Quantitative Methods for Communication Students[1]

Petro K. Poutanen[2]

10.2.2014

[1]This material can be seen at http://blogs.helsinki.fi/quantitative-communication/. Note that all the materials are licensed under the Attribution-NonCommercial-ShareAlike 2.5 Generic.

[2]I want to thank Sam Kingsley for correcting the language, and Heikki Hyhkö and Olli Parviainen for their comments to some sections of this work. These contributions improved the quality of this work. However, any errors are mine, and feedback and comments about this material can be sent to me at petro.k.poutanen@gmail.com.

Contents

Chapter 1

Introduction

This document consists of introductory material on quantitative methods. It is aimed at students of communication but is suitable for the other social scientific fields as well. The emphasis will, though, be on observational studies, with a focus on surveys and content analysis. Experimental research settings are not discussed in this document, although they form an essential part of the scientific study of communication. This limitation is due to framing issues, since this material has been prepared for a specific communication methods university course.

The examples presented in this paper are not from real studies, but specifically calculated for the purposes of the discussed themes, so they should not be taken as real research results. I have tried to avoid using any formulas and presented mathematical expressions only when I think it is absolutely necessary for the understanding of the explained matter. However, even then I have done so informally, with a focus on getting an intuitive grasp of the matter.

What will follow in this document is mainly based on my own experiences in teaching quantitative methods courses for communication students during three semesters. Hence, the choices made on what is included and excluded from the presentations of various methods are based on my comprehension of how much information is necessary for students whose background knowledge of statistics and research methods is not that strong yet. I have added a lot of links and options for further readings for those who would like to deepen their knowledge on some issues.

The presentations of issues and methods is not by any means meant to be exhaustive; rather the purpose has been to provide an easy and self-contained

introduction to various options that are available if one chooses to conduct quantitative research. The focus will be heavily on the *hands-on* approach that I have adopted in my own studies of statistics and quantitative methods.

Chapter 2

Quantitative Research

This section covers basic protocols and starting points for developing a research design for quantitative inquiries in communication.

2.1 Starting to think statistically

An important first step is to try to tune one's brain in to seeing things through the prism of statistics. This is not easy though, since human beings are not statistical thinkers. In general, statistical reasoning is based on absolute and relative frequencies and relations between two or more phenomenon. A second step is to try to avoid anxiety and accept the fact that learning statistical methods will take some time and effort. Especially when one is inexperienced, everything feels disconnected and thus incomprehensible. But step by step, things will begin to come together, and some insights and *a-ha moments* will follow.

In statistics everything is connected. To understand regression analysis one needs to understand the concept of correlation. To understand correlation one needs to know what continuous variables, distributions, and linearity mean. Once the knowledge is built gradually and the relevant steps are taken, it will be much easier to begin to understand more complicated methods, such as regression analysis.

First of all, it is relevant to note that the kind of statistical research done in the social sciences is *empirical* in its nature, i.e. we are not doing mathematics with numbers, but using the statistical toolbox (i.e. methods) for making

observations and inferences about empirical phenomena in a rigorous way. Secondly, all the information that is collected should be in a measurable form, i.e. transferable into numbers. By empirical I did not mean the kind of *empiricism* without any theoretical concepts that was conducted by the early positivists in the 19th century. Theories have an imperative role in scientific empirical research in formulating research hypotheses and interpreting results.

2.2 Theories and models

Theories and models form the basis for empirical inquiries. In empirical research we are not only interested in variations in our data (what we observe happening in the world), but we also want to test whether the data fits to our model and a theory. We also want to develop theories or even build new ones based on the previous ones and/or new empirical evidence.

Social scientists usually ask questions concerning humans' social behavior, attitudes, and beliefs. How do people behave in certain social circumstances? Are there any regular patterns of behavior that can be repeatedly observed? For example, what attitudes are associated with a high degree of online participation? What makes people vote for certain candidates in presidential elections? What is the effect of media coverage? How do socioeconomic factors predict and influence media consumption? And so on.

In general, there are two strategies to follow when conducting quantitative research. The first, in which researchers formulate hypotheses on the basis of the previous research and test them against empirical data, is called *confirmatory research.* Confirmatory research confirms (or rejects) hypotheses. The other, possibly supplementary, strategy is *exploratory research.* Exploratory research is by definition exploration, a kind of adventure into the data. This method starts with the data and exploring in order to formulate hypotheses and theories based on the understanding derived from the data.

Some data analyzers may suggest that exploratory data analysis could be followed by confirmatory analysis: first the researcher explores the data, finds some interesting associations, then finds theory to support the observed variations, and finally performs a statistical test to verify what has just been found. However, there is a danger in this reasoning. The data can include correlations which do not really exist but are in the data just by chance. In theory, one can find arbitrary evidence from the data and develop a theory around it to "con-

firm" the observations. This is how human brains mostly work: we see patterns ("evidences") around us and try to explain them. However, patterns are also seen where there is actually nothing at all going on!

This confirmation bias can be harmful in scientific research, and statistical methods are exactly the way of avoiding succumbing to it. Therefore, good scientific research is based on some type of theoretical reasoning, either taken as given and tested against empirical data or developed over the course of data exploration and carefully linked with an existing body of empirical research and theories.

2.3 Research design

Research design consists of the different options and choices to be made when conducting empirical research. A study comprises a sequence of choices, which all influence the validity and plausibility of the research. Among things to consider are research questions and objectives, operationalization of variables, reliability and validity, and data-gathering methods. Some of these areas will be covered in the further sections. Here, we will briefly make a bold separation between two common research designs in social and communication research-experimental and non-experimental designs – and briefly discuss validity and reliability.

Experimental design is based on a well-prepared and framed experiment in which some particular causal relationship is tested under controlled conditions. In experiments people are usually randomly divided into separate groups thus controlling the possible bias caused by the variables that are not studied. Then the interesting variable is manipulated and the possible effect is observed and measured. Good experimental studies test real causal mechanisms.

Non-experimental research design refers to observational studies, such as a survey or a content analysis. In observational studies a researcher collects observations using a research instrument, such as a questionnaire in surveys, and then performs statistical tests on the data. Good sampling techniques can make observational studies reliable and generalizable to the population within certain limits. No causal mechanisms can, however, be tested as reliably as in controlled trials.

There are two important criteria that should always be considered with respect to a given research design. The first of them is *validity* – is the study

really measuring what it claims to be measuring? For example, if the study design is a survey that intends to measure people's attitude towards social media, we can evaluate how well the theoretical constructs are operationalized into survey questions and how well the items used describe attitudes towards social media. Validity will in the end determine how well the study can predict the behaviors or attitudes it measures.

Reliability is the other criteria, which could be called "repeatability". It describes how accurate the measurement is, i.e. if the same study were conducted again, would similar results be drawn? It is useful to start thinking about reliability by asking how consistent the measurable constructs would be from one sample to another. There are many ways to test reliability, such as *test-retest* (how much two different samples correlate together for the tested questions) or reliability estimates, such as *Cronbach's alpha*.

2.4 Research process

In textbooks (and guides like this) one can often find an idealistic and unrealistic model for how the research process should proceed. These process models are, however, useful in describing what different stages there are in conducting a piece of research, and especially when planning one. Since we are focusing here on non-experimental research designs, a process illustrating a non-experimental research process is presented. The following outline of a research process includes 8 steps of which steps 1–3 are related to the preparation of the study, 4 to the data collection, 5–7 to the actual analysis, and 8 to presenting the results.

1. Exploring theory (i.e. reading books and scientific articles)
2. Determining the research questions and formulating hypotheses
3. Defining what/who the observational units are (i.e. people), the variables (what is to be measured/asked) and the sample (how the observational units will be selected, i.e. the criteria for inclusion and how many there will be of them)
4. Collecting data and preparing it for analysis
5. Describing the data through frequencies and relations between different variables
6. Conducting statistical tests and inferences

7. Making any required additional inquiries to the data and testing cause and effect relations

8. Interpreting the results and writing the research report.

2.5 Further resources

A good place to start training in statistical thinking and reasoning is of course any text book. Good "open textbooks" can be found online, such as OpenIntro Statistics.

Plenty of great online introductory courses on statistics, probability, data analysis, etc. are also available. For example, Coursera provides high-quality massive open online courses (MOOC's) on topics such as data analysis, computational analysis, R (statistical programming language), and network analysis. To begin with, take a look at Scott E. Page's *Model Thinking* (University of Michigan), Andrew Conway's *Statistics One* (Princeton University), or *Data Analysis and Statistical Inference* (Duke University) taught by Mine Çetinkaya-Rundel.

Khan Academy's free tutorials on statistics and probability are of great quality and interest for those who want to learn more about the mathematics of statistical reasoning and methods. Khan Academy is an educational website hosted by Salman Khan, consisting of 4,000 of his *micro lectures* on various topics.

Good online sources for familiarizing oneself with communication theories and models are also available. Communication theories can serve as testable theories or models for building up own one's own theoretical framework.

- *Communication theories*is a comprehensive resource of different communication theories maintained by Twente University. It is a well-organized catalog of different theories arranged according to their level and purpose. Also the most important references and practical examples are presented for some theories.

- *Theory for communication*is a "work in progress" web page by the University of Colorado Boulder and it comprises a massive amount of information about communication theories. There are plenty links and lectures concerning important theorists, different research paradigms, and so forth.

- *A First look at Communication Theory*is a web site for a communication theory text book by Em Griffin including an overview of many of the

important theories in the area of communication studies.

Chapter 3

Methods

This section introduces three types of different methodological approaches for quantitative research in communication: survey, content analysis, and network analysis.

3.1 Survey

A survey is one of the most popular data collection and research methods across the social sciences and market research. So much so, in fact, that it has become difficult to get people to answer survey questionnaires. Moreover, online surveys have arrived and changed the game a little, making data collection easier and cheaper, but at the same time subject to many biases.

In essence, a survey is a research instrument that usually takes the form of a structured questionnaire. The questionnaire data can be collected through face-to-face or phone interviews or by sending an invitation to answer to the questions by post or via email. In research methodology jargon, survey design is called a *cross-sectional study* if it is conducted only once, or a *longitudinal study* if data is collected more than once over time.

3.2 Content analysis

Content analysis is a widely used technique in communication sciences. It is a method by which some observable *contents*, such as texts, images, objects, etc. are transferred through coding into a measurable and verifiable form. The

process of conducting a content analysis follows the general research process (see section 2.4) of quantitative inquiries. Perhaps the most distinctive part is the construction of a coding scheme, a detailed instruction for identifying and classifying the *units of analysis*.

The unit of analysis refers to what is counted, what is the "analytical whole" taken as a separate unit. Originally, content analysis was introduced to the communication sciences to bring analytical rigor and objectivity. Therefore, the "content" was originally defined as *manifest*, i.e. observable and as unambiguously counted as possible. Examples of such units are the lengths of newspaper articles as measured in inches or the numbers of different sexes in different scenes of a soap opera. Content can also be visual, non-verbal or even audible. Manifest content is easy to determine beforehand and to identify from the data.[1]

The identification of contents can be subject to the subjective biases of the coder. More objectivity can be brought to the analysis by using more than one coder and calculating an *inter-coder coefficient*. In this way, the reproducibility of the results by other researchers can be used as a measure of reliability. There are many ways to do the calculation, and the whole data does not need to be double-coded, but a fraction of it, such as 10–20%. One of the most-used reliability measures in this context is *Cohen's Kappa*.

3.3 Network analysis

Network analysis, and especially social network analysis, has always been a popular method in communication sciences, although recently it has become even more popular. Nowadays, the idea of *social networks* is so widespread that almost everyone has some notion of what a social network is. For example, one of the most well-known ideas about social networks is that every two people are, on average, six steps away from each other. This *six degrees of separation* thesis was famously tested empirically by sociologist Stanley Milgram in the late sixties. In the age of Facebook, the virtual length between every two Facebook users has shrunk to 4.7 steps. This illustrates the "small-world structure" of social networks, meaning that social networks comprise highly connected clusters of nodes (people), which have several "weak ties" between them holding the whole

[1] Units of analysis can be less manifest though, i.e. based on the subjective interpretation of the coder. An example of such unit of analysis could be attractiveness or an emotional tone of voice, positivity or negativity of a speech act, etc.

network together and making it possible to reach almost anyone quickly.

A network is comprised of nodes (e.g. people) and edges (e.g. social connections). When planning to conduct a network analysis the researcher will need to consider what the units of analysis are, i.e. who are the actors. They can be people in a small workplace or in a community, but they can also be nation states or power plants. One also needs to consider the sample: what is the population of actors that comprises the network? The other crucial thing to consider is to determine the form of the tie between the two units of analysis: what makes the relation? Usually the connection takes the form of some type interaction between the actors, which can be directional or mutual.

For beginners, many great examples and easy instructions for conducting a social network analysis with Facebook or Twitter data can be found atthis blog, for example. Network analysis of one's own Facebook friends provides both fun and excitement and is a good step towards conducting serious research using network analysis. For these purposes this blog provides a lot of good resources to begin with.

Chapter 4

Data

This section concerns issues related to data gathering and sources of data.

4.1 Data and variables

Data refers to a set of values, which are usually organized by *variables* (what is being measured) and *observational units* (members of the sample/population). An example of data is a data matrix in a spreadsheet program, such as *Excel* or *SPSS* (see figure 4.1). Along the upper horizontal line there are the variables (e.g. survey questions) and down the first vertical line there are the observations (e.g. people). In each cell there is a value that is the given observational unit's value of a given variable.

Variables are of different types and can be classified in many ways, for example as *numerical* and *categorical* variables. Numerical variables are measured by some (usually existing) measures, whereas categorical variables are qualitative, and not necessarily more or less, or bigger and smaller than one another. Another way of classifying variables is according to their *measurement scale*.

The *continuous variable* is numerical and it can take, in theory, an infinite amount of values. An example of such a variable is length in centimeters or inches. The *discrete variable* is also numerical, but differs from a continuous variable in that it takes a finite number of values. An example of a discrete variable is a performance score from 1 to 10. Each value is, in theory, equally far from the subsequent value, so that 4 is exactly the same increase from 3 as 9 is from 8.

Figure 4.1: An example of a data matrix in SPSS program. The figure shows a data matrix of World Economic Freedom data. The countries on each row are observational units or "cases", and the columns of different freedom indexes, such as *business freedom*, are variables. Each country then has its individual cell scores for the variables.

	Country	Overall	Business	Trade	Fiscal	GovtSize	Monetary
1	Afghanistan	.	.	.	.	.	.
2	Albania	63,7	67	75,8	92,8	75,6	79,6
3	Algeria	56,6	73	68,6	77,2	74,1	78,6
4	Angola	47,0	44	72,0	85,2	62,2	61,8
5	Argentina	52,3	62	70,0	70,3	75,6	60,6
6	Armenia	69,9	84	86,4	90,0	89,7	77,8
7	Australia	82,6	91	84,8	61,4	64,3	84,7
8	Austria	71,2	79	85,8	49,9	27,1	80,9
9	Azerbaijan	58,0	75	78,4	79,7	77,5	66,3
10	Bahamas	70,3	75	56,0	96,5	84,4	75,5
11	Bahrain	74,8	80	80,0	99,9	79,4	74,0
12	Bangladesh	47,5	63	40,2	72,8	94,2	67,3
13	Barbados	71,5	90	64,6	70,9	66,3	74,5
14	Belarus	45,0	64	67,2	79,4	30,9	66,8
15	Belgium	72,1	93	85,8	41,5	28,3	81,0
16	Belize	63,0	76	69,6	68,8	77,6	78,4
17	Benin	55,4	44	67,4	67,2	88,5	79,6
18	Bhutan	57,7	62	42,0	83,6	58,3	75,9
19	Bolivia	53,6	59	81,8	84,8	71,9	67,9
20	Bosnia_Herzegovina	53,1	60	77,2	71,8	37,6	79,0
21	Botswana	69,7	71	69,0	75,1	70,6	71,8
22	Brazil	56,7	54	71,6	65,8	50,3	77,2
23	Bulgaria	64,6	74	85,8	86,2	58,7	72,8
24	Burkina_Faso	59,5	59	70,4	80,7	83,7	83,7
25	Burma	37,7	20	72,2	81,8	98,5	45,3
26	Burundi	48,8	34	63,0	72,5	55,8	72,7
27	Cambodia	56,6	43	63,4	91,4	94,5	80,0
28	Cameroon	53,0	39	56,0	69,6	93,6	76,0
29	Canada	80,5	97	88,2	76,6	53,7	80,8
30	Cape_Verde	61,3	57	65,4	65,5	62,2	76,9
31	Central_African_Republic	48,3	40	50,4	65,4	94,0	75,0
32	Chad	47,5	36	58,4	50,5	90,8	78,6
33	Chile	78,3	66	85,8	78,2	90,1	77,3
34	China	53,2	52	71,4	70,6	88,9	72,9
35	Colombia	62,3	77	72,4	72,9	65,9	70,6
36	Comoros	43,3	47	27,2	64,6	88,1	78,9

If a numerical variable has an absolute zero, the variable can be measured on a *ratio scale*. A typical example is weight. It can be zero, but definitely not less than that. Therefore we can say that one observation is twice as heavy as the other one. If the variable has no zero point, it has an *interval scale*, meaning that the distances between different values are same (e.g. from 10 to 20 and from 40 to 50), but the zero point is arbitrary. An example of an interval scale is temperature. A centigrade thermometer can show "0", but the quality itself, temperature, does not cancel out. The zero point is arbitrary. Therefore we cannot really say that it is "twice as cold" as yesterday.

Categorical variables, in turn, can be *nominal*, in which case there is no order at all: each category has its unique meaning ("What domestic pets do you like most: 1 = cats, 2 = dogs, 3 = hamsters, 4 = bunnies?"). If there is a sense of order there, the variables are called *ordinal*. A *Likert-type scale* represents an ordinal measurement: 1 = “not like me at all”, 2 = “not like me”, 3 = “not sure”, 4 = “somewhat like me”, 5 = “very much like me”. A special type of variable is the *dichotomous* one. It can have only two values (e.g. gender) and it can be interpreted both as numerical and categorical.

4.2 Sampling

It is relatively easy to observe people’s behaviors or ask them what they think about something. However, it is another thing completely to claim that these observations (e.g. ticks on a survey form) apply to other people than those who we have asked as well. The problem of how general our results are can be solved by sampling methods. The size and representativeness of a sample define the limits for inferences that are made on the basis of the sample data. In general, researchers collect from a few hundred to thousands or even tens of thousands of observations. But even with a sample as small as 30 cases some conclusions can be drawn.

One of the most critical issues in data analysis is to understand what the difference between the *sample* and the *population* is. Basically, a sample is subset of a population. It is not just any subset, however, as a researcher usually has some kind of idea of what kind of subset it is, i.e. how the sample has been selected. In the ideal case, 1) each member of the population has the same probability of being selected for the sample and 2) they are independent with regard to the measurable characteristic. Then we refer to a *random sample*.

This is an ideal case and we can define the exact limits within which the results drawn from a random sample can represent the whole population.

The fact that a random subset of a larger population can be representative is based on *the central limit theorem* (CLT), a finding belonging to the area of probability theory. The central idea of CLT is that given the sample size and the variation in population, a sufficiently large number of samples (or their mean values) summed up will yield a bell curve -shape *normal distribution*. As we know what the properties of this distribution are, we can assume, under certain conditions, that the mean of the sample we have, is near to most of the other samples' means, meaning that we don't need to take hundreds of samples but only one, randomly selected. This finding forms a basis for statistical estimation and testing, which are covered in the analysis section 5.

In reality, a random sample is not easily achievable, and there are other sampling strategies too. For example, in a *convenience sample* the members of the population are selected on the basis of their availability. Results drawn from studies based on this type of sample are not generalizable to any population. If the study is conducted for a sufficiently small population, such as a firm with 500 workers, sampling need not be used. All the workers can be sent an email inquiry to participate in the study. Researchers need to evaluate then how well those who participated represent the whole firm (for example, are all the departments included) and whether there are any possible sources for systematic bias (such as people without computers, who only rarely read their emails).

4.3 Getting the data

In addition to traditional data-gathering methods such as interviewing and observing, the Internet, digitalization, and the recent development of computational techniques have opened up new possibilities. The use of mobile phones, credit cards, customer-loyalty cards, social media, web browsing – all leave behind digital traces that can be used as research material on our social behavior. The massive accumulation of this type of data has led some to talk about big data and its possibilities for changing our society. However, although not just anyone can get access to such data files, there are many new intriguing ways for gathering data from the Internet that are accessible to all. For example, data from social media forms an important part of communication research today.

We will not cover the techniques for computational data mining here, but

refer to some examples of collecting digital data easily. For example, data on real social networks can be collected from social network services such as Twitter and Facebook. For example, an *Excel* plug-in called *NodeXL* enables users with direct connections to social networks, such as Twitter, Facebook, YouTube, Flickr, Wikis, and emails to mine data (take look at these handouts on how to use NodeXL). An easy option for mining network data from Facebook is also provided by an app called *Nettvizz*. Many more tools and techniques for data collection can be found at DIRT: Digital Research Tools Wiki that has gathered together hundreds of tools for conducting research in the digital environment. It lists tools from data mining to reference management, from quantitative to qualitative data, and from commercial to open software.

If one is interested in extracting information from websites, *web scraping* is the order of the day. OutWit Hub is an example of a tool for scraping. By this tool a researcher can collect data from web pages directly and automatically without painful copy-pasting. Scrapers can be used directly via web browsers. OutWit Hub is available for Mozilla Firefox, and for Google Chrome there is another option that can be used. Software requiring programming skills, such as R, offers several possibilities for gathering different types of online content. Take a look at this nice example of gathering data on consumer attitudes towards an airline company through Twitter.

Last but not least, it goes without saying that there is tons of open and free data online that has already been gathered by someone and is made available for anyone to analyze. Data collected by governments, NGOs, and think tanks is readily available. Sources for open data sets can be found through governmental data pages and, for example, through a platform held by the *Open knowledge foundation*, an organizations promoting for the idea of open data.

4.4 Preparing the data

Before doing any analysis it is necessary to prepare the data so that it is as error-free as possible and organized so that it is suitable for spreadsheet programs to read it. There are many sources of error, such as coding errors, that should be checked before starting any analysis. Otherwise the risk of drawing incorrect conclusions because of biased data remains. The first thing to do is to browse through the data and try to find the possible errors manually. All the variables should also be checked for their range of possible values - for example, there

are no 999-years-old people - so checking the minimum and maximum values for all variables is useful. The convention for coding "missing values" (i.e. non-responses) is to mark them as 9999, N/A (standing for "no answer"), or blank. You should check whether the program you are using interprets missing values correctly.

Another type of data-preparation problem can occur if the data has been gathered by hand or scraped from a website, discussion boards, Facebook, etc. Data that comprises text, such as names of people or companies, can include spelling errors, missing letters, etc. Usually such errors have to be dealt with by hand, but nowadays there are good computational resources for the purposes of data cleaning, such as Google's *Open refine* or *Fusion Tables*. Open refine is easy to use, and there are many good tutorials on its main functions. Open refine includes algorithms that can, for instance, help to identify spelling errors and unify textual values. Once ready, the data can be transferred into a CSV file and uploaded into any spreadsheet program.

Chapter 5

Analysis

This section discusses some of the most common statistics and statistical analysis methods.

5.1 Descriptive statistics

Broadly speaking there are two types of research strategies: exploratory and confirmatory (see: Theories and models 2.2). In statistical data analysis, the descriptive part of the analysis can be seen as a kind of exploration. Its purpose is to get a general view of the data and the distributions of the variables by diagrams, tables, and basic statistics, such as mean and standard deviation. The descriptive analysis is a necessary part of the research and is always conducted before doing any statistical tests or more complicated modeling. This part presents some common techniques for descriptive data analysis, while the next section Inferential statistics 5.2 focuses on statistical testing and modeling.

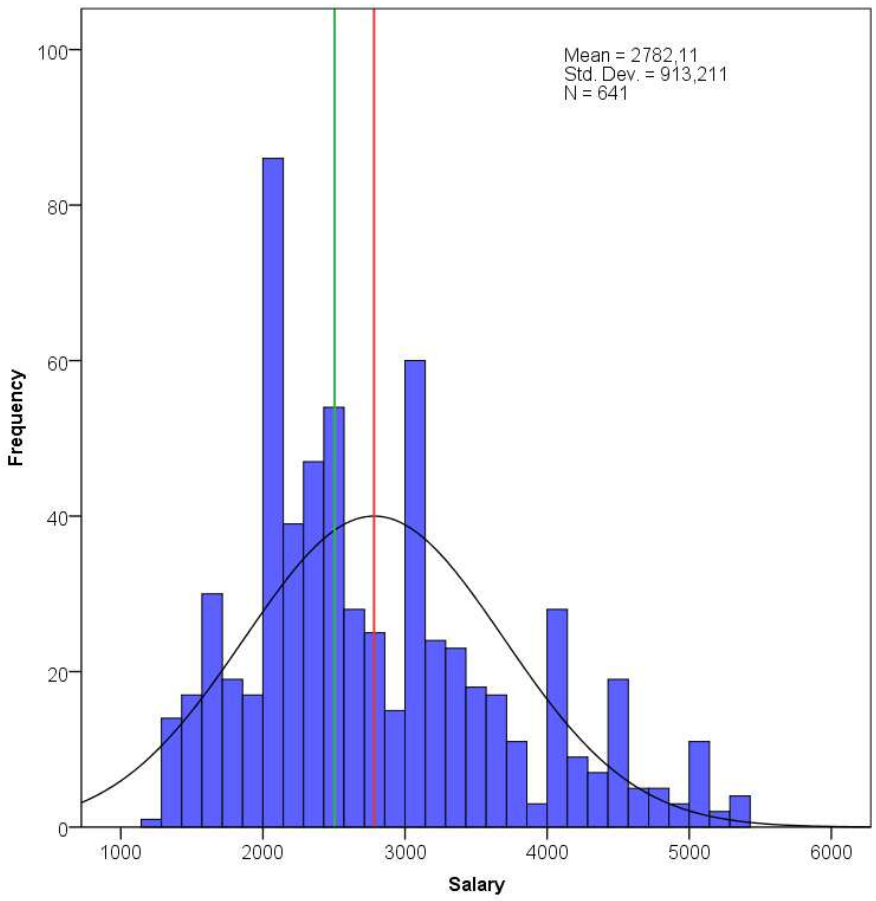
5.1.1 Statistics for single variables

Visual presentations are indispensable means for exploring the distributions of different variables. A variable's distribution is basically a set of all the different values that the variable can take and their observed frequencies. For visualizing the distribution of a numerical variable a *histogram* is the usual choice (5.1).

The histogram in the figure 5.1 shows the frequencies of the observed salaries.[1]

[1] Salaries lower than 1238€ representing the minimum wage, and bigger than 5300€, representing the top 10% of the observed values, are cut off for making the variable's distribution easier to analyze for the purposes of these examples.

Figure 5.1: A histogram of salaries with the mean (red line) and median (green line) and normal curve.



On the horizontal line are salaries and on the vertical line the corresponding frequencies (how many times a certain value has been observed). Not all observed values are shown in a histogram, but adjacent values are grouped together into several bins that are represented as the bars in the histogram. The important thing to consider is how *symmetrical* the distribution is. The more symmetrical it is, the easier it will be to describe using simple statistics. The distribution of salaries is clearly skewed to the right, as there are relatively very few people who earn a lot (the tail of the distribution goes right) and the bulk of the observations are located near to the mean (red vertical line) and the median (green vertical line) a litle bit left from the center of the distribution.

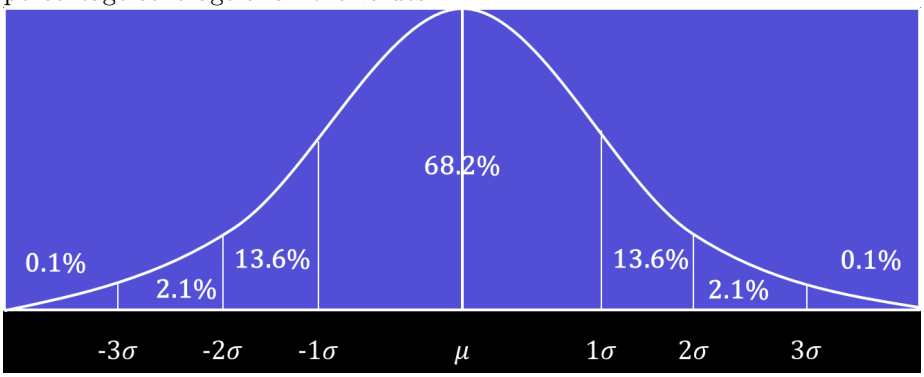
Among the simplest possible descriptive statistics are the measures of *central tendency* and *dispersion*. The measurement scale (see section 4.1) and distribution of the variable generally determine which statistics would best describe the variable. For numerical variables we can calculate an *arithmetic mean*. The symbol for population mean is μ (the Greek letter "mu") and for the sample mean $\bar{x}$. The sample mean is what we can calculate on the basis of our data, and it serves as an estimate for the population mean, which is a theoretical "real" value in the population. The sample mean is calculated by summing all the individual values together and dividing that sum by the number of observations: $(x_1 + x_2 + \ldots + x_n)/n$.

For the distribution of salaries above the mean is about 2782, which is located slightly to the right of the peak of the frequencies, since the distribution is skewed to the right (the extreme values "attract" the mean). It is easy to see that the mean value does not capture the skewness of the distribution. Actually, the mean best describes the central tendency of a variable when the variable is fully symmetrical. However, if it is not, we can use the median (Md), which is the middle value of the all values ranked on a scale from minimum to maximum (or a mean of the two middle values in case of an even number of different values). In this case, the median is 2505 and probably a better measure of central tendency, since it is robust to the effect of extreme values.

The most basic measure for dispersion is called *standard deviation*. The symbol for population standard deviation is σ (the Greek letter "sigma") and for sample it is s. Standard deviation describes how far the values are on average from their mean value. The standard deviation of the salaries is 913, meaning that the values are dispersed on average 913 euros away from the mean value.

Symmetrical distributions are useful, because if we know the mean and the standard deviation, we know quite a lot of other things about the distribution

Figure 5.2: A symmetrical curve with standard deviations and corresponding percentage coverage of all the values.



as well. In the case of a symmetrical distribution about 68% of the all values are within one standard deviation from the mean, and almost 95 % of all the values are two standard deviations away from the mean (see figure 5.2).

For describing the distribution of a categorical variable a *bar chart* is used instead of a histogram. The bar chart (see figure 5.3) displays different categories of a variable measuring how often people say they use the Internet for personal matters. The variable varies from non-users to those who spend time on the Internet daily. The measurement scale is ordinal, since categories can be meaningfully ordered, but the intervals are not necessarily equal. The first category "no access" is different from the other categories in the way that it could be omitted from the analysis or handled independently.

For categorical variables central tendency measurements are not usually calculated, except mode (Mo). Mode gives us the frequency of the biggest category. In this case it is the category "Every day", which has 1012 observations in it. The bar chart could also be drawn with relative frequencies on the vertical axis, in which case each category would have a value in percentages instead of absolute frequencies. If the distribution is fully symmetrical, all the measures of central tendency – mean, median and mode – are situated in the middle of the distribution where the peak of the frequencies are.

5.1.2 Statistics for two variables

When describing distributions of more than one variable at the same time, the interest is usually in the relationship between the two variables. For example,

Figure 5.3: A bar chart of personal use of the Internet.

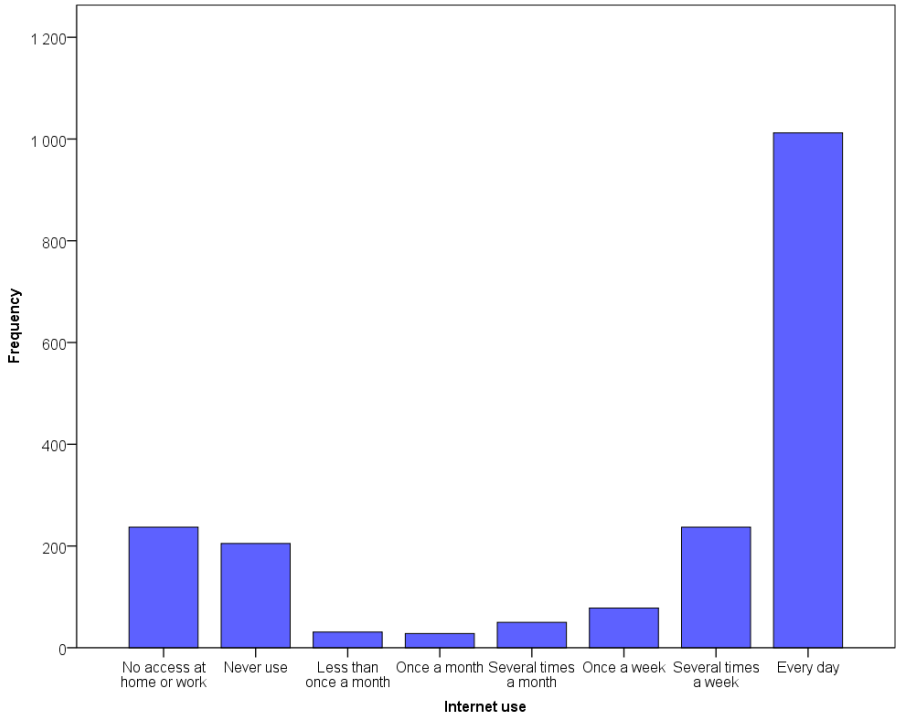
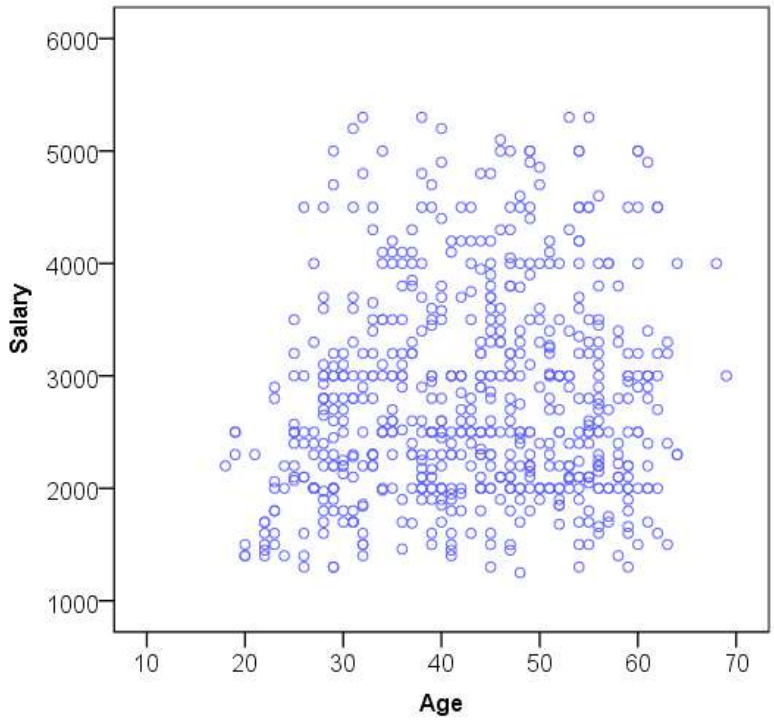


Figure 5.4: A scatterplot of age and salary.



one could be interested to find out if age and salary are somehow related. Do the values of age get higher, for example, as salary increases? In the case of two continuous variables we can draw a scatterplot (5.4) figure and study how the values of the two variables vary together, i.e. whether there is any recognizable pattern. However, it is important to note that no causal claims (such as X causes Y: $X \rightarrow Y$) should be made on the basis of this kind analysis.

In the scatterplot (see figure 5.4), the association seems to be relatively weak over the different ages. Salary hovers around its median value (2505) for all the ages from 25 onwards. Some higher salaries can be seen across all ages. These exceptional observations are grouped together in the tail that was pointing right in the histogram (5.1). A very slight ascending pattern could be observed.

A scatterplot is the standard method for exploring the relationship between two variables, but it works only for numerical continuous variables. For exploring the relationships between two categorical variables we need a contingency table or a special kind of bar chart. A contingency table is a table with absolute

Table 5.1: A contingency table of gender and the personal use of the Internet with absolute and relative column frequencies.

internetuse * gender Crosstabulation

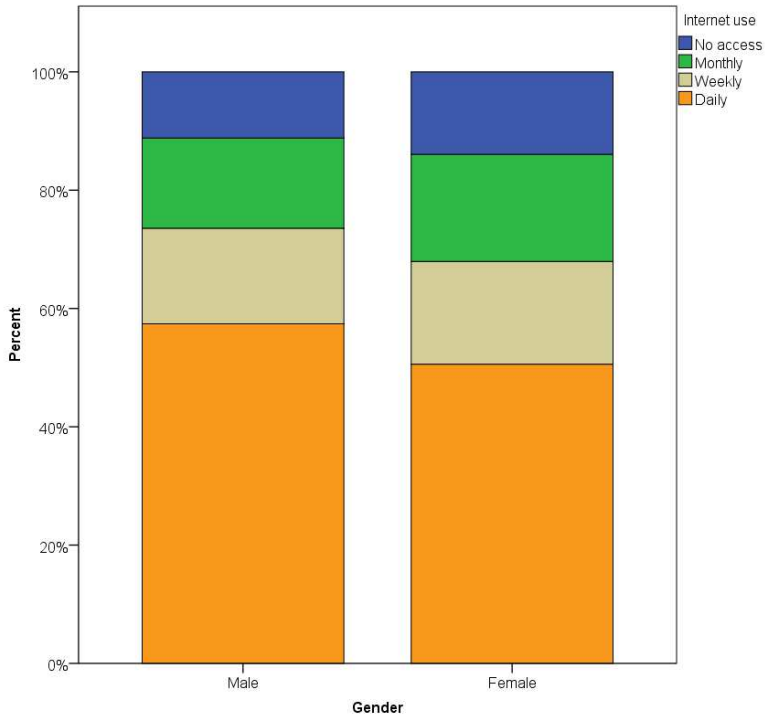
		gender		
		Male	Female	Total
internetuse	no access	102	135	237
		11,2%	14,0%	12,6%
	never use	91	114	205
		10,0%	11,8%	10,9%
	monthly or less	48	61	109
		5,3%	6,3%	5,8%
	at least weekly	147	168	315
		16,1%	17,4%	16,8%
	daily	523	489	1012
		57,4%	50,6%	53,9%
Total		911	967	1878
		100,0%	100,0%	100,0%

or relative frequencies and it can be used for 2–3 categorical variables. It depicts the frequencies of one variable across all the classes of another variable. Usually of the interest are the relative frequencies (i.e. percentages), since the sample sizes of the classes that we are interested in comparing can vary.

An example will clarify this. Let us take "gender" as the first variable and "personal use of the Internet" as the second variable. Now, we are interested in finding out whether the values of males and females (that is, the classes of the first variable "gender") vary similarly across the different values of the personal use of the Internet. To do this we need to use relative frequencies, as the absolute frequencies are not comparable due to different amounts of males and females in the sample. The relative frequencies are usually presented so that they are summed up at the bottom of the columns, i.e. for each of the classes individually that are compared together (see table 5.1).

Relative frequencies can also be depicted visually by a bar stacked chart (see figure 5.5). In this case, bars represent different gender classes, and Internet use is presented as relative areas within the bars, summing up to 100% for each bar. On the basis of the table or the bar chart, we have good reason to suspect that there are slightly more daily Internet users in the male category than

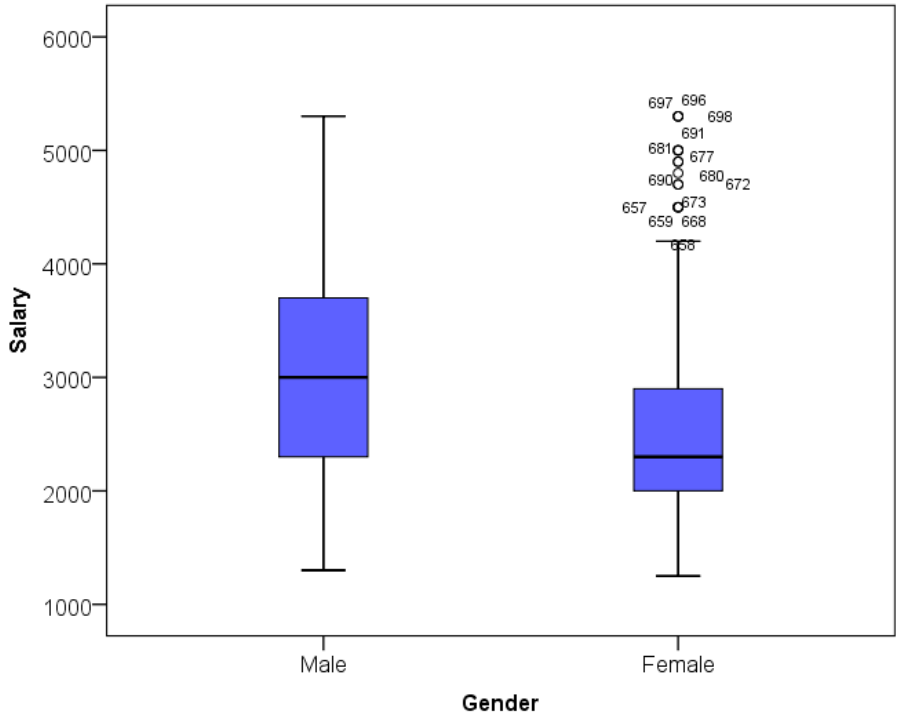
Figure 5.5: A bar chart of gender and the personal use of internet.



the female, and slightly more those who have no access at all among females. This relationship could, however, diminish or change if a third variable, such as different age groups, is introduced to the table. Therefore it is often not enough to only study the relationship between two variables, rather several variables together with different combinations should be explored. The results should also be tested statistically to verify that the observed differences are arguably more than just random variation, which we will discuss in section 5.2.

There is a third form of visual presentation, the *box-plot*, that is useful when a categorical variable needs to be studied against a numerical variable. The box-plot has the categorical variable on the horizontal axis and the numerical variable on the vertical axis. Each category of the categorical variable gets its own box with two whiskers. The boxes and whiskers illustrate the distributions of each class of categorical variable across the values of the numerical variable. The ends of the whiskers represent maximum and minimum values, the stars and points are outliers, the box contains 50 % of all the observations, and the

Figure 5.6: A box plot of gender and salary.



black line within the box is the median value. In the figure 5.6 salaries of males and females are explored by plotting the categorical gender variable against the continuous salary variable.

This plot tells us that the median salaries of males (3000 euros) are higher and that there is more dispersal in the salaries of males (the box including 50 % of the observations is taller). We can also spot numerous *outliers* (exceptional observations) in the high end of females.

5.2 Inferential statistics

Whereas descriptive statistics are necessary for making sense of the data and exploring it, inferential statistics, especially significance testing is what statistics is all about. The idea of doing statistical tests is related to the discussion on sample: if we have a representative sub-set of a population as our sample, we can make statistical inferences concerning the population on the basis of that

sample. These inferences are then generalizable within particular limits, called confidence intervals, that need to be specified.

5.2.1 Confidence intervals

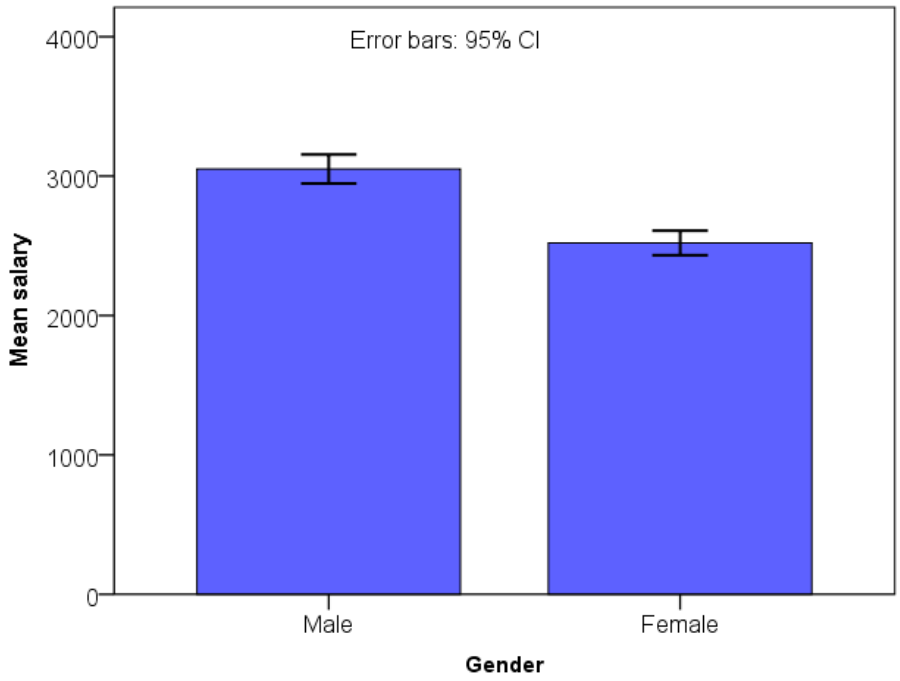
Let us start with a concrete example. In the descriptive statistics section5.1 we calculated a mean value for a cropped income variable and got 2782 euros as the overall mean of salaries. The number of observations (n) was 641. If we randomly take one of the cases from the data and check whether that person's salary is close to the mean, it is likely that it is not exactly the same. There is some distance between the mean and the individual value, either up or down from the mean value. In order to get a picture of all the differences in our sample, we usually calculate the standard deviation, which describes the average distance of data points from the mean. The standard deviation (s) of our salary variable is 913 euros.

Using this information it is possible to calculate the *standard error* (SE) of the mean, which is a measure of how far the sample means deviate from the true population mean on average from sample to sample. SE is the standard deviation for the theoretical *sampling distribution* containing all the possible means from all the different samples. According to the central limit theorem (see section 4.2), the sampling distribution is nearly normal and centered at the population mean. The equation for SE is: $s/\sqrt{n}$, i.e. the sample standard deviation divided by the square root of the sample size. This information helps us to define the confidence intervals (CI) for the mean, i.e. the interval within which the true population mean most likely lies.

Let us calculate an example in which we want to be 95% sure about the exact interval within which the true population mean lies. For the upper boundary we add SE multiplied by 1.96 to the sample mean. For the lower boundary we subtract the same value from the sample mean, i.e: $CI = \pm(1.96 \times SE)$. Where does the multiplier "1.96" come from? This value is based on the idea that the sampling distribution is symmetrical, and the value describes the critical limit at both tails of a *standardized normal distribution*, telling us that there are no more than 2.5% of observations above/below these values. Therefore, the multiplier times SE gives us the limits for mean values that 95% of any sample we could possibly take from the population would yield.

The SE for our sample mean is $913/\sqrt{641} \approx 36$. Let us calculate the 95% CI for our sample mean: $2782 \pm (1.96 \times 36) \approx \pm 80.7$. This allows us to say

Figure 5.7: A plot of mean salaries for males and females with 95% confidence intervals.



that in 95 of 100 samples the real population mean value lies between 2711 and 2853 euros, and yet we haven't done more than one sample! The 95% confidence intervals for the salaries of males and females (these groups have different means, standard deviations, and sample sizes, and hence different standard errors) are depicted in the figure 5.7.

In figure 5.7 the T-bars show the ranges within which the population means for both males and females will fall. The bars show the observed sample means. An important observation is that the error bars are not overlapping, i.e. they do not cover same range of values. This observation makes us more confident in saying that the population means of males and females are different from each other and that the difference is not due to random deviations between different samples.

5.2.2 Testing for a difference between means

In the theories and models section 2.2 we briefly touched on the topic of hypothesis testing, when making the distinction between confirmatory and exploratory research. Statistical testing is a method by which hypothesis testing is conducted. Usually we assume the null hypothesis H_0 and the alternative hypothesis H_1. The alternative hypothesis H_1 is what we are interested in. Hypotheses are the research questions that take the form of a "claim". The null hypothesis is what we test and try to falsify, i.e. to refute by empirical evidence, that is, the data. The statistical test gives us the assurance for saying that the null hypothesis is probably not true, but it also gives us the probability for being wrong (i.e. "risk") when making such a claim. When we have a reason to suspect the claim of the null hypothesis, then we say that we have found evidence for the alternative hypothesis (if not, however, "proved" it to be true). Most typically, H_0 takes the following form: "there is no difference between A and B" or "A and B are equal". In turn, H_1 can be expressed in the following way: "A is bigger/smaller than B" or "A is not zero in the population".

Let us perform a test by which we can be sure that the mean salaries of males and females are different from each other, as the figure 5.7 suggested. We begin by setting the null hypothesis H_0 : *there is no difference between the mean salaries of males and females.* The alternative hypothesis H_1 would hence be *mean salaries differ from each other.* Since we are not testing whether one or the other group has bigger or lower salary, but simply if they differ from each other, the test is called a *two-tailed* test. The two "tails" refer to the 2.5% tails of the sampling distribution. If we test whether one group's mean salary is bigger or smaller than the other one's, we call it a *one-tailed* test, in which case we would have one 5% tail. The logic of hypothesis testing is illustrated in the figure 5.8, in which there is a sampling distribution with a one and two-tailed tests and regions for rejection and acceptance of the null hypothesis.

To compare two means from independent samples we can use a statistical test called a *t-test.* The logic is the same as above when we calculated confidence intervals for the sample mean, but this time we use a different formula for calculating how probable it would be to observe a particular difference between the means of the two samples (males and females) given the null hypothesis is true. In our sample the difference happens to be 529.9 euros. The results of the t-test can be seen in the table 5.2.

The test starts from the assumption made in H_0 that there is no difference

Figure 5.8: A sampling distribution with regions for acceptance and rejection of the null hypothesis.

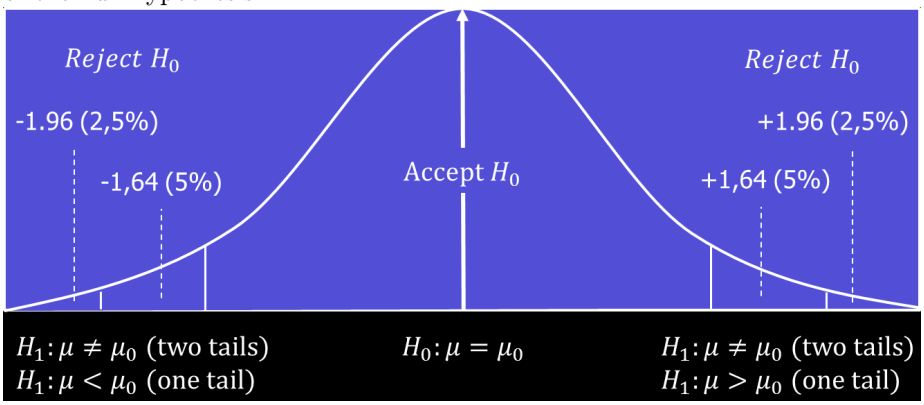


Table 5.2: t-test statistics.

Independent Samples Test

	t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
salary	7,670	639	,000	529,911	69,093	394,235	665,587

between the salaries, and tests how likely it would be to get 529.9 as the difference given the null hypothesis. It appears that given the sample sizes and standard deviations, it would be nearly impossible to observe such a difference by chance. This is indicated by the t-test value which is 7.670, being way up from the 95% critical value. According to the p-value ("Sig. (2-tailed)" column), less than 0.001% of the all possible samples could generate such a difference. The last columns of the table gives the 95% confidence intervals of the difference, indicating that out of 100 samples 95 will yield a difference between 394.2 and 665.6. Since almost every sample, if properly conducted, would yield such a difference, we have a very good reason to suspect the null hypothesis and hence get support for the alternative hypothesis.

It is good to note that the t-test and the comparison of means is just one type of statistical test. There are many possibilities available, of which the most important are the t-test, F-test and χ^2 ("chi-sqare") -test. For example, when doing a contingency table, the test used for examining the statistical significance of the observed differences is the chi-square test. Tests may look different and be calculated differently but the basic idea behind statistical testing is as presented

here. The Decision Tree for Statistics is a helpful aid when finding out what statistical test should be used given the data and measurement levels of the variables.

5.3 Correlation and regression

In the Descriptive statistics section 5.1 we used a scatter plot to draw two continuous variables, age and salary, against each other. On the basis of the picture we were not able to determine if there was any association between the variables. For studying the linear relationship between two continuous variables a measure called the *Pearson product-moment correlation coefficient* ("correlation" in short) can be used. It is a measure of a linear relationship, not a curvilinear one. It does not take a stand on any kind of causality (such as X is a cause of Y), but describes the strength of the association.

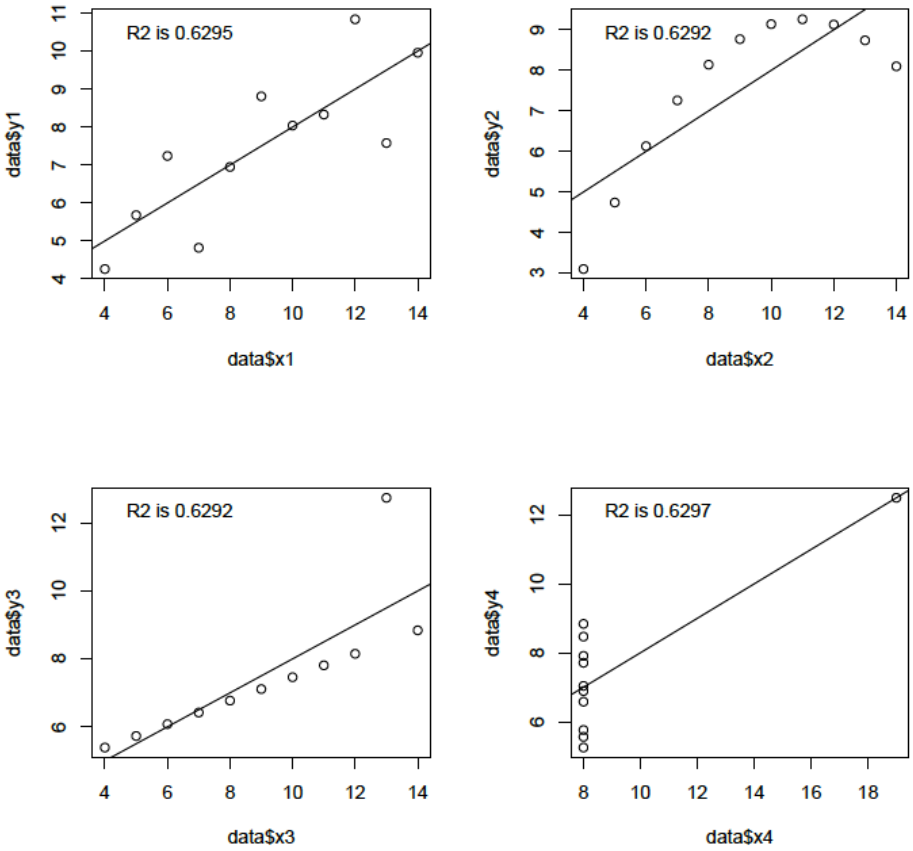
The symbol of the Pearson correlation is ρ or r and in reality it is a standardized measure of *covariance* ("co-variation"). The standardized values can vary between -1 and +1, where 1 indicates perfect positive (linear) relationships, -1 a perfect negative (liner) relationship, and 0 stands for no correlation at all. For the correlation to be considered "weak" or "strong" depends on the study and data, but usually more than $\pm$ 0.3 to 0.5 is expected. For example, correlation between age and salary in our example is 0.124, which is, although statistically significant ($p<.05$), very weak.

The correlation coefficient is easily calculated by any statistical package, but the results are not meaningful unless there is a linear relationship between the variables. The distributions of the variables should also be approximately symmetrical for correlation to be meaningful. For example, the presence of outliers would severely harm the results of a correlation coefficient as seen in the classical example by Anscombe[2] in figure 5.9.

It is good to note how misleading correlation coefficients (or any other single statistics) can be. Therefore, graphical explorations are necessary. In the Anscombe's quartet (fig 5.9) the first plot shows the correlation as it should be, the second shows a curvilinear relationship, the third shows the effect of an outlier, and the fourth shows the effect of an extreme outlier. If the data is biased and the outliers are a necessary part of the data (not random errors), the Pearson correlation should not be used. In such cases, *Spearman's rank*

[2]Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27(1), 17-21. Retrieved from http://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf

Figure 5.9: Anscombe's quartet. Correlation coefficients, means and standard deviations of X's and Y's and regression lines are almost identical although the data sets are considerably different.



correlation can be used, which is a similar measure to the Pearson correlation, but is not influenced by the effect of outliers.

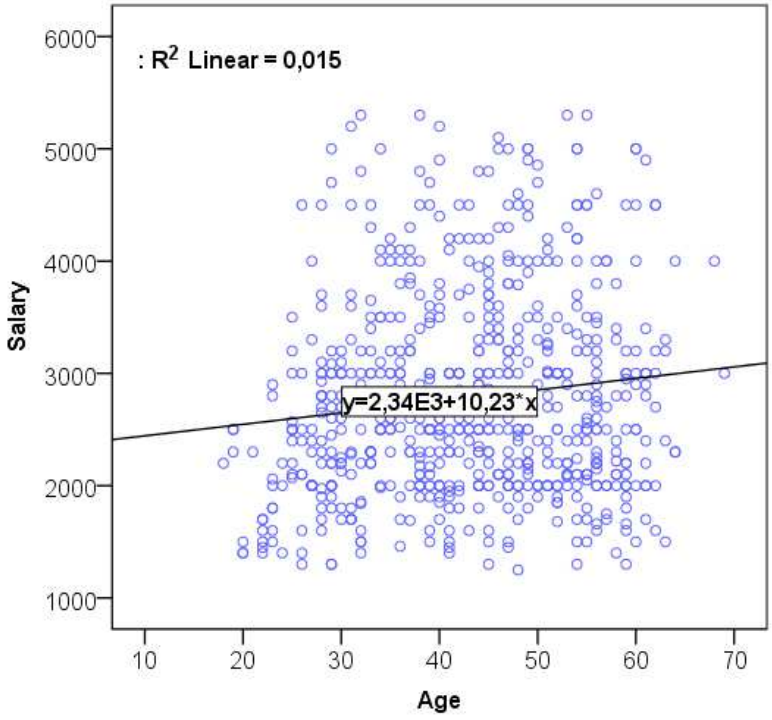
5.3.1 Linear regression

Once we have familiarized ourselves with the basic idea of correlation, it is time to move on to linear regression, which is a technique for modeling the relationship between two or more variables. The scatterplot in the figure 5.10 depicts the regression line that passes through the dots, representing the observations. This regression line is similar to a linear polynomial function that can be represented as a linear equation having a constant "a" (the intersection with Y-axis), the slope "b" (coefficient for defining how steep the line is), and the variables x and y, whose relationship the function describes. This type of equation for a linear line takes the following form: $y = a + bx$, and can be thought of as what goes in as x comes out as y. It is important to note that theoretically speaking we are not dealing with associations any more, but making a causal claim that from x follows y. When x changes, y changes in a way specified by the right side of the equation. In the language of regression analysis, x is called the *independent variable* (the one that is not influenced; also "predictor") and y is called the *dependent variable* (the one that is dependent on x). The symbol for constant is β_0 ("beta") and for coefficient(s) β_x.

If you take a look at the plot in figure 5.10, age and salary are depicted as a linear line, and the equation describing the position and slope of the line is also labeled. In the equation y stands for "salary", and x for "age". The numbers 2340 and 10.23 are estimates for β_0 and β_1. The R^2 in the top left of the plot is the *coefficient of determination*, which describes how well the line – the linear model – fits the data, i.e. how well the line accounts for the variation of the observations. R^2 can be interpreted as a percentage. In this instance the model captures no more than 1.5% of the variation.

The equation can be written in the following way: $y_{salary} = 2340 + 10.23 \times x_{age}$. The constant β_0 is 2340 and the coefficient β_1 is 10.23. How to interpret the model? First, the coefficient is positive, so age has a positive effect on salary. Secondly, the constant (or intercept) is 2340. So if age is 0, the salary would be 2340 when β_1 cancels out. The interesting part, the effect of the age on salary (read: the effect of x on y) is described by the coefficient β_1. It is interpreted in the following way: when x increases by 1 unit (which is 1 year in this case), y will increase by 10.23. So each time we get one year older, our salary increases

Figure 5.10: A plot showing the regression line and equation for age and salary.



by 10.23 euros according to the model.

The idea of a regression model is to predict the values of y on the basis of the values of one or more x variables. Let's try to predict the salary for a 40-year-old person. We start by inserting 40 in place of x so that the right of the equation will be 2340+10.23*40, which yields 2749. This a very simple linear model. Considering the low coefficient of determination (98.5% was not explained), the next task of a researcher is to find better predictors, i.e. x-variables that can predict changes in y. The motive for using regression analysis is that it allows us to add as many x variables into the equation as needed. The regression analysis will account for them all together, considering their simultaneous effect on y. The motive for adding new predictors to the equation comes both from theoretical and statistical reasons, i.e. we know that x influences y under certain conditions and R^2 increases accordingly. Linear regression analysis also has some constraints, however. Most importantly, the dependent variable need to be continuous. Independent variable(s) need to be discrete/continuous or dichotomous. All variables need to be more or less normally distributed, and independent variables should not correlate strongly together (a condition called *multicollinearity*).

Let's add two new variables into the regression equation. We introduce a new variable called "education", which is dichotomous and hence can take two values: "1" standing for "person has a university/college degree" and "0" for "person does not have a university/college degree" (i.e. they have a degree lower than that or no degree at all). We are interested in testing the effect of a university/college degree on the predicted salary. The second new variable is "gender", testing the effect of "being female" (=1) on the predicted salary. The regression equation is now $y_{salary} = \beta_0 + \beta_1 x_{1(age)} + \beta_2 x_{2(education)} + \beta_3 x_{3(gender)}$. The output of the regression analysis is in the table 5.3.

From the table 5.3 we can see that the effect of age is 16.787, meaning that each year yields this much extra to the salary. The constant is now 2117.362, meaning that this is the value when both age, education, and gender are zero. The effect of education is that if a person has a college/university degree, 766.969 will be multiplied by 1, and hence added to the salary. The same goes if a person is female, but this time we subtract 592.928 from the salary. Let's make the same prediction as above for a 40-year-old female with a college/university degree. The salary is now 2117.362 + 16.787*40 + 766.969*1 - 592.928*1 = 2962.883. Which one of the predictors is stronger? Beta values in the table are standardized values for B's, making it possible to compare their effects on y. It

Table 5.3: Regression coefficients.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2117,362	133,951		15,807	,000
	age	16,787	2,865	,204	5,859	,000
	high_edu	766,969	68,264	,391	11,235	,000
	gender	-592,928	62,786	-,325	-9,444	,000

a. Dependent Variable: salary

Table 5.4: Model summary table.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,502[a]	,252	,248	791,659

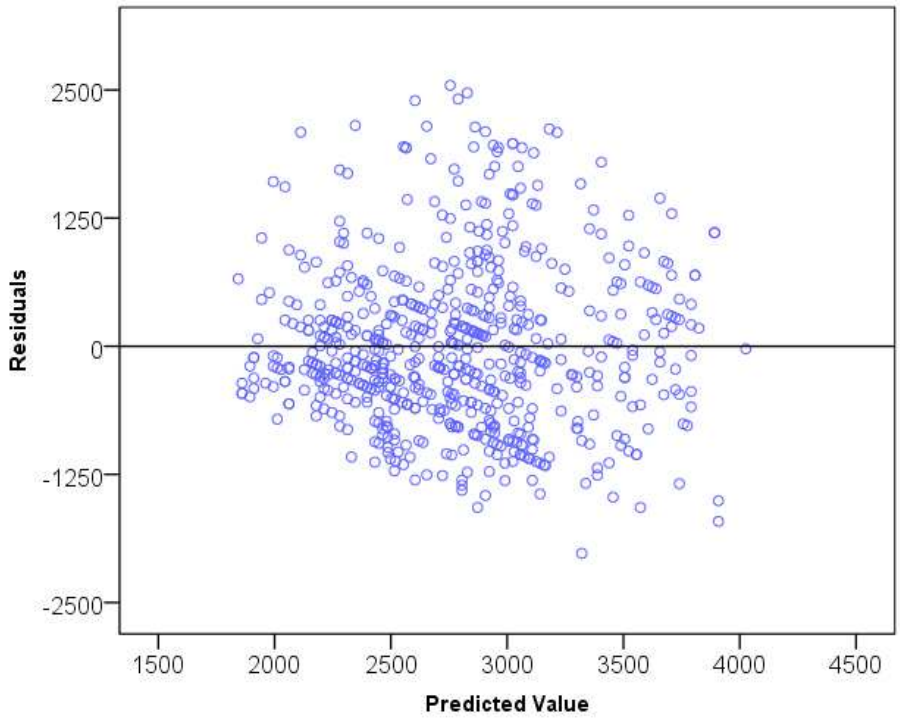
seems to be that education is the strongest one, having almost twice as big an effect as age. t-tests at the right end of the table show whether the predictors are statistically significant in the model (p-values should be <0.05).

From the other output table 5.4 we can also see that R^2 increased from 1.5% to 24.8% ("adjusted R^2" accounts for the error caused by increasing the number of predictors), so it is a big change in the goodness of fit of the model. The standard error of estimate is 791.659, and it is a measure of error in the model. It can be interpreted as the average deviation of the predicted values (salaries given by the model) from the observed values (the observed salaries in the data).

The difference between observed and predicted values is called *residual variation* in regression analysis and it can be marked with e for "error": $y = \beta_0 + \beta_1 x_1 + e$. This means that y is equal to the model plus the rest of the variation of y that is *not* captured by the model. If we modify the equation slightly, we can say that $e = y - (\beta_0 + \beta_1 x_1)$. Technically speaking, the goal of a regression analysis is to minimize the amount of errors. This happens usually by the procedure called *least square estimation*, which is done by finding such values for β_0 and β_1 that the sum of the squared differences between y and $(\beta_0 + \beta_1 x_1)$ become as small as possible.

Finally, a researcher needs to assess how well the model actually fits to the

Figure 5.11: A residual plot should show a homogenous cloud around the center line.



data. This is typically done by examining the residual variance, i.e. the part of the variation that was not explained by the model. For this purpose a residual plot can be drawn (figure 5.11). If the residual variance is normally distributed and varies in an unsystematic manner around the predicted values, we can state that the model is quite a good predictor of the values for y.