Workshop: Basic Analysis of Survey Data

Martin Mölder

November 23, 2017

Contents

1 Introduction and general remarks

The goal of the current workshop is to give a basic introduction to the main concepts and basic statistical procedures that are underlying most quantitative analyses, including and especially survey data. The workshop assumes no prior knowledge of the subject matter and will start from the very beginning of the whole topic. The content of the workshop has the same range as a typical introductory qunantitative methods class in the social sciences would have (at the MA level), but what would normally be covered in a whole semester is here squeezed into about 6 hours.

The focus is on substance and concepts, but we will also use hypothetical data as well as examples from some of the main international surveys in social and political science to familiarise ourselves with the basic statistical procedures.

1.1 Further reference

Most of the topics that are covers in these materials are included in any introductory statistics textbook. Almost all of the topics are also covered in the following book:

- Gravetter and Wallnau, "Essentials of Statistics for the Behavioral Sciences". There are various edition of the book available, all are equally good.

1.2 Statistical software

The range of software to handle numerical data is vast and different sectors and disciplines have different preferences in this regard. For examples, Excel is for some reason more often used in business and the public sector, in science software like SPSS, STATA and R are preferred. Among other things, these differences come from the purposes of the analyses that are conducted as well as the nature of the data. Each have their benefits and weaknesses. However, for scientific analysis, R is preferred, as it is both powerful and open source (free to use).

Benefits and problems of Excel:

- Easy to use basic functionality.
- Ability to see and directly work with your data.
- Less efficient for more complicated tasks (analysis and visualisation)
- Limitations to automate and integrate

Benefits and problems of R:

- Potentially limitless capabilities
- Easy to integrate
- Easy to automate
- Scripts are diary of your analysis
- Very hard to learn at first

For many purposes, SPSS is very user friendly and sufficient. However, it is also very expensive. Thus, if you do not have access to SPSS or other proprietary statistical software and you do not have the time or the need to familiarise yourself with R, then I recommend giving PSPP a chance. It is open source (free of charge), compatible with SPSS files and allows to the very basic data visualisations and analyses, as well as basic data management that are tedious and unnecessarily complicated with Excel.

1.3 Getting Excel ready

As this is software that everyone is familiar with, we will be using Excel for some of the examples in this workshop. Excel is good for manually managing your data and it does have some statistical capabilities by default, but for any kinds of statistical analysis with Excel I would recommend expanding its functionality with some of the available add-ons.

For the purposes of this workshop, let's download and install XLSTAT. It is not free of charge, but it does have a free trial period, so we can use this here. If you are going to be doing more data analysis, I would, however, recommend ultimately adopting R.

Why this is not a class in R?

1.4 Overview of main concepts

- **Census** & **survey**.
- **Population** & **sample.**

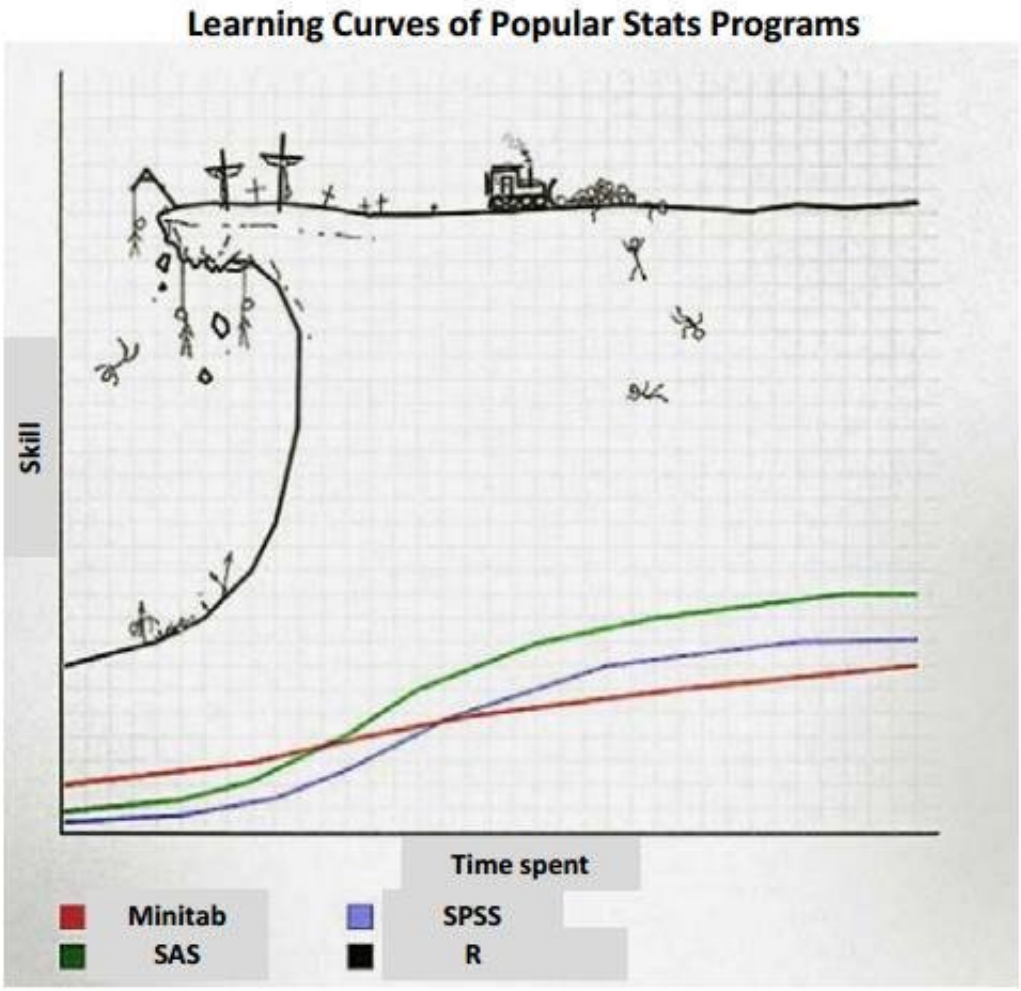


Figure 1: *The learning curve of R.* Source: https://pbs.twimg.com/media/CiSMsEpXEAATXry.jpg

- **Random sampling.**
- **Response rate.**
- **Variable.**
- **Parameter.**
- **Statistic.**
- **Census value** & **survey value.**
- **Descriptive statistics.**
- **Inferential statistics.**
- **Sampling error.**
- **Reliability.**
- **Margin of error.**
- **Non-sampling error.**
- **Validity.**
- **Self-selection bias.**
- **Observational and experimental studies.**
- **Independent variable.**
- **Dependent variable.**
- **Conceptualisation.**
- **Operationalisation.**
- **Operational or effective definition.**

1.5 Goals of surveys and the problem of uncertainty

Many of the concepts above emphasise the gap between what is out there in reality and what we can and want to know about it. This is also the central question of survey analysis. Even if we assume that everything else is perfect (questionnaire design, collecting the sample, the interview situation, etc.), we have to take into account the fact that we have information from only a part of the whole range of people we are interested in and therefore there is always a certain irreducible uncertainty that we have to account for, that we have to determine. This is the fundamental problem on any analysis involving survey data. Fortunately, given certain assumptions, it is rather straightforward to put a number on this uncertainty. Much of what we are looking at below, will involve focusing on this number.

1.6 About data

1.6.1 Example data

Some of the example variables below will come from:

- Comparative Study of Electoral Systems (CSES). The CSES is a major international survey about elections and stands out for its excellent data quality and preparation. Except for a few minor details, it is exemplary in terms of how data should be cleaned, stored and made available.
- European Values Study. This is another major international survey with a more sociological focus.

1.6.2 File types

Survey data rarely comes in the .xls or .xlsx formats that are native to Excel and often they are either in the form of a specific type of a text file or in the format of some statistical analysis software. Here are some of the most common file types:

- **.csv** Comma separated values. A text file, where each row is a row in the data table and columns are separated by commas. They can also be separated by tabs or some other character. Normally, Excel opens them automatically. If it does not then under the "Data" tab, one can import them with the

option “From Text/CSV”. Because it is a text file, you can open this with any text editor and have a look at how the data in the file looks like.
- **.rda or .rdata** R data files. Can be processed only with R. Excellent data compression.
- **.sas** File type of SAS software.
- **.sav** File type of SPSS software.
- **.dta** File type of STATA software.

Excel is only able to comfortably handle its own native format and CSV files. The proprietary software that I have mentioned tries to put restrictions on where else you can use their file types. R, however, through its extensions is able to handle almost all file types.

CSV files are the most simple and compatible across all systems and platforms (because they are essentially text files), but they take a lot of space (for the very same reason) and they are unable to store nothing, but the numerical or textual values of cells.

1.6.3 A clean data file and a codebook

Although here opinions might differ, one could argue that it is best to keep your data file as “clean” as possible and that you store everything else about your data in a separate file that is your codebook. What does this mean?

- Data is only stored as numbers (except for explicitly textual data).
- Each row is one case (individual) and each column is one variable.
- Column names are short, but informative (user is able to deduce the content of the variable from the name).
- Codebook contains for each column name the full wording of the question and a definition for each of the categories that are represented by numbers in the data file.
- All missing data is coded as NA (unless you need to differentiate between various kinds of missing data; what are they?).

1.6.4 Missing data

Missing data poses a problem for the survey as a whole as well as for data analysis. What is the problem and what can we do about it?

- Too much missing data → biased sample. Why? Missing at random or missing not at random?
- Cases excluded for any analysis → less certainty even if no bias.
- Imputation. Filling in the blanks as intelligently as possible. How?

1.6.5 Types of variables

It is always important to keep in mind what the nature of the measurement scale that you are working with is. This determined what it makes sense to do with the data and what it does not. Choices are made more complicated by the fact that the lines between the scales are often blurred and it might make sense to treat one type of data as if it was of another type.

What are the different scales of measurement?

What kinds of plots are useful for visualising different types of variables?

The scale of measurement is also something that one should think about before any survey is conducted. We have to consider what the nature of the underlying phenomenon or characteristic is that we want to study and we must choose out survey questions accordingly. The main dilemma here is between ease of analysis and the nature of social reality.

What is this dilemma?

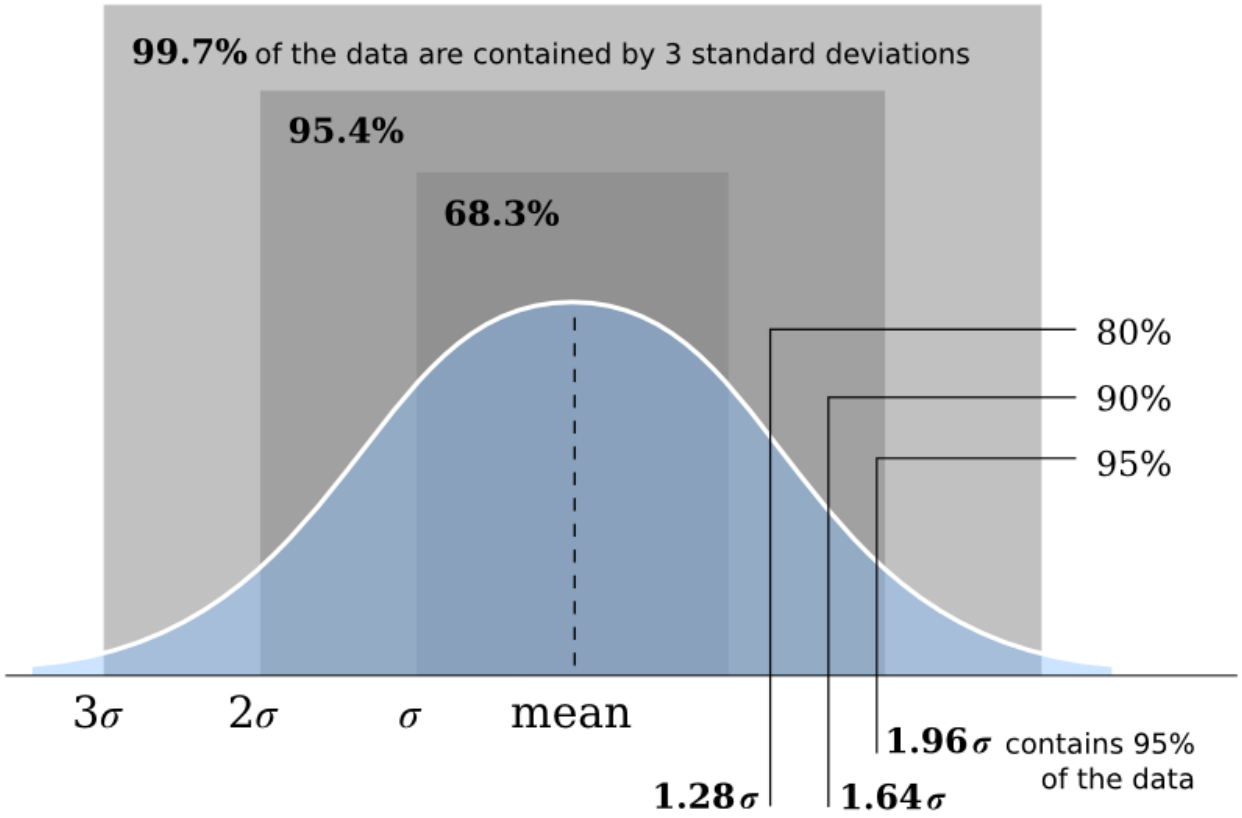


Figure 2: The normal distribution. Source: http://subsurfwiki.org/wiki/Normal_distribution

What is the nature of social reality? Is it continuous or categorical?

It is also important to keep in mind that it is possible to only move "down" different measurement scales, from the more informative to the less. You can re-code your variables only "down", and not "up".

2 Getting to know your data

2.1 The normal distribution

At the very root of all survey analysis is the normal distribution (or something very similar to it). It is the underlying assumption of much of what follows, thus it makes sense to have a look at it at the very outset.

How to evaluate the "normality" of a distribution?

- **Skewness.** One of the two tails is stretched out.
- **Kurtosis.** The distribution of more flat or more peaked.

2.2 Central tendency

What is central tendency?

Any analysis, of single variables or variables together, is about summarising patters in a data, representing with few numbers what is in the data set perhaps recorded through thousands. Central tendency is the one value that best describes the distribution of a single variable. What measure of central tendency is more appropriate depends on the measurement scale of the variable and the nature of its distribution.

2.2.1 The mean

The mean is the balance point of the distribution. The total distance below the mean and above the mean is the same.

- Population mean: $\mu = \frac{\sum X}{N}$
- Sample mean: $\bar{X} = \frac{\sum X}{n}$ or $M = \frac{\sum X}{n}$ (Notations can differ wildly across various sources, thus it is important to keep in mind what refers to what in which context.)

Why should we differentiate between the two?

Because the mean is based on distances, it should be used only for continuous variables. It is impossible to measure distance if you scale is not interval or ratio.

2.2.2 The median

The median divides the distribution into two. The number of cases above and below the median is the same, 50%. When the mean balances distance, then the median balances the number of cases.

The median can be applied both to continuous and categorical variables.

When is it more useful to focus on the mean and when on the median if both are applicable?

2.2.3 The mode

The mode is the value or the category that has the greatest frequency in a distribution.

The mode can be applied both to continuous and categorical variables. However, it is the only measure that can be applied to nominal scales, because the median, which also applies to some categorical variables, assumes direction.

2.3 Variability

What is variability? Why is it not enough?

Central tendency alone is a very poor characterisation of your data, because it disregards the most important thing - uncertainty. The latter is represented in how much the values of a variable vary around their central tendency. It is important to know the variability of your data (and where it comes from, both substantively and mathematically) for its own sake, but also for the fact that it is be basis of estimating how well e.g. the mean in the sample represents the mean in the population.

Variability is the extent to which the values of a variable as spread out or concentrated. If we are thinking in the context of surveys and samples, then variability determines how certain we can be that the values or associations we see in the sample are also there in the population. Our survey can be useless, if variability is high, even if measures of central tendency of some association point to the direction that we like.

How to characterise variability? This also depends on the measurement scale.

For nominal data and the mode:

- It's simplest just to have a look at the **frequency distribution** of your variable. There is no simple measure of variability. We can focus on the modal category and the associated confidence interval. Will come back to this later.

For ordinal (median) and continuous data (mean):

- **Range.** The difference between the largest and the smallest value.
- **Inter-quartile range.** The distance between the third and the first quartile. The region where the middle 50% of the data is located.

If you have continuous data, however, it makes most sense to work with **variances** and **standard deviations**, because they are a fundamental part of the puzzle of assessing the uncertainty related to our statistics.

2.3.1 Variance and standard deviation

Deviation: distance from the mean $= X - \mu$

Variance: is the mean squared deviation, the average squared distance from the mean.

$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$

The sum of the squared distances from the mean is called simply the **sum of squares**.

$SS = \sum(X - \mu)^2$

About the summation notation.

Standard deviation is the square root of the average squared distance from the mean, the square root of the variance.

$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$

NB! Notice the big N and the μ, we are talking about the population variance and the population standard deviation. If we are talking about the sample, then that is usually referred to with a small n and the mean of X in the sample can be referred to as $\bar{X}$

Now we can come back to the normal distribution, it will make much more sense.

Standard deviation is the **typical distance from the mean of the normal distribution**. About 70% of the values are located within one standard deviation from the mean and about 95% of the values are located within two standard deviations. Thus if you see the mean and the standard deviation, you will immediately know two of the essential features of the data - the central tendency and the variability. This is as simple as you can get, you cannot merge these two numbers into one.

2.3.2 Z-scores and the standard normal distribution

What is a z-score?

A z-score is the value of a variable that has been transformed into how many standard deviations above or below the mean it is. For each score, you calculate the distance from the mean and divide it by the standard deviation of the distribution.

$z = \frac{X-\bar{X}}{\sigma}$

A standard normal distribution is the distribution of such a variable. It has a mean of 0 and a standard deviation of 1. Why can this be useful?

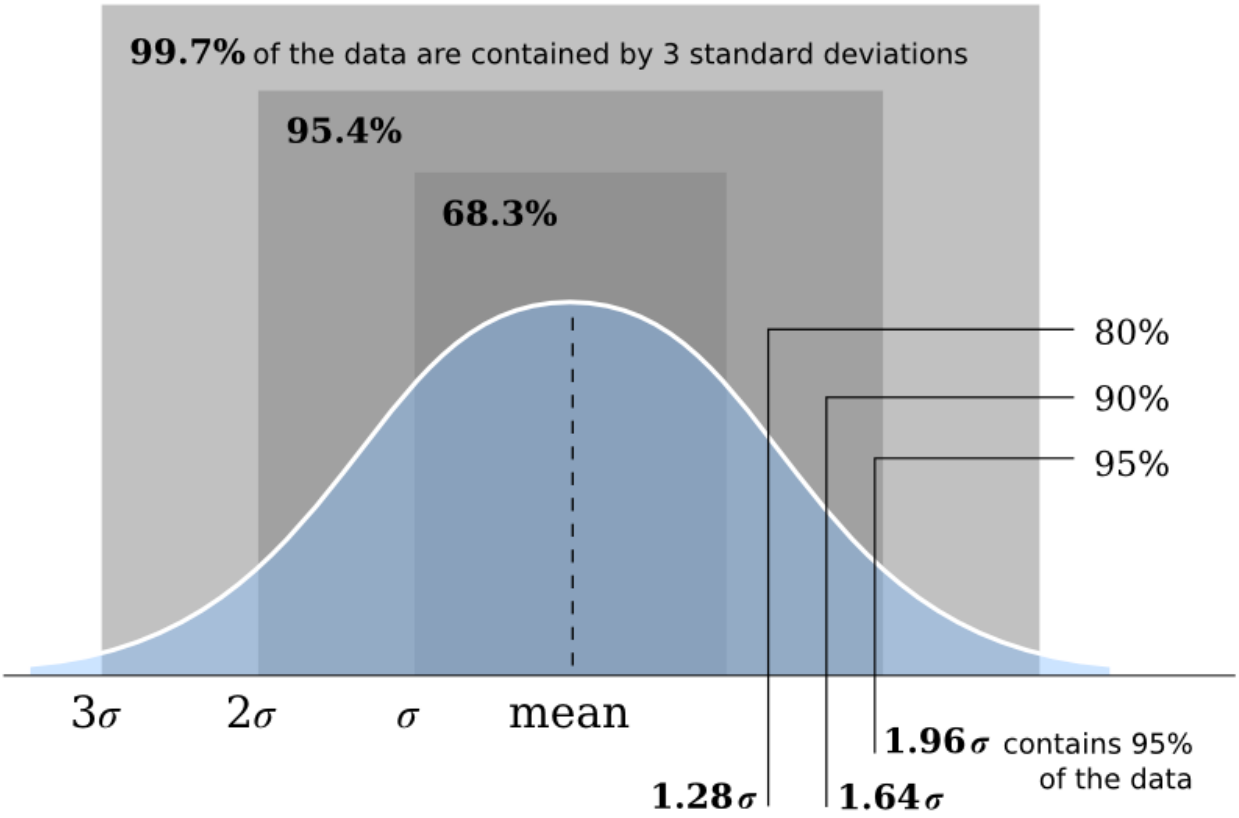
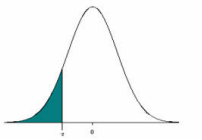


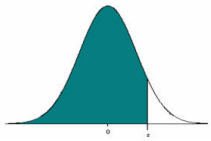
Figure 3: *The normal distribution.* Source: http://subsurfwiki.org/wiki/Normal_distribution

Table of Standard Normal Probabilities for Negative Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

Figure 4: *Z-scores.* Source: http://sixsigmastudyguide.com/z-scores-z-table-z-transformations/

2.4 From the sample mean to the population mean

Sampling error - the amount of discrepancy between the sample statistic and the population parameter.

Let's imagine a hypothetical situation where we draw a sample of a certain size many times in a row and each time we calculate the mean of a variable in the sample. The means are different, because each random sample is different. But they will follow a certain mathematically proven logic. This also means that we do not have to draw an infinite amount of samples to evaluate how far our sample mean can be from the population mean. We can use the information that we have for and from that single sample. We just have to keep in mind, that we are talking about **random samples** from a large population. The is the very basic assumption. **Coming back to a topic that we touched upon above, how often is that actually the case?**

In short, the distribution of sample means is the distribution of the means that are calculated for a sample of size n drawn an infinite amount of time. The means are different, because each random sample is different. A **sampling distribution** is such a distribution of sample statistics. It also has a mean and a standard deviation.

Thus, if we are talking about a sample statistic like a sample mean and we want to make an inference about the population, which we ultimately should want to do (because what is the point of surveys otherwise), then we should be thinking about this sampling distribution. Its variance and its standard deviation help us to understand how far the population mean can be from our sample mean. But how do we calculate this?

Central limit theorem. What is it?

For a population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$ and it will approach a normal distribution as n approaches infinity.

n does not have to be large, the central limit theorem works already rather well, if sample size is 30 or more and if the population that we are working with is normally distributed.

Let us look at the formula again: $SE = \frac{\sigma}{\sqrt{n}}$

The standard error is a function of sample size.

> ***Example***
> **Sheet:** SE and sample size
> The standard error depends on the sample size, but not linearly. Let's see how that looks like.

2.4.1 Confidence interval

Now we can finally answer the question - if we have a sample mean $\bar{X}$, then how far from this sample mean and with what certainty is the population mean? How can we answer that?

- The standard error is the standard deviation of the sampling distribution of the mean.
- The sampling distribution is a normal distribution.
- We know what proportion of the values of a normal distribution are within a certain number of standard deviations from the mean.
- We just have to choose the right number (of standard deviations), multiply the standard error with that number and the **confidence interval (CI)** is that much to the either side of the mean.
- The level of confidence of the CI refers to the amount of the normal distribution that is contained in that interval.
- **There is one more nuance that we will get to in a bit.** Hint: here we are still assuming that we know the population standard deviation.

2.5 Variance and standard error for a proportion

With surveys perhaps the even more common quantity that we need to evaluate is a proportion, the proportion of respondents that gave a particular (categorical, binary) answer to a question. There is a standard error also associated with that proportion. We will not go into the background "mechanics" of this like we did in the case of the mean and the normal distribution, but it is good to remember that the overall logic is the same. We are simply dealing with a different distribution, which has a different formula for variance and thus also for standard deviation.

The sample variance of a proportion is: $p(1-p)$

Standard error of sample proportion (standard deviation of the sampling distribution) is thus:

$SE = \sqrt{\frac{p(1-p)}{n}}$

where p is the sample proportion and n is the sample size.

Our standard 95% confidence intervals are thus:

$CI_{95\%} = \pm 1.96 \times SE$

> ***Example***
> **Sheet:** SE proportion
> Let's have a look at how to do these calculations for a range of proportions and sample sizes. Since you are quite likely very often working with proportions like this, it is useful to have a longer look.

2.6 Looking at the distribution of variables

The mean and the standard deviation are just two numbers, but very often it is useful to plot the whole distribution of your variable. This will give an intuitive sense of the nature of the data that you are working with. Let's have a quick look at how the XLSTAT extension allows us to make some very basic visualisations.

2.6.1 Continuous variables

To visualise the distribution of a continuous variable, use a histogram.

> ***Example***
> **Sheet:** EVS1
> XLSTAT -> Visualising data -> Histograms

2.6.2 Categorical data

To visualise the distribution of a categorical variable, use a bar chart.

> ***Example***
> **Sheet:** EVS1
> XLSTAT -> Visualising data -> Univariate plots

How to add error bars to the chart?

> ***Example***
> **Sheet:** EVS1
> XLSTAT -> Visualising data -> Error bars
> Use the output from the previous operation for the data.

Tip: If you are making graphs, the height should be between 2/3 and 3/4 of the width, at least in most cases. This provides the best visual outcome.

How should a good visualisation look like?

3 Statistical inference and hypothesis testing

A statistical test or a hypothesis test means asking and answering questions about the population on the basis of the data in the sample, taking into account the uncertainty that comes from sampling error. We looked at elements of this already, but now we can put the whole picture together.

Alpha level or **the level of significance**. Determines the width of the confidence interval for our statistic, that we deem acceptable. But here we refer to the region that is outside the confidence interval. The simplest interpretation: If we take a **random sample** of size n an infinite amount of times, what is the probability that the true value in the population is outside the confidence interval determined by the alpha level.

The basic questions that we ask with the hypothesis test: - Is one value in the population different from another value? - Is the value in the population different from 0?

We already basically got there, when we were looking at the standard error of the mean and of the proportion and the corresponding confidence intervals. A 95% confidence interval corresponds to 0.05 level of significance.

Hypothesis and **null hypothesis** - what are they and what is the difference?

3.1 Statistical significance and substantive significance

When something is statistically significant then that means that our hypothesis test according to our level of significance gave us a desireable answer. But only in the form of yes/no. We are able to say that the population parameter is different from 0.

But remember: All else being equal, the standard error depends on the sample size, more specifically the square root of the sample size. Thus, statistical significance is purely an artifact of sample size. With enough data, everything is statistically significant and if we simply focus on whether something is statistically significant, then we might be emphasising a difference or an association that is trivial.

Therefore, what is much more important is **substantive significance**. What is substantive significance?

What is the nature of the difference? Is it big enough to care about or so small as to be practically irrelevant? In this case we have to think about the measurement scale and range of the data, interpret what the numbers actually stand for. This can be more difficult (especially if one does not know how to interpret some of the numbers that come out of the models), but it is the only way to go.

The obsession with statistical significance has caused a sort of a crisis on social and behavioural sciences (unimportant effects are overemphasised; publication bias - results that are significant by accident get published and true results that are not, don't; studies are not replicable). By now there are many journals that have banned the reporting of "stars" (that indicate the levels of statistical significance) and encourage authors to abandon this concept.

How can z-scores be useful in this case?

3.2 Type I and type II errors

Type I error: concluding that there is an "effect" when there is not. Rejecting the null hypothesis, when it is actually true. This is when our alpha level lets us down. The alpha level is the probability of Type I error.

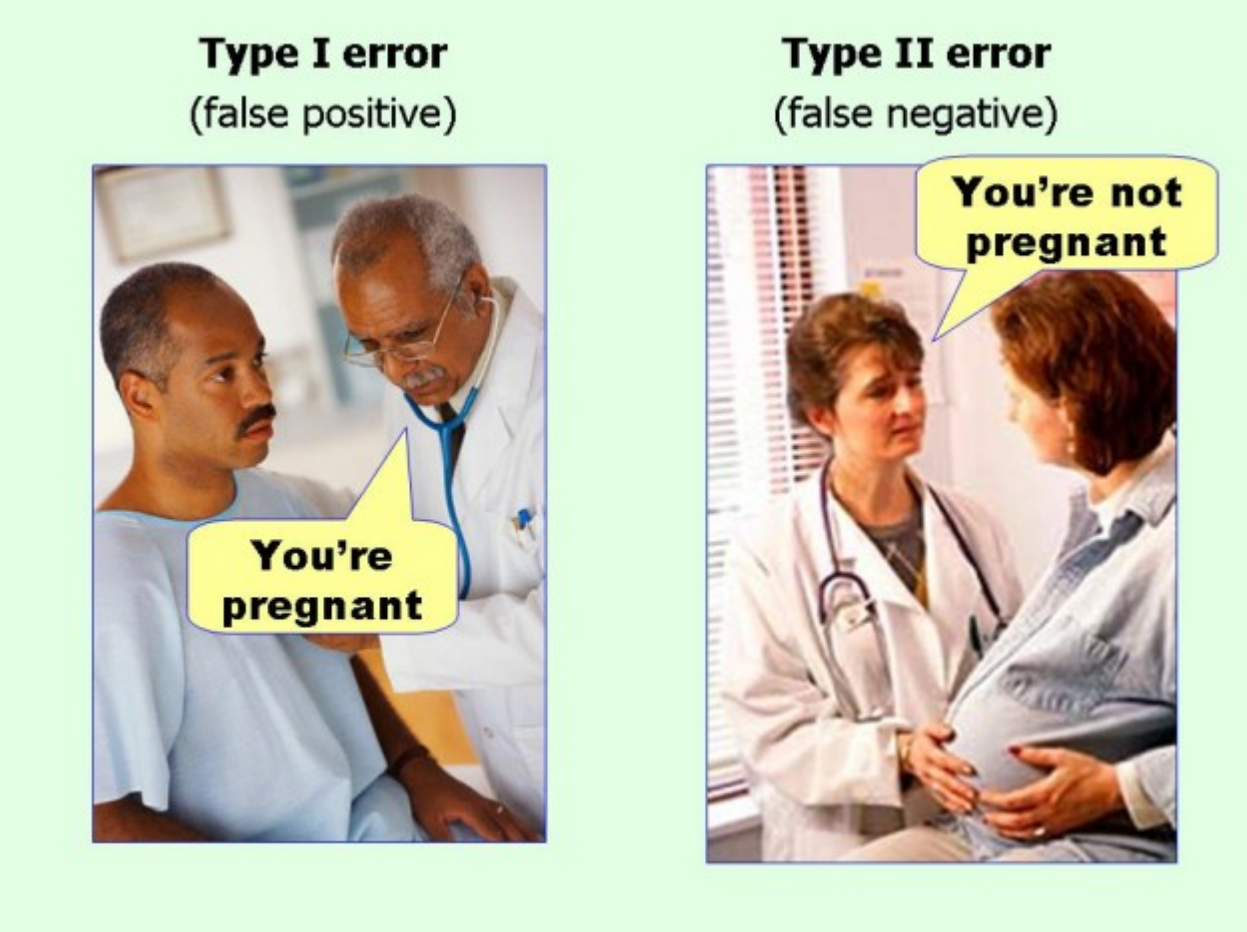


Figure 5: *Type I and Tye II errors.* Source: http://flowingdata.com/2014/05/09/type-i-and-ii-errors-simplified/

Type II error: concluding that there is no effect, while there actually is. We fail to reject the null hypothesis when it is actually not true. This can happen when the effects are really small and we fail to distinguish them from 0. It is not straightforward to determine what the probability of type II error is, as it can depend on many things, including the true value of the effect in the population, which we almost never know (but for which we can make educated guesses).

Statistical power: The power of a statistical test refers to the probability of rejecting a false null hypothesis. The probability of identifying an effect if there really is one. It depends on effect size, sample size and the desired alpha level. Power is one minus the probability of Type II error.

If we have an effect with a certain size in mind, then it is possible to calculate the sample size that we would need to detect that effect.

3.3 The t-statistic and the t-test

Now we come back to that one piece of the puzzle that was missing above. We assumed that we know the population standard deviation for the calculations of the standard error. But this we do not know.

3.3.1 Single sample t-test

Usually we are not interested in the sample as such, but in the population value that we can guess from the sample. And it is here that certain problems come in. Because the sample does not contain the whole population, the variation in the sample tends to be less than the variation in the population. Therefore, calculating the variance for a sample, like was shown above, would give us a biased estimate of the population variance. That is, after all, what we are ultimately interested in.

What is the solution? If you subtract 1 from the sample size and do the calculations, then you will get an unbiased estimate of population variance and standard deviation. To differentiate, we can name them s and s^2.

$s^2 = \frac{\sum(X-\mu)^2}{n-1}$

$s = \sqrt{\frac{\sum(X-\mu)^2}{n-1}}$

Thus, the estimated standard error is:

$SE_M = \frac{s}{\sqrt{n}}$

where $s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$

We can use this to calculate the t statistic:

$t = \frac{\bar{X}-\mu}{SE_M}$

The difference here is that we are not using the standard deviation, but the sample estimate of the population standard deviation. This can be used to test hypothesis about an unknown population mean.

If the distribution of z-scores was the standard normal distribution, then what is the distribution of t-scores or t-statistics? It is not very different and it depends on the number of **degrees of freedom**.

Degrees of freedom. The number of values in the sample that are free to vary. You see this concept a lot and although ultimately it is very difficult to explicate its meaning in a non-statistical way, it helps to keep in mind some "hints". One can also think of it as the amount on "units" in our data (I am not saying "cases", because it can also to some extent refer to other things, e.g. the number of categories in a variable) minus the pieces of information that we have "used". Or the pieces of information that are free to vary? **What does this mean?**

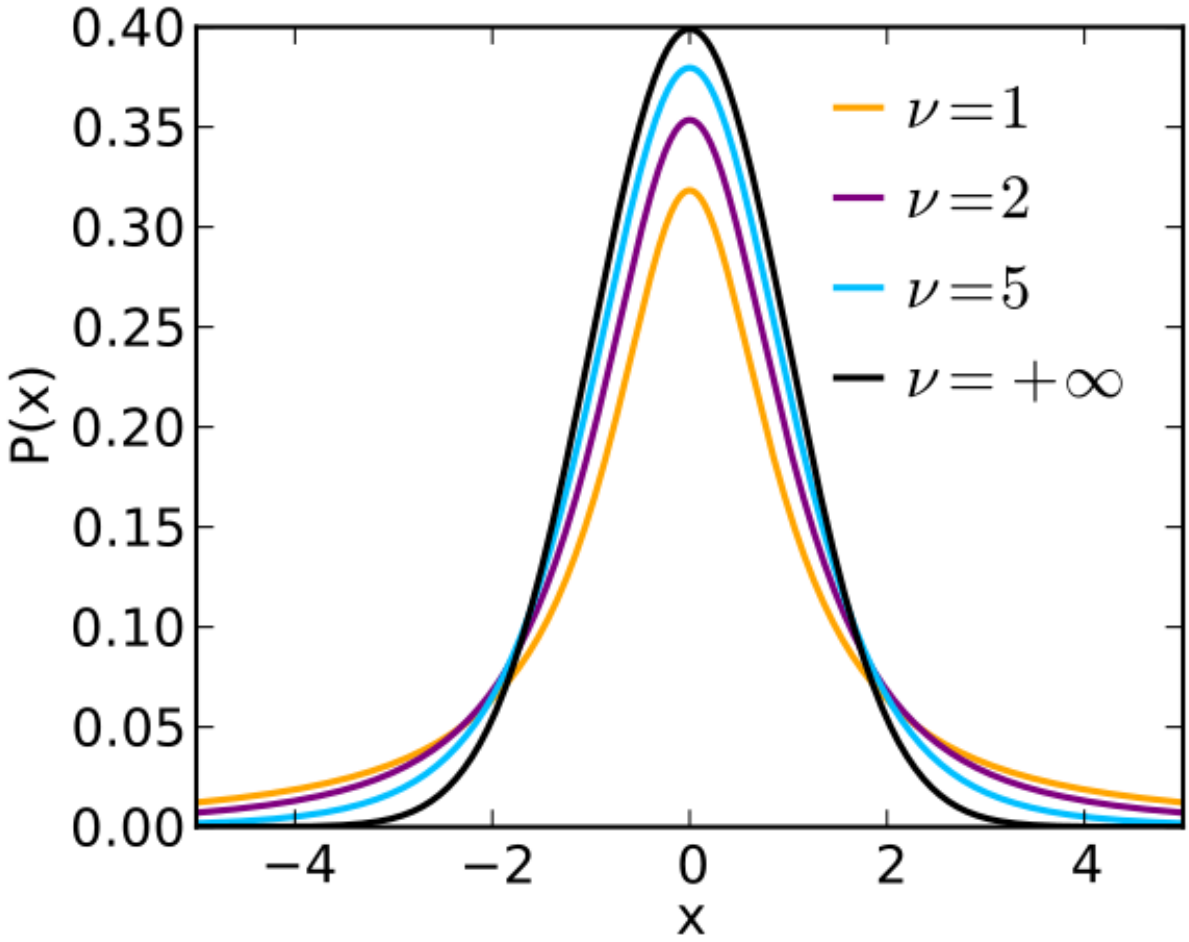


Figure 6: *The t distribution.* Source: https://commons.wikimedia.org/wiki/File:Student_t_pdf.svg

In this case, $df = n - 1$, because this is the degrees of freedom for the sample estimate of the population standard deviation.

For large enough degrees of freedom, the t distribution and the normal distribution are indistinguishable. So roughly the same cutoff points than for z-scores apply, e.g. the value of the t-statistic of 2 indicates roughly the 95% confidence level.

And now we are ready to do a basic t-test.

$t = \frac{sample\ mean - population\ mean\ (hypothesised)}{estimated\ standard\ error}$

Example
Sheet: t-test

This was a single sample t-test. We will also quickly look at the related samples t-test and the independent samples t-test. **What are they?**

3.3.2 Related samples t-test

This is exactly like the single sample t-test, but instead of values for a variable, we are working with the difference between two measures. Let's say you have a survey of the same people twice and you want to test

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Figure 7: *Table of t values.* Source: http://www.z-table.com/t-value-table.html

if the difference that you see in the values of a variable is significantly different from 0. That is where this test would come in.

$t = \frac{M_D - \mu_D}{SE_{M_D}}$

where $SE_{M_D} = \sqrt{\frac{s^2}{n}}$ and s^2 refers to the variation of the difference scores.

3.3.3 Independent samples t-test

The independent samples t-test means that you are comparing two unrelated groups of values for the same variable. Perhaps this is the most common case for the t-test and this happens whenever we are comparing one group to another, for example men and women.

What makes the calculation of this test a bit more complicated is the fact that we are dealing with two samples, thus we have two different means with their corresponding standard deviations and standard errors. Therefore, also the formula differs slightly to take this into account. The general form of the test is the following:

$t = \frac{sample\ mean\ difference - population\ mean\ difference}{estimated\ standard\ error} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{SE_{(M_1 - M_2)}}$

$SE_{(M_1 - M_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

However, there is a problem here. This approach works only if the two samples are exactly the same size. This is almost never the case. The problem comes from the fact that the formula takes the two sample variances equally, but when they are not the same size, the should not be taken equally.

We would need to calculate something that is called the pooled variance:

$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$

The formula thus becomes:

$SE_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$

The degrees of freedom for the t-statistic in this case is the degrees of freedom for the first sample plus the degrees of freedom for the second sample.

3.3.4 Point estimate and confidence interval

We were looking at the confidence interval for the proportion before, let's have a look here at the same for a mean.

The point estimate is simply the sample mean, but this lacks any information about the uncertainty that is related to the value.

$\mu = M \pm t \times SE_M$

t is the value of the t statistic (according to the degrees of freedom that we have), that corresponds to the level of significance that we want to portray. E.g. if we want to construct a 95% confidence interval, we would have to choose a t value that corresponds to the 0.05 significance level for the corresponding degrees of freedom.

3.4 ANOVA (analysis of variance)

The t-test enables us to test if there is a difference between two group means. But often we have more than one group to compare. The analysis of variance (ANOVA) is a test, which tells us if, among several groups, one group is significantly different from the others. The test itself does not tell us which group it is, for that

we have to look at the estimated means of the groups and the confidence intervals or plot the data or the results of the test to see which group stands out.

There are many types of ANOVA, we are going to look at here on only the most simple one - comparing independent samples (similar to the independent samples t-test) and one grouping variable.

Why not just do multiple t-tests?

Terminology:

- **Factor**
- **Levels of a factor**
- **Treatment effect**

The F-ratio and the F-test:

$F = \frac{variance\ between\ sample\ means}{variance\ expected\ with\ no\ treatment\ effect}$

$F = \frac{mean\ SS\ betweengroups}{mean\ SS\ within\ groups}$

We are basically comparing the within treatment variance to the between treatment variance.

Degrees of freedom:

- Degrees of freedom total: $N - 1$
- Degrees of freedom within: $N - k$
- Degrees of freedom between: $k - 1$

Procedure:

1. Calculate the mean for each group.
2. Calculate the overall mean.
3. Calculate the between group sum of squares (difference between the group mean and the overall mean multiplied by the number of observations in the group) and the mean square (using the degrees of freedom).
4. Calculate the within group sum of squares. Sum of the squared values of the difference of each individual observation from its group mean. Calculate the mean square for that.
5. Calculate the F-ratio.

3.5 Correlation

Correlation is a measure of the strength of a linear association between two variables.

However, the value of the correlation coefficient and the strength of the association themselves do not have a linear relationship. We will come back to this in a second.

Correlation is a ratio of the degree of covariance to the degree of independent variance.

Sum of products:

$SP = \sum(X - M_X)(Y - M_Y)$

The rest is familiar to us already from the above.

$r = \frac{SP}{\sqrt{SS_X \times SS_Y}}$

The square of the correlation coefficient is called the coefficient of determination and represents the proportion of variance in one variables that is associated with the variance of the other variable.

r	r^2
0.1	0.01
0.2	0.04

r	r^2
0.3	0.09
0.4	0.16
0.5	0.25
0.7	0.49
0.8	0.64
0.9	0.81

Example
Sheet: Correlation

But also remember, a correlation can be deceiving. Have a look at your data!

How to do a hypothesis test in this case? This is again a t-test, because we are using the t-distribution, but this does not mean that there is anything similar in the calculations.

Degrees of freedom: $n - 2$

$t = r \times \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$

Let's stop here for a second on the slogan **correlation is not causation**.

Useful to keep in mind, types of correlations:

- Pearson correlation. Continuous data.
- Spearman correlation. Ranks.
- Point biserial correlation. Dichotomous variable and continuous variable.

3.6 Linear bivariate regression

A linear relationship can be represented by a line. How can we determine what is the best line to describe the relationship. The least squared method.

$Y = bX + a$

Distance between the predicted and the actual Y, the amount of error.

total squared error: $\sum(Y - \hat{Y})^2$

$\hat{Y} = bX + a$

$b = \frac{SP}{SS_X}$

$a = M_Y - bM_X$

Example
Sheet: Correlation

3.6.1 Multivariate regression

Example
Sheet: Mult reg (CSES)

3.6.2 Assumptions of regression

- **No specification error**.
- **Interval measures with no measurement error**.

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

Figure 8: *Table of F statistics.* Source: http://www.statisticslectures.com/tables/ftable/

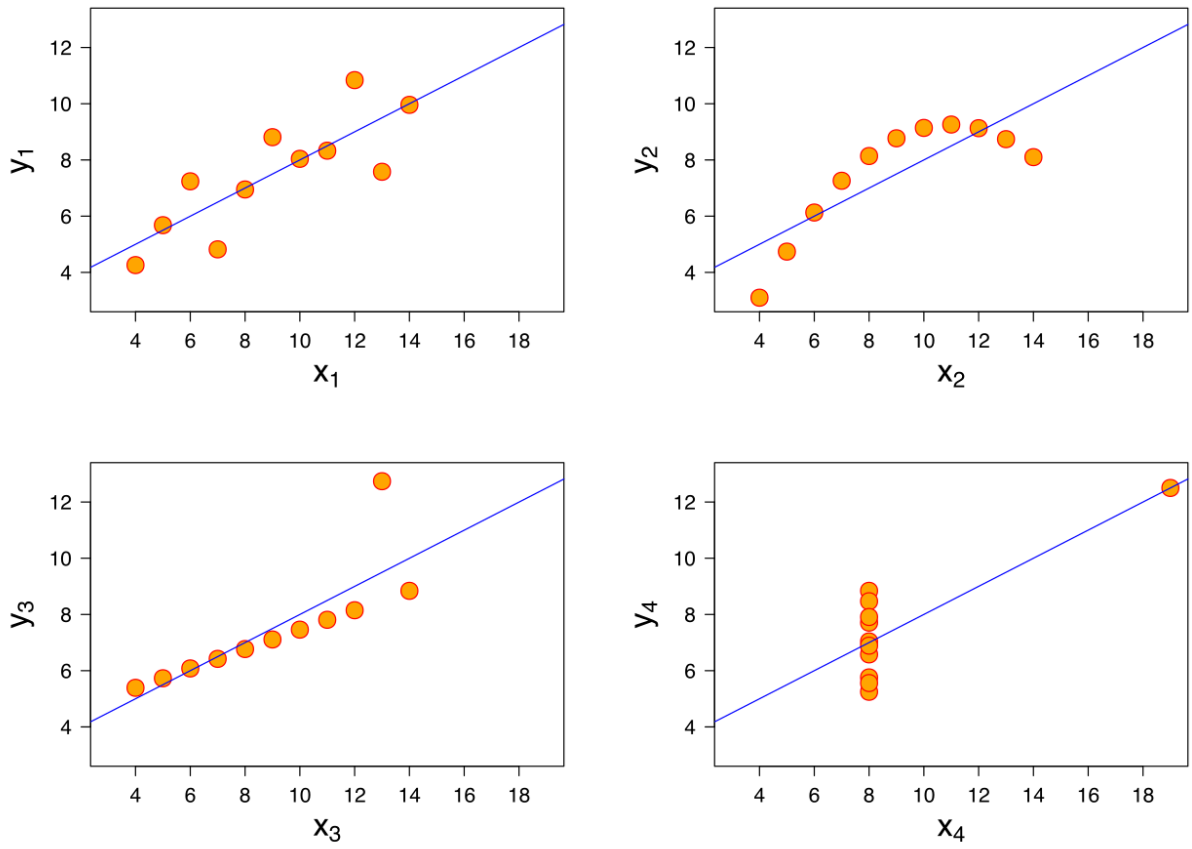


Figure 9: *Anscombe's quartet.* Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512

Figure 10: *Correlations and significance.* Source: http://www.real-statistics.com/statistics-tables/pearsons-correlation-table/

- **Homoskedasticity.**
- **No autocorrelation (errors independent of each other).**
- **Relationships are linear.**
- **No collinearity.**
- **No influential outliers.**
- **Normal distribution of variables.**

3.7 Chi-square test

With survey data, where most of the variables if not all are categorical, it is very often the case that we would want to know if the distribution of a variable is statistically different from some benchmark distribution (are respondents evenly distributed across the categories of a variable, or is there some kind of a pattern?) or is there is an association between two categorical variables (does the distribution of one variable depend on the distribution of the other?). In this context, a chi-square test will give us an answer of whether there is an association or not.

In short, chi-square test in the context of contingency tables (crosstabs, frequency tables) is a way to evaluate the distributions of cases across the categories of a single variable (a one-way table) or the association of two categorical variables (two-way table). It is based on observed and expected frequencies.

3.7.1 Goodness of fit

For the distribution of one variable. How does this look like for a one-way table?

$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

$df = C - 1$

	Proportion in Critical Region				
df	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59

Figure 11: *Chi-square distribution.* Source: Gravetter and Wallnau, "Essentials of Statistics for Behavioural Sciences".

Example
Sheet: Chi 1

3.7.2 Chi-square test of independence

For the joint distribution of two variables, focuses on the distribution of one variable across the categories of another. How does this look like for a two-way contingency table?

The expected frequencies here must reflect the marginal distributions (the row and column totals). If we focus on the rows, then the row totals must be distributed across the column categories according to the column total proportions. The rows can have different numbers of cases, but the relative frequencies across the rows must be the same as the relative frequencies of the column totals. Calculations for that are more simple, we do not have to focus on the proportions, although they are in there as well (column proportion times the row proportion).

$f_e = \frac{f_c f_r}{n}$

The degrees of freedom is the total number of cells in the table minus the number of margins.

$df = (R-1)(C-1)$

And the chi-square statistic is calculated in exactly the same way.

$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Example
Sheet: Chi 2

3.8 Odds and odds-ratios

While the chi-square test only tells us if there is an association, then odds and odds ratios are a measure of the strength of association for contingency tables. They are closely related to probability and perhaps in the beginning they are not entirely intuitive, but nevertheless they are ultimately very useful. Being comfortable with the concept of odds is also necessary to understand logistic regression, which we will look at last.

Let's start with probability, then. The probability associated with a cell in a table is the number of cases in the cell relative to the total (whether it is row or column total or the grand total, depends on what probability we want to know).

What are odds?

$odds = \frac{p}{1-p}$

How can we go from odds back to probability.

$probability = \frac{1}{1+odds}$

There is also a confidence interval associated with odds ratios. How could one calculate it (without going into the details of the calculation)?

$ln(CI) = ln(OR) \pm z\sqrt{\sum \frac{1}{f_{ij}}}$

This gives us the confidence interval on the scale of logarithm of odds. To get the corresponding odds ratios, we need to take the antilog of them, i.e we need to exponentiate.

$CI = e^{ln(OR) \pm z\sqrt{\sum \frac{1}{f_{ij}}}}$

Example
Sheet: OR

What are logarithms and what is a natural logarithm?

3.9 Logistic regression

Binary outcome variables.

Like linear regression, except that we estimate the logarithm of the odds of the outcome variable. This complicates the formula and the interpretation of the coefficients, but not beyond reason.

$logit(p) = ln(\frac{p}{1-p})$

Why logit?

Our outcome is bounded by 0 and 1, but we need to estimate something that can vary from + to - infinity.

Odds can vary from 0 to + infinity. Logarithm of odds can vary from minus to plus infinity, because as a number approaches 0, the logarithm of that number approaches minus infinity. Logarithm on 1 is 0 and the odds of 1 mean that the probability is 50-50. Thus, the logit function is symmetric around 0 with respect to probability.

$ln(\frac{p}{1-p}) = a + bX$

Example
Sheet: Logistic (CSES)

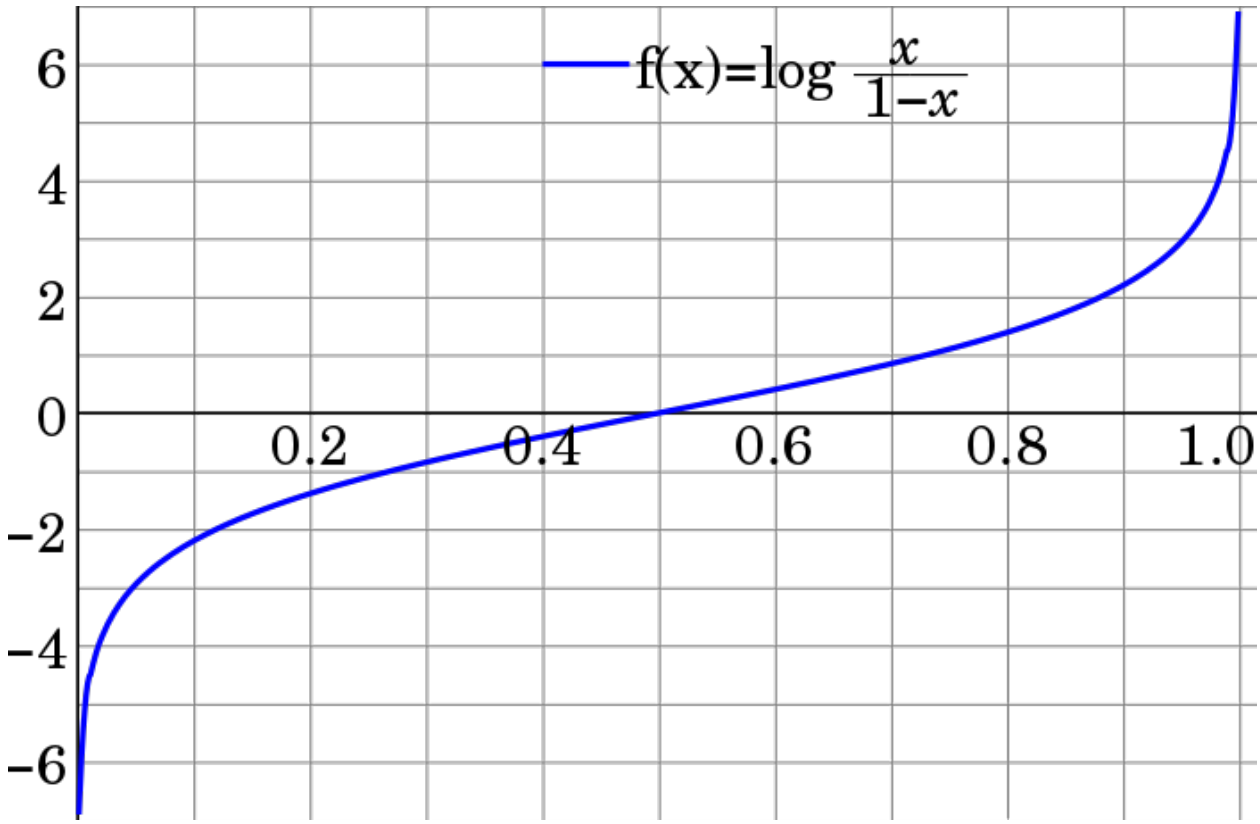


Figure 12: *The logistic function.* Source: https://en.wikipedia.org/wiki/Logit