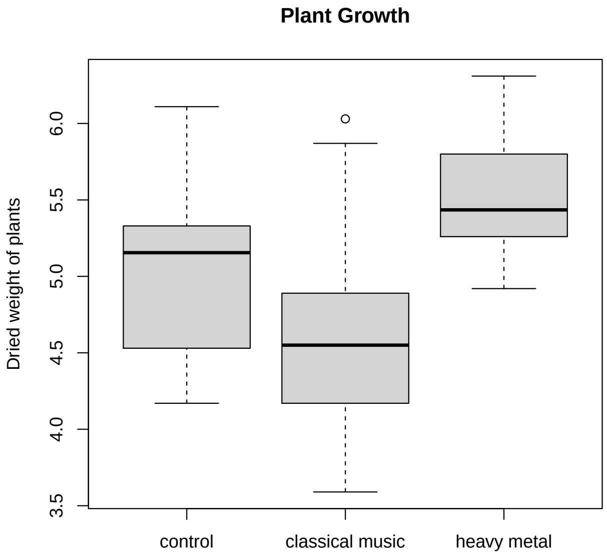


Figure 4.1: Groth measurements under three different treatments

There seem to be some differences between the three cases, but there is also quite a bit of variation in plant growth within each treatment (we have on average seven observations per treatment

Remember the interpretation of the boxplots - the strong line in the middle is the median. The box covers the central 0.5 quantile of the distribution.

here). Hence, it is still possible that the differences that we observe have arisen by chance.

If we want to say something definite about the probability of a difference between those two treatments and the control, we need to make a model that describes the stochastic variation in the data, which then allows us to calculate things such as the probability of the observed differences arising by chance. These assumptions are what we call a statistical model (eqivalent: stochastic process, data-generating model). .

A statistical model describes how the response variable arise as a functiono of the predictors as well as some random (stochastic) processes

The more common class of models used for this purpose are parametric statistical models. For the data that we have here, a parametric statistical model might, for example, make the assumptions that there is a mean growth rate for each treatment (control, classical and heavy metal), but that the growth of each individual plant varies with a normal distribution around the mean growth of its respective group. The parameters of this model are the unknown mean growth rates and the variance of the normal distribution. Those parameters are then fit to the data with methods explained in this chapter, and based on the fit one can calculate, for example, the probability that the data would arise if there was no difference between the groups.

Parametric statistics uses statistical models that describe the data-generating process in terms of functions and distributions that have parameters that then need to be fit

The other option to obtain a data-generating model are non-parametric methods. Non-parametric methods bypass the necessity of making assumptions, e.g. about the distribution of the data, typically by randomizing or resampling the data itself. How does this work? For the plant growth, for example, we could answer the question of how likely it is to see the observed data if there is no difference between groups also without making an assumption about the distribution - we simply throw all observations in one pot, irrespective of their treatment, and then re-distribute them randomly on the three treatments. If we do this many times (e.g. 1000 times), we can get a good idea how likely it would be to obtain the observed differences if treatment has no effect.

non-parametric statistics tries to avoid making assumptions about the data-generating process. Typically, the data-generating process is emulated by using the data itself, e.g. by resampling methods.

Non-parametric methods are an important branch of modern statistics. Their advantage is obviously that they don't make assumptions. On the other hand, parametric methods are typically much faster, and if their assumptions are correct, they are more sensitive and powerful, meaning that, with the same amount of data, they are more likely to detect an effect if it is there. For the latter two reasons, parametric methods are currently the basis of most statistical analysis.

The higher sensitivity of parametric methods relies on the fact that they make assumptions, which, in a sense, are like additional data. Of course, all results then rely on those assumptions being correct. We will talk about how to check parametric assumptions later in the chapter.

4.2 Inferential outputs

Based on the data-generating model (parametric or non-parametric), we can now apply different inferential procedures to draw conclusion about our data (in our case: to decide if music makes a difference or not). In standard statistics, there are two main inferential procedures that are applied in all kind of settings and models, and the outputs of these two procedures are: p-values and maximum likelihood estimates. A third procedure, the posterior calculated by Bayesian inference, has become more fashionable lately. I will mention it shortly at the end of this section

p-values

The use of p-values is connected to the inferential method of null hypothesis significance testing (NHST). The idea is the following: if we have some data observed, and we have a statistical model, we can use this statistical model to specify a fixed hypothesis about how the data did arise. For the example with the plants and music, this hypothesis could be: music has no influence on plants, all differences we see are due to random variation between individuals. Such a scenario is called the null hypothesis. Although it is very typical to use the assumption of no effect as null-hypothesis, note that it is really your choice, and you could use anything as null hypothesis, also the assumption: "classical music doubles the growth of plants". It's the analyst's choice what to fix as null hypothesis, which is part of the reason why you can choose among such a large number of available tests. We will see a few of them in the following chapter about important hypothesis tests.

A null hypothesis H_0 is a fixed scenario that makes predictions about the expected probabilities of different observations.

If we have a null hypothesis, we calculate the probability that we would see the observed data or data more extreme under this scenario. This allows us to test whether our null hypothesis is compatible with the data, and thus called a hypothesis tests. We call the probability to see the data or more extreme the p-value.

The p-value is the probability to see the data or more extreme data under the null-hypothesis.

$$p := p(d >= D_{obs}|H_0) \tag{4.1}$$

If the p-value falls under a certain level (the significance level α), we say we have significant evidence to reject the null hypothesis. The level of α is a convention, in ecology we chose typically 0.05, so if a p-value falls below 0.05, we can reject the null hypothesis.

if $p<0.05$, we say we have significant evidence to reject the null-hypothesis.

A problem with hypothesis tests and p-values is that their results are notoriously misinterpreted. The p-value is NOT the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false, although many authors have made the mistake of interpreting it like that (Cohen, 1994). Rather, the idea of p-values is to control the rate of false positives (Type I error). When doing hypothesis tests on random data, with an α level of 0.05, one will get exactly 5% false positives. Not more and not less.

Parameter estimates

The second type of output that is reported by most statistical methods are maximum-likelihood parameter estimates. In a nutshell, the maximum-likelihood estimate (MLE) is our best estimate for the parameters in our model (e.g. a difference between the treatments and the control in our example).

In a bit more detail: in statistics, we define the likelihood as a function of the model parameters θ as

$$L(\theta) := p(dD_{obs}|M(\theta)) \tag{4.2}$$

, i.e. as the function that is obtained by calculating the probability of obtaining the observed data when we vary the model parameters.

The maximum-likelihood estimate (MLE) is then defined as the combination of parameters or model assumptions for which the likelihood is maximal. So, while the p-value is the probability of the observed data or more extreme data under a fixed (null) hypothesis, the MLE gives us the "hypothesis" that has the highest probability to produce the observed data

It is important to note that the MLE is the parameter set for which the data is most likely, not the most likely parameter set!

The MLE is only a single parameter value. This type of estimate is often called a point estimate. However, a point estimate is typically of little use if we don't know how certain it is. Therefore, parameter estimates are usually accompanied by confidence intervals. Sloppily you can think of the 95% confidence interval of a parameter as the probable range for this area. Somewhat confusing for many, it's not the interval in which the true parameter lies with 95% probability. Rather, in a repeated experiments, the standard 95% CI will contain the true value in 95% of all cases. It's a subtle, but important difference. However, for now, don't worry about it. The CI interval is roughly the range within which we expect the true parameter to be.

A point estimate is something like a single best estimate.

Confidence intervals provide an estimate of uncertainty around the point estimate.

Bayesian estimates

To conclude our overview on inferential products, one method is still lacking - Bayesian methods calculate a quantity that is called the posterior parameter estimate. It is similar, but not identical to the parameter estimates discussed previously. You won't need this here, but if you want to know more about those, have a look at (Gelman et al., 2003) and at my website Learning Bayes.

Bayesian methods calculate a third quantity, the posterior probability. Although they can be used for any model, Bayesian methods tend to be used for more advanced statistics.

Different methods != different models

You know now that there are three different things that statisticians typically calculate: p-values, MLE, and the posterior. For a given data-generating process, you can always calcuate any of the three.

I make this point because many people wrongly assume that they use different models when they are indeed only using different ways to evaluate them. An example is the case of ANOVA, t-tests and linear regression. All of them are based on the same data-generating process - some fixed effects between groups and an iid normal observation error on top. ANOVA and t-tests specify different null-hypothesis, and the linear regression searches for the MLEs. You could additionally calculate the Bayesian posterior if you wanted.

ANOVA , t-tests and linear regression are only different evaluations of the same model

4.3 Important hypothesis tests

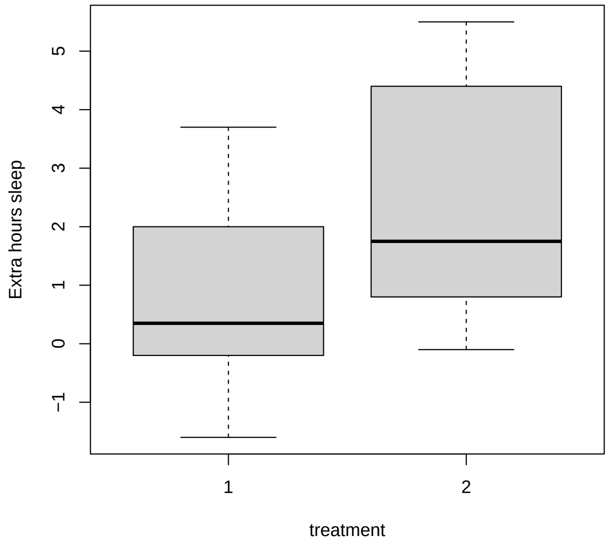
After having discussed the basic statistical outputs, lets move to practice and get to know the two probably most applied hypothesis tests, the t-test and ANOVA. I told you already that they are based on the same data-generating process, but specify slightly different null-hypotheses.

t-test

A t-test tests for differences between the means of two normally distributed samples; or if there is only one sample, between 0 and the mean of the sample. Again, the statistical model underlying is that of a normally distributed response, and the null hypothesis is that there is no difference in the mean of this normal distribution for the two groups, respectively that the sample mean is 0 if we have only one group. Also, depending on the software used, there are usually a number of adjustments possible, e.g. relaxing the assumption that the two groups have the same variance. Here, I show an example in R, using the classical data from Student (1908). The data show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.

To stay in our previous classification, we would say the response variable in continuous, and the predictor is categorical (group 1 or group 2), or, if there is only one group, there is no predictor.

Figure 4.2: Data from Student (1908)



```
> ## Traditional interface
> with(sleep, t.test(sleep$extra[sleep$group == 1], extra[group == 2]))

> ## Formula interface
> t.test(extra ~ group, data = sleep)

        Welch Two Sample t-test
```

```
data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

Note that the output provides a p-value (H0 = no difference), but also the maximum-likelihood estimate for the difference of the means, together with the confidence intervals. This is goes beyond the classical t-test, but probably the programmers assumed that you would also want to have the best estimate for the difference of the means.

Suggestion for reporting this result: p>0.05: differences between groups were not significant. p<0.05: we found a difference of X +- Confidence interval between the groups (p-value for difference from a t-test was X).

Analysis of variance (ANOVA)

ANOVA or analysis of variance can mean different things to different people. The standard ANOVA makes basically the same assumptions as a t-test (normally distributed responses), but allows for more than two groups. More precisely, it tests if the measured response (i.e. the dependent variable) is influenced by one or several categorical variables that could have two or more levels could also interact. An interaction between two variables means that the value of one explanatory variable affects how strongly another explanatory variable affects the response.

An interaction = one variable modifies the effect of another variable

While the word ANOVA is generally associated with the assumption explained above (which correspond to a t-test / linear regression, see next chapter), the concept of ANOVA can be extended in the same way as linear regression models can be extended to generalized linear models etc. Hence, we can do ANOVA also for models with non-normally distributed errors (of course you have to tell this the software, it won't do it automatically). You therefore have to read carefully what people mean if they use the term.

Here a simple example with a standard ANOVA (normal errors), testing whether weight (of chicken) depends on their diet, where diet is a factor variable with four levels:

```
> aovresult <- aov(weight~Diet, ChickWeight)
> summary(aovresult)

             Df  Sum Sq Mean Sq F value   Pr(>F)
Diet          3  155863   51954   10.81 6.43e-07 ***
Residuals   574 2758693    4806
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>
```

We find a p-value of 6.43e-07, which is highly significant at an α level of 0.05. Hence, we can reject the null hypothesis that the diet has no influence on the response "weight". Note that in this case, we don't get any parameter estimates, and we can't say anything about which of the diets differs from which. If you want those, there are two options:

- Either you apply what is called post-hoc testing, which means that you test for differences (e.g. with a t-test) between the diets.
- Or you switch to a regression, which is described in the next chapter

If you do post-hoc testing, you are doing multiple tests on the same data. This is a problem - the idea of the p-value is that you calculate the probability of seeing the data under ONE null hypothesis. If you do this, you will get at most 5% error at an α level of 0.05. However, if we do multiple tests, we are testing multiple null hypotheses, and there are more options for the test statistics to get significant just by chance. Hence, we need to correct the p-values for multiple testing. There are a number of options to do so, google is your friend.

When doing multiple tests on the same data, we need to correct the p-values for multiple testing.

4.4 Other important tests

t-test and ANOVA are the commonly needed tests in the context of the research skills module, but there are many more tests that could be potentially important. A list of tests that have wikipedia articles can be found at here

4.5 Regression

As explained earlier, regression does not necessarily mean using a different statistical model as in hypothesis testing (ANOVA and the linear regression model in R use the same assumptions). However, the goal of regression is a different one. While hypothesis tests are all about seeing whether the data would be compatible with a null-hypothesis, regression is about finding the best-fitting hypothesis or parameters (the MLE). Hence, a regression model tries to find the parameter combination that produces the highest probability to create the observed data, given the model assumptions.

Linear regression

The most basic regression model is the linear regression. The assumption here is that we have a response that depends on the predictors in a form of

$$y \sim a \cdot x + b + \epsilon \tag{4.3}$$

where y is the response, x is a predictor, a is the parameter that fits how strongly the predictor influences the response, b is the intercept, and ϵ is the random variation, which in a linear regression is assumed to be normally distributed.

In R, we can do such a regression by typing

```
> fit = lm(airquality$Temp~airquality$Ozone)
> summary(fit)

Call:
lm(formula = airquality$Temp ~ airquality$Ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-22.147  -4.858   1.828   4.342  12.328

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      69.41072    1.02971   67.41   <2e-16 ***
airquality$Ozone  0.20081    0.01928   10.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.819 on 114 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.4877,       Adjusted R-squared:  0.4832
F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

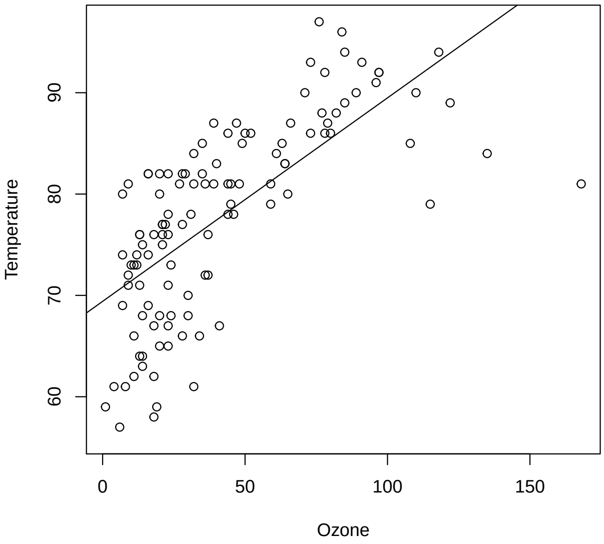


Figure 4.3: Airquality data: Temperature plotted against Ozone. Relationship indicated by the regression line.

You can use the same code regardless of whether your predictor is continuous or categorical. In case of continuous variable, a line is fit to the data. In case of a categorical variable with n levels, the first level is set as the reference (intercept), and n-1 parameters are fitted for the following levels that describe the difference to the reference.

The fitted parameters appear in the column ”Estimate”. This tells us how much the predictor, in this case Ozone, affects the response, in this case the Temperature: for each unit of Ozone more, temperature increases by a 0.208 units, with a standard error (confidence interval) of 0.019. Apart from seeing how a regression output looks like, this teaches us another valuable lesson: the fact that we have used temperature as a response here and ozone as a predictor doesn't mean that ozone causally affects temperature. In fact, it is likely the other way around: if we have more sun, it's hotter, and we tend to have more ozone as well. Regression, as most other statistical analysis, doesn't establish causality, it establishes correlation. What we are saying is that if our ozone measurements go up, we can be pretty sure that it is hotter as well. Doesn't mean that ozone creates heat. Correlation is not causality.

Correlation is not causality.

The regression results gives us a lot of p-values as well. These are results of various hypothesis tests that are performed automatically for you after the regression is done. For example, we get a p-value for each parameter. This p-value is based on a particular type of t-tests where the full model is tested against the model with the parameter set to 0. There is also other p-value, based on a different test statistics at the end of the regression output. This tests the null hypothesis that all parameters are zero.

Specifying different model assumptions in R:

Response y depends linearly on a variable a (continous or categorical)

```
> fit = lm(y~a)
> summary(fit)
```

Response y depends linearly on two variables a and b (continous or categorical), but the value of either variable doesn't influence the effect the other variable has on the response (no interaction)

```
> fit = lm(y~a+b)
> summary(fit)
```

Response y depends linearly on two variables a and b (continous or categorical), but the value of one variable does influence the effect the other variable on the response (interaction)

```
> fit = lm(y~a+b)
> summary(fit)
```

Response y depends as in $a + a^2$ on a variable a (continous or categorical)

```
> fit = lm(y~a + I(a^2))
> summary(fit)
```

the I() notation means that the following expression is interpreted as a mathematical formula.

Checking the assumptions

Formally, we can fit any data with a linear model. However, as in any statistical inference procedure the results (i.e. parameter estimates, p-values) are conditional on the assumptions that we have made. Hence, the p-value we get is conditional on the assumption that the data is actually from a process that conforms to eq. 4.4. If it doesn't the p-value could be completely wrong. Hence, we have to check whether those assumptions are actually met.

So, what were the assumptions of a linear regression? One problem I often encounter is that students remember that the assumptions were normal distribution. Hence, they look at whether the response variable is normally distributed. However, if you look sharply at eq. 4.4, you see that this was actually not the point. If we shift around the terms in eq. 4.4, we see that what is actually supposed to be normally distributed is

$$y - (a \cdot x + b) \sim \epsilon \tag{4.4}$$

i.e. the difference between the observed value and the model predictions. These differences are called the residuals, and, according to the assumptions of our model, they should be normally distributed. You get basic residual diagnostics by typing plot(fit), where fit is your fitted model object. For further details on residual diagnostics, see here.

4.6 Generalized linear regression models

The general ideas of a linear regression was that 1) The response is continuous, theoretically from -infinity to + infinity, and 2) residuals are normally distributed around the model predictions. The idea of the GLM framework is take the linear regression framework is to allow you to work as before in the linear regression example, but relaxing both the assumptions about response values from - to + infinity, and the normality. To do this, we have to do two things

- To get the output values on the range that we want, we wrap the linear model in a transformation function that forces the response on the right interval (typical intervals are positive, or between 0 and 1). This transformation is called the link function
- To fit other distributions, we have to tell the model to use something else than the Gaussian error function.

Let's talk about these points in a bit more detail.

The link function

We said above that a linear regression takes the form

$$y \sim a \cdot x + b \tag{4.5}$$

That means that if x gets large, y could take any value, positive or negative. A trick to ensure that all predictions for y are positive, or within a certain range is using a link function of the form

$$y \sim f^{link}(a \cdot x + b) \tag{4.6}$$

Any function is possible, but as we see later typical choices are the exponential function, which ensures positives outcomes, and the inverse logit, which ensures are range between 0 and 1.

Other distributions

Well, this is conceptually the easy part, but maybe you are not yet aware what kind of distributions exist beside the normal. Two typical choices that we use below are the binomial (the distribution for coin flipping), and the Poisson distribution (a discrete probability distribution). There are many other choices available. Maybe it becomes more clear when we move to the actual examples in the next sections.

0/1 data - logistic regression

Logistic regression is the most common analysis for binary data (presence/absence; survived/dead; infected/not infected). Logistic regression assumes that the distribution is binomial (coin flip model). To get the linear predictor on a scale between 0 and 1 that is necessary for the binomial distribution, we use the logistic link function (or inverse logit).

evspace1cm

In R:

Here an example with the data of the Titanic survivors. Note that if you tell R to use the binomial distribution, the logit link is automatically selected. If you wanted, you could overrule this choice.

```
> library(effects)
> fmt <- glm(survived ~ age + I(age^2) + I(age^3), family=binomial, data = TitanicSurvival)
> summary(fmt)
```

```
Call:
glm(formula = survived ~ age + I(age^2) + I(age^3), family = binomial,
    data = TitanicSurvival)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5062  -0.9978  -0.9695   1.3483   2.0135

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.855e-01  3.031e-01   2.592 0.009549 **
age         -1.189e-01  3.291e-02  -3.613 0.000303 ***
I(age^2)     3.414e-03  1.113e-03   3.066 0.002171 **
I(age^3)    -2.931e-05  1.107e-05  -2.648 0.008109 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.6  on 1045  degrees of freedom
Residual deviance: 1398.5  on 1042  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 1406.5

Number of Fisher Scoring iterations: 4
```

count data - poisson regression

Poisson regression is the standard choice for working with count data, although there are a few other options available as well. In poisson regression, the standard choice is to use a exponential function to make all values positive. The inverse of the exponential is the log, so we call this the log link. As before, R is choosing this automatically if you specify the distribution to be poisson.

In R:

An example, using some data on the feeding of bird nestlings, in relation to their attractiveness:

```
> schnaepper <- read.csv("schnaepper.txt", sep="")
> fm <- glm(stuecke ˜ attrakt, family=poisson, data = schnaepper)
> summary(fm)
```

```
Call:
glm(formula = stuecke ˜ attrakt, family = poisson, data = schnaepper)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.55377  -0.72834   0.03699   0.59093   1.54584

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.47459    0.19443   7.584 3.34e-14 ***
attrakt      0.14794    0.05437   2.721  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25.829  on 24  degrees of freedom
Residual deviance: 18.320  on 23  degrees of freedom
AIC: 115.42

Number of Fisher Scoring iterations: 4
```

Residual checks in GLMs

Residuals in glms are not supposed to be normally distributed, so don't use standard checks for normality such as normal qq plots to check for the appropriateness of the residuals. For not too complicate models, a good way to deal with this problem is to use the so-called pearsons residuals, which scale the observed differences between model and data by the expected variance of the model [1]

One standard concern in poisson or binomial glms is that the variance of the poisson and binomial distribution cannot be adjusted, but is fixed by the mean. This is a problem you don't encounter in the normal linear model, because here the random part is modelled by a normal distribution, which has a parameter for the variance. A problem that appears very frequently in Poisson or binomial glms is overdispersion, i.e. that the residuals show more variance than expected under the fitted model. The easiest way to correct for this is to use the quasi-Poisson and quasi-binomial models available in the glm function. These models fit an additional parameter that modifies the variance of the Poisson and binomial glm.

You can check for overdispersion by looking at the fitted deviance, or apply an overdispersion test

[1] in R, you can specify the option pearson in many fuctions, including the residual() function that you can apply to a fitted object

5 Predictive statistics - machine learning

A third class of statistical procedures that have become very important in recent years are predictive methods, often called machine-learning algorithms. The basic goal of this methods is to to be able to make predictions from a given dataset with the lowest possible error. In doing so, they typically use relatively complicated, often non-parametric methods that typically don't allow to calculate inferential products such as the MLE or p-values.

There is considerable tension between the more classical field of inferential statistics and the more modern field of machine-learning. For classical, inferential statisticians, machine learning methods have abandoned the idea of "learning from data", in the sense of comparing hypotheses to data, in favor of simply making predictions. A statistician concentrating on machine-learning would reply that in many applied problems, there is nothing to learn [1]. The goal is to build an algorithm that is able to correctly predict given a complex dataset. The distinction between the goals of inferential statistics and predictive statistics, as well as the tension between these fields, are nicely summarized in the abstract of the extrememly recommendable article "Statistical Modeling: The Two Cultures" by Breiman (2001):

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated bya given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move awayfrom exclusive dependence on data models and adopt a more diverse set of tools.

I have included this short chapter because of the importance of predictive methods in modern statistics. A detailed explanation of the methods of machine-learning, however, is beyond this introduction. If you are interested in starting to learn more about predictive

[1] Typical machine learning applications include predicting the interests of customers in web shops, the association of feature-rich satelite data with ground signals, or speech / face recognition. Machine-learning experts are currently sought-after by technology companies such as google, facebook and so on.

methods, I would recommend to start with the textbook by James et al. (2013) that I also recommend at the end of this primer for further reading.

6
Design of experiments

Let's come back to one of the first point in this script: the data. If we have to collect data ourselves, we have to answer a number of questions. Which variables should we collect? At which values of those variables should we collect data? And how many replicates do we need?

6.1 Selection of variables

In a practical setting, we are typically interested in how a response is affected by a number of predictor variables. Clearly, we need to measure both response and this predictors of interest across a few of those predictor values to say something about the effect of the predictors. If we only wanted to know whether there is a correlation between predictors and response, our list of variables would be complete at this point. However, typically, we want to know not only if there is a correlation, but also whether we can say with some confidence that this correlation is causal. If we want to make this claim, we have to exclude that there are counfounding factors, also called confounding variables.

Correlation is not causality. For suggesting causality, we have to exclude confounding effects.

What is a confounding variable?

Imagine we are interested in a response A, and we have hypothesized that A B. Imagine there is a second predictor variable C that has an influence on A, but in which we are not interested in for the purpose of the question under consideration. Such a variable that is not of interest for the question is also called "extraneous variables". . So we also have A C, but we are not interested in this relationship. If we now take data, and don't measure C, it's usually not a bit problem as long as C is uncorrelated with B - it might create a bit more variability in the response, but by and large the effect of C should average out and we should still be able to detect the effect of B.

An extraneous variable is a variable that can influence the response, but is not of interest for the experimenter

The problem of confounding appears when the extraneous variable C is for some reason correlated with the predictor variable of interest B. In that case, if we only measure B, we see both the effect of B and C. In this case, we may attribute the effect of C on A wrongly to the effect of B on A. Auch a correlation that is caused by an

A confounding variable is an extraneous variable that correlated to both the response and a predictor variable of interest.
A spurious correlation is a correlation that is caused by a confounding variable.

unmeasured confounding variable is called a spurious correlation.

What do do about confounding variables

If we think there is a factor that could be confounding, we basically have three options

1. Best: control the value of these factors. Either fix the value (preferred if we are not interested in this factor), else vary the value in a controlled way (see below).

2. Second best: randomize and measure them

3. Third best: only randomize or only measure them

Randomization means that we try to ensure that the confounding factor is not systematically correlated with the variable of interest (but can still cause problems with interactions and nonlinear relationships).

Measuring allows us to account for the effect in a statistical analysis, but cost power (see below) and, and we can't measure everything.

Variables that we include but that are not interesting to us are often called nuisance variables.

6.2 Definition and bias of variables

A common mistake at this step of the design is to take the variable definition and measurements for granted, and continue with considering the selection of replicates and so on. The step that is missing is to think about the following two questions.

The consideration of these two questions is often referred to as construct validity, see also main RS script.

1. Do my variables measure what I want to measure

2. What is the expected statistical (stochastic) error in my measurements, and what is the possible systematic error in my measurements

The first item may seem a bit odd, because one would think that we know what we measure. However, in many cases in ecological statistics and beyond, we do not measure directly the variable that we are interested in, but rather a proxy. So, for example, we want temperature on the plot, and we use temperature from a weather station 5 km away. Or, we want to look at functional diversity, but how can we exactly express this in terms of variables that we measure in the field.

The second questions relates to considering how much two measurements would differ if we do them repeatedly (stochastic), and how much measurements could be off systematically (e.g. because a method or instrument is systematically wrong, or because humans show particular biases).

See main RS skript for more info on biases.

6.3 Selection of values for the independent (predictor) variables

If we have decided which variables to measure, we have to decide for the values at which we want to measure them. In an observational study, this may not always be possible completely, but it's usually possible to ensure sufficient variation in the predictor variables. A few points to consider

Vary all variables independently

A common problem in practice is that we have two variables, but their values change in a correlated way. Imagine we test for the presence of a species, but we have only warm dry and cold wet sites. We say the two variables a collinear. In this case we don't know whether any observed effect is due to temperature or water availability. The bottom-line: if you want to separate two effects, you need them to vary independently.

Interactions

To be able to detect interactions between variables, it's not enough to vary all, you also need to have certain combinations. The buzzword here is called factorial design. Google will help you.

Nonlinear effects

The connection of two points is a line. If you want to see whether the response to a variable is nonlinear, you therefore need more than two values of each variable.

6.4 How many replicates?

We said before that the significance level α is the probability of finding false positives. This is called the type I error. There is another error we can make: failing to find significance for a true effect. This is called the type II error, and the probability of finding an effect is called power. . For standard statistical methods, power can be calculated. You have to look it up for your particular method, but in general assume that

Power is the probability of finding significance for an effect if it's there

1. Power goes up with increasing effect size
2. Power goes down with increasing variability in the response

This means that, unlike for the type I error which is fixed, calculation of power requires knowledge about the expected effect and the variability. This sounds really bad, but in most cases you can estimate from previous experience how much variation there will be, and in most cases you also know how big the effect has to be at least to be interesting. Based on that, you can then calculate how many samples you need.

Checklist experimental desig

( ) Clear, logically consistent question? Write it down.

( ) What variables need to be measured to answer this question? Check that the measured variables really correspond to the main question

( ) Confounding factors controlled, randomised or measured? Are you sure they are confounding (correlated to response AND one or several of the predictors)

( ) Define the exact hypothesis (statistical model) to be tested. Write it down, as in $height \sim age + soil * precipitation + precipitation^2$. Based on that, decide how many different levels of each variable need to be measured.

( ) Decide on the number of replicates. Make a guess for effect size and variability of the data, and either calculate or guess the number of replicates necessary to get sufficient power. What sufficient means depends on the field, but I would say you want to have a good chance to see an effect if it's there, so a power of $>$ 80% would be good.

7
Good to knows and further reading

7.1 Reproducibility and good scientific practice

Reproducibility means that each step of your analysis is repeatable. Experience shows that it is not as trivial as it sounds to ensure reproducibility. Here some hints for making your data analysis reproducible

- Once you have your raw data produced, NEVER change it. Store it in a save location, make a backup, and never touch it again
- Typically you will have to do some cleaning, renaming etc. before the data analysis. If possible at all, make this through a script (e.g. R, python, perl). Store the script with the analysis.
- Use a version control system for your code, and note for each output the revision number that the output was produced with.
- When running the analysis, store the random seed and the settings of your computer to ensure reproducibility. In R, the easiest way to do this is to set the random seed by random.seed(123), and store the results of sessionInfo() which provides you with the version numbers of all the packages that you use
- Think about running your code within an reporting environment such as Rmd or sweave

7.2 How to learn more about statistics

- To complete this primer, I would recommend that you go through the practicals for this script.
- If you want one further practical textbook for beginners, I recommend Dormann (2013) for German speakers (ebook free of charge for students from Freiburg, contact me) and Gotelli and Ellison (2004) for English speakers.
- For the technically slightly more ambitious (it's still very elementary), I recommend James et al. (2013). You can download the pdf for free, and there is a MOOC available for the book with lectures and exercises.

- For more help and references, see the stats help website of our department here, in particular the recommendations regarding R scripts and statistics textbooks.

8 Bibliography

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician 27*(1), 17–21.

Breiman, L. (2001, 08). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci. 16*(3), 199–231.

Cohen, J. (1994). The earth is round (p<. 05). *Am. Psychol. 49*(12), 997.

Dormann, C. F. (2013). *Parametrische Statistik.* Springer.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian data analysis* (2nd ed.). Chapman & Hall, London.

Gotelli, N. J. and A. M. Ellison (2004). *A Primer Of Ecological Statistics.* Sinauer Associates.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

Student (1908). The probable error of a mean. *Biometrika*, 1–25.