Statistics

Wikibooks.org

April 20, 2012

This PDF was generated by a program written by Dirk Hünniger, which is freely available under an open source license from HTTP://DE.WIKIBOOKS.ORG/WIKI/BENUTZER:DIRK_HUENNIGER/WB2PDF.

Contents

1 Introduction

1.1 What is Statistics

> *Your company has created a new drug that may cure arthritis. How would you conduct a test to confirm the drug's effectiveness?*
> *The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you* not*look for?*
> *You and a friend are at a baseball game, and out of the blue he offers you a bet that neither team will hit a home run in that game. Should you take the bet?*
> *You want to conduct a poll on whether your school should use its funding to build a new athletic complex or a new library. How many people do you have to poll? How do you ensure that your poll is free of bias? How do you interpret your results?*
> *A widget maker in your factory that normally breaks 4 widgets for every 100 it produces has recently started breaking 5 widgets for every 100. When is it time to buy a new widget maker? (And just what is a widget, anyway?)*

These are some of the many real-world examples that require the use of statistics.

1.1.1 General Definition

Statistics, in short, is the study of DATA[1]. It includes **descriptive statistics** (the study of methods and tools for collecting data, and mathematical models to describe and interpret data) and **inferential statistics** (the systems and techniques for making probability-based decisions and accurate predictions based on incomplete (sample) data).

1.1.2 Etymology

As its name implies, statistics has its roots in the idea of "the state of things". The word itself comes from the ancient Latin term *statisticum collegium*, meaning "a lecture on the state of affairs". Eventually, this evolved into the Italian word *statista*, meaning "statesman", and the German word *Statistik*, meaning "collection of data involving the State". Gradually, the term came to be used to describe the collection of any sort of data.

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/DATA

1.1.3 Statistics as a subset of mathematics

As one would expect, statistics is largely grounded in mathematics, and the study of statistics has lent itself to many major concepts in mathematics: probability, distributions, samples and populations, the bell curve, estimation, and data analysis.

1.1.4 Up ahead

Up ahead, we will learn about subjects in modern statistics and some practical applications of statistics. We will also lay out some of the background mathematical concepts required to begin studying statistics.

1.2 Subjects in Modern Statistics

A remarkable amount of today's modern statistics comes from the original work of R.A. FISHER[2] in the early 20th Century. Although there are a dizzying number of minor disciplines in the field, there are some basic, fundamental studies.

The beginning student of statistics will be more interested in one topic or another depending on his or her outside interest. The following is a list of some of the primary branches of statistics.

1.2.1 Probability Theory and Mathematical Statistics

Those of us who are purists and philosophers may be interested in the intersection between pure mathematics and the messy realities of the world. A rigorous study of probability—especially the probability distributions and the distribution of errors—can provide an understanding of where all these statistical procedures and equations come from. Although this sort of rigor is likely to get in the way of a psychologist (for example) learning and using statistics effectively, it is important if one wants to do serious (i.e. graduate-level) work in the field.

That being said, there is good reason for all students to have a fundamental understanding of where all these "statistical techniques and equations" are coming from! We're always more adept at using a tool if we can understand *why* we're using that tool. The challenge is getting these important ideas to the non-mathematician without the student's eyes glazing over. One can take this argument a step further to claim that a vast number of students will never actually use a t-test—he or she will never plug those numbers into a calculator and churn through some esoteric equations—but by having a fundamental understanding of such a test, he or she will be able to understand (and question) the results of someone else's findings.

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/RONALD%20FISHER

1.2.2 Design of Experiments

One of the most neglected aspects of statistics—and maybe the single greatest reason that Statisticians drink—is Experimental Design. So often a scientist will bring the results of an important experiment to a statistician and ask for help analyzing results only to find that a flaw in the experimental design rendered the results useless. So often we statisticians have researchers come to us hoping that we will somehow magically "rescue" their experiments.

A friend provided me with a classic example of this. In his psychology class he was required to conduct an experiment and summarize its results. He decided to study whether music had an impact on problem solving. He had a large number of subjects (myself included) solve a puzzle first in silence, then while listening to classical music and finally listening to rock and roll, and finally in silence. He measured how long it would take to complete each of the tasks and then summarized the results.

What my friend failed to consider was that the results were highly impacted by a *learning effect* he hadn't considered. The first puzzle always took longer because the subjects were first learning how to work the puzzle. By the third try (when subjected to rock and roll) the subjects were much more adept at solving the puzzle, thus the results of the experiment would seem to suggest that people were much better at solving problems while listening to rock and roll!

The simple act of randomizing the order of the tests would have isolated the "learning effect" and in fact, a well-designed experiment would have allowed him to measure both the effects of each type of music *and* the effect of learning. Instead, his results were meaningless. A careful experimental design can help preserve the results of an experiment, and in fact some designs can save huge amounts of time and money, maximize the results of an experiment, and sometimes yield additional information the researcher had never even considered!

1.2.3 Sampling

Similar to the Design of Experiments, the study of sampling allows us to find a most effective statistical design that will optimize the amount of information we can collect while minimizing the level of effort. Sampling is very different from experimental design however. In a laboratory we can design an experiment and control it from start to finish. But often we want to study something outside of the laboratory, over which we have much less control.

If we wanted to measure the population of some harmful beetle and its effect on trees, we would be forced to travel into some forest land and make observations, for example: measuring the population of the beetles in different locations, noting which trees they were infesting, measuring the health and size of these trees, etc.

Sampling design gets involved in questions like "How many measurements do I have to take?" or "How do I select the locations from which I take my measurements?" Without planning for these issues, researchers might spend a lot of time and money only to discover that they really have to sample ten times as many points to get meaningful results or that some of their sample points were in some landscape (like a marsh) where the beetles thrived more or the trees grew better.

1.2.4 Modern Regression

Regression models relate variables to each other in a linear fashion. For example, if you recorded the heights and weights of several people and plotted them against each other, you would find that as height increases, weight tends to increase too. You would probably also see that a straight line through the data is about as good a way of approximating the relationship as you will be able to find, though there will be some variability about the line. Such linear models are possibly the most important tool available to statisticians. They have a long history and many of the more detailed theoretical aspects were discovered in the 1970s. The usual method for fitting such models is by "least squares" estimation, though other methods are available and are often more appropriate, especially when the data are not normally distributed.

What happens, though, if the relationship is not a straight line? How can a curve be fit to the data? There are many answers to this question. One simple solution is to fit a quadratic relationship, but in practice such a curve is often not flexible enough. Also, what if you have many variables and relationships between them are dissimilar and complicated?

Modern regression methods aim at addressing these problems. Methods such as generalized additive models, projection pursuit regression, neural networks and boosting allow for very general relationships between explanatory variables and response variables, and modern computing power makes these methods a practical option for many applications

1.2.5 Classification

Some things are different from others. How? That is, how are objects classified into their respective groups? Consider a bank that is hoping to lend money to customers. Some customers who borrow money will be unable or unwilling to pay it back, though most will pay it back as regular repayments. How is the bank to classify customers into these two groups when deciding which ones to lend money to?

The answer to this question no doubt is influenced by many things, including a customer's income, credit history, assets, already existing debt, age and profession. There may be other influential, measurable characteristics that can be used to predict what kind of customer a particular individual is. How should the bank decide which characteristics are important, and how should it combine this information into a rule that tells it whether or not to lend the money?

This is an example of a classification problem, and statistical classification is a large field containing methods such as linear discriminant analysis, classification trees, neural networks and other methods.

1.2.6 Time Series

Many types of research look at data that are gathered over time, where an observation taken today may have some correlation with the observation taken tomorrow. Two prominent examples of this are the fields of finance (the stock market) and atmospheric science.

We've all seen those line graphs of stock prices as they meander up and down over time. Investors are interested in predicting which stocks are likely to keep climbing (i.e. when to buy) and when a stock in their portfolio is falling. It is easy to be misled by a sudden jolt of good news or a simple "market correction" into inferring—incorrectly—that one or the other is taking place!

In meteorology scientists are concerned with the venerable science of predicting the weather. Whether trying to predict if tomorrow will be sunny or determining whether we are experiencing true climate changes (i.e. global warming) it is important to analyze weather data over time.

1.2.7 Survival Analysis

Suppose that a pharmaceutical company is studying a new drug which it is hoped will cause people to live longer (whether by curing them of cancer, reducing their blood pressure or cholesterol and thereby their risk of heart disease, or by some other mechanism). The company will recruit patients into a clinical trial, give some patients the drug and others a placebo, and follow them until they have amassed enough data to answer the question of whether, and by how long, the new drug extends life expectancy.

Such data present problems for analysis. Some patients will have died earlier than others, and often some patients will not have died before the clinical trial completes. Clearly, patients who live longer contribute informative data about the ability (or not) of the drug to extend life expectancy. So how should such data be analyzed?

Survival analysis provides answers to this question and gives statisticians the tools necessary to make full use of the available data to correctly interpret the treatment effect.

1.2.8 Categorical Analysis

In laboratories we can measure the weight of fruit that a plant bears, or the temperature of a chemical reaction. These data points are easily measured with a yardstick or a thermometer, but what about the color of a person's eyes or her attitudes regarding the taste of broccoli? Psychologists can't measure someone's anger with a measuring stick, but they can ask their patients if they feel "very angry" or "a little angry" or "indifferent". Entirely different methodologies must be used in statistical analysis from these sorts of experiments. Categorical Analysis is used in a myriad of places, from political polls to analysis of census data to genetics and medicine.

1.2.9 Clinical Trials

In the United States, the FDA[3] requires that pharmaceutical companies undergo rigorous procedures called Clinical Trials[4] and statistical analyses to assure public safety before

[3] http://en.wikipedia.org/wiki/FDA

[4] http://en.wikipedia.org/wiki/Clinical%20Trials

allowing the sale of use of new drugs. In fact, the pharmaceutical industry employs more statisticians than any other business!

1.2.10 Further reading

- Econometric Theory[5]
- Classification[6]

1.3 Why Should I Learn Statistics?

Imagine reading a book for the first few chapters and then becoming able to get a sense of what the ending will be like - this is one of the great reasons to learn statistics. With the appropriate tools and solid grounding in statistics, one can use a limited sample (e.g. read the first five chapters of Pride & Prejudice) to make intelligent and accurate statements about the population (e.g. predict the ending of Pride & Prejudice). This is what knowing statistics and statistical tools can do for you.

In today's information-overloaded age, statistics is one of the most useful subjects anyone can learn. Newspapers are filled with statistical data, and anyone who is ignorant of statistics is at risk of being seriously misled about important real-life decisions such as what to eat, who is leading the polls, how dangerous smoking is, etc. Knowing a little about statistics will help one to make more informed decisions about these and other important questions. Furthermore, statistics are often used by politicians, advertisers, and others to twist the truth for their own gain. For example, a company selling the cat food brand "Cato" (a fictitious name here), may claim quite truthfully in their advertisements that eight out of ten cat owners said that their cats preferred Cato brand cat food to "the other leading brand" cat food. What they may not mention is that the cat owners questioned were those they found in a supermarket buying Cato.

"The best thing about being a statistician is that you get to play in everyone else's backyard." John Tukey, Princeton University[7]

More seriously, those proceeding to higher education will learn that statistics is the most powerful tool available for assessing the significance of experimental data, and for drawing the right conclusions from the vast amounts of data faced by engineers, scientists, sociologists, and other professionals in most spheres of learning. There is no study with scientific, clinical, social, health, environmental or political goals that does not rely on statistical methodologies. The basic reason for that is that variation is ubiquitous in nature and probability[8] and statistics[9] are the fields that allow us to study, understand, model, embrace and interpret variation.

[5] http://en.wikibooks.org/wiki/Econometric%20Theory

[6] http://en.wikibooks.org/wiki/Optimal%20Classification%20

[7] http://en.wikipedia.org/wiki/John%20W.%20Tukey%20

[8] http://en.wikibooks.org/wiki/probability

[9] http://en.wikibooks.org/wiki/statistics

1.3.1 See Also

UCLA Brochure on **Why Study Probability & Statistics**[10]

1.4 What Do I Need to Know to Learn Statistics?

Statistics is a diverse subject and thus the mathematics that are required depend on the kind of statistics we are studying. A strong background in linear algebra[11] is needed for most multivariate statistics, but is not necessary for introductory statistics. A background in Calculus[12] is useful no matter what branch of statistics is being studied, but is not required for most introductory statistics classes.

At a bare minimum the student should have a grasp of basic concepts taught in Algebra[13] and be comfortable with "moving things around" and solving for an unknown. Most of the statistics here will derive from a few basic things that the reader should become acquainted with.

1.4.1 Absolute Value

$$|x| \equiv \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$

If the number is zero or positive, then the absolute value of the number is simply the same number. If the number is negative, then take away the negative sign to get the absolute value.

Examples

- $|42| = 42$
- $|$-5$| = 5$
- $|2.21| = 2.21$

1.4.2 Factorials

A factorial is a calculation that gets used a lot in probability. It is defined only for integers greater-than-or-equal-to zero as:

[10] http://www.stat.ucla.edu/%7Edinov/WhyStudyStatisticsBrochure/WhyStudyStatisticsBrochure.html

[11] http://en.wikibooks.org/wiki/Algebra%23Linear_algebra

[12] http://en.wikibooks.org/wiki/Calculus

[13] http://en.wikibooks.org/wiki/Algebra

$$n! \equiv \begin{cases} n \cdot (n-1)!, & n \geq 1 \\ 1, & n = 0 \end{cases}$$

Examples

In short, this means that:

$$\begin{aligned}
0! &= & 1 & &= 1 \\
1! &= & 1 \cdot 1 & &= 1 \\
2! &= & 2 \cdot 1 & &= 2 \\
3! &= & 3 \cdot 2 \cdot 1 & &= 6 \\
4! &= & 4 \cdot 3 \cdot 2 \cdot 1 & &= 24 \\
5! &= & 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 & &= 120 \\
6! &= & 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 & &= 720
\end{aligned}$$

1.4.3 Summation

The summation (also known as a series) is used more than almost any other technique in statistics. It is a method of representing addition over lots of values without putting + after +. We represent summation using a big uppercase sigma: $\sum$.

Examples

Very often in statistics we will sum a list of related variables:

$$\sum_{i=0}^{n} x_i = x_0 + x_1 + x_2 + \cdots + x_n$$

Here we are adding all the x variables (which will hopefully all have values by the time we calculate this). The expression below the $\sum$ (i=0, in this case) represents the index variable and what its starting value is (i with a starting value of 0) while the number above the $\sum$ represents the number that the variable will increment to (stepping by 1, so i = 0, 1, 2, 3, and then 4). Another example:

$$\sum_{i=1}^{4} 2i = 2(1) + 2(2) + 2(3) + 2(4) = 2 + 4 + 6 + 8 = 20$$

Notice that we would get the same value by moving the 2 outside of the summation (perform the summation and then multiply by 2, rather than multiplying each component of the summation by 2).

Infinite series

There is no reason, of course, that a series has to count on any determined, or even finite value—it can keep going without end. These series are called "infinite series" and sometimes they can even converge to a finite value, eventually becoming equal to that value as the number of items in your series approaches infinity (∞).

Examples

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}, \qquad |r| < 1$$

This example is the famous geometric series[14]. Note both that the series goes to ∞ (infinity, that means it does not stop) and that it is only valid for certain values of the variable *r*. This means that if *r* is between the values of -1 and 1 (-1 < *r* < 1) then the summation will get closer to (i.e., converge on) ^{1} / $_{1\text{-}r}$ the further you take the series out.

1.4.4 Linear Approximation

v / *α*	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
40	0.85070	1.30308	1.68385	2.02108	2.42326	2.70446
50	0.84887	1.29871	1.67591	2.00856	2.40327	2.67779
60	0.84765	1.29582	1.67065	2.00030	2.39012	2.66028
70	0.84679	1.29376	1.66691	1.99444	2.38081	2.64790
80	0.84614	1.29222	1.66412	1.99006	2.37387	2.63869
90	0.84563	1.29103	1.66196	1.98667	2.36850	2.63157
100	0.84523	1.29007	1.66023	1.98397	2.36422	2.62589

Student-t Distribution at various critical values with varying degrees of freedom.

Let us say that you are looking at a table of values, such as the one above. You want to approximate (get a good estimate of) the values at 63, but you do not have those values

[14] http://en.wikipedia.org/wiki/Geometric%20series

on your table. A good solution here is use a linear approximation to get a value which is probably close to the one that you really want, without having to go through all of the trouble of calculating the extra step in the table.

$$f\left(x_i\right) \approx \frac{f\left(x_{\lceil i \rceil}\right) - f\left(x_{\lfloor i \rfloor}\right)}{x_{\lceil i \rceil} - x_{\lfloor i \rfloor}} \cdot \left(x_i - x_{\lfloor i \rfloor}\right) + f\left(x_{\lfloor i \rfloor}\right)$$

This is just the equation for a line applied to the table of data. x_i represents the data point you want to know about, $x_{\lfloor i \rfloor}$ is the known data point *beneath* the one you want to know about, and $x_{\lceil i \rceil}$ is the known data point *above* the one you want to know about.

Examples

Find the value at 63 for the 0.05 column, using the values on the table above.

First we confirm on the above table that we need to approximate the value. If we know it exactly, then there really is no need to approximate it. As it stands this is going to rest on the table somewhere between 60 and 70. Everything else we can get from the table:

$$f(63) \approx \frac{f(70) - f(60)}{70 - 60} \cdot (63 - 60) + f(60) = \frac{1.66691 - 1.67065}{10} \cdot 3 + 1.67065 = 1.669528$$

Using software, we calculate the actual value of $f(63)$ to be 1.669402, a difference of around 0.00013. Close enough for our purposes.

2 Different Types of Data

Data are assignments of values onto observations of events and objects. They can be classified by their coding properties and the characteristics of their domains and their ranges.

2.1 Identifying data type

When a given data set is numerical in nature, it is necessary to carefully distinguish the actual nature of the variable being quantified. Statistical tests are generally specific for the kind of data being handled.

2.1.1 Data on a nominal (or categorical) scale

Identifying the true nature of numerals applied to attributes that are not "measures" is usually straightforward and apparent. Examples in everyday use include road, car, house, book and telephone numbers. A simple test would be to ask if re-assigning the numbers among the set would alter the nature of the collection. If the plates on a car are changed, for example, it still remains the same car.

2.1.2 Data on an Ordinal Scale

An ordinal scale is a scale with ranks. Those ranks only have sense in that they are ordered, that is what makes it ordinal scale. The distance [rank *n*] minus [rank *n-1*] is not guaranteed to be equal to [rank *n-1*] minus [rank *n-2*], but [rank *n*] will be greater than [rank *n-1*] in the same way [rank *n-1*] is greater than [rank *n-2*] for all *n* where [rank *n*], [rank *n-1*], and [rank *n-2*] exist. Ranks of an ordinal scale may be represented by a system with numbers or names and an agreed order.

We can illustrate this with a common example: the Likert scale. Consider five possible responses to a question, perhaps *Our president is a great man*, with answers on this scale

Response:	**Strongly disagree**	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
Code:	1	2	3	4	5

Here the answers are a ranked scale reflected in the choice of numeric code. There is however no sense in which the distance between *Strongly agree* and *Agree* is the same as between *Strongly disagree* and *Disagree.*

Numerical ranked data should be distinguished from measurement data.

2.1.3 Measurement data

Numerical measurements exist in two forms, Meristic and continuous, and may present themselves in three kinds of scale: interval, ratio and circular.

Meristic or **discrete** variables are generally counts and can take on only discrete values. Normally they are represented by natural numbers. The number of plants found in a botanist's quadrant would be an example. (Note that if the edge of the quadrant falls partially over one or more plants, the investigator may choose to include these as halves, but the data will still be meristic as doubling the total will remove any fraction).

Continuous variables are those whose measurement precision is limited only by the investigator and his equipment. The length of a leaf measured by a botanist with a ruler will be less precise than the same measurement taken by micrometer. (Notionally, at least, the leaf could be measured even more precisely using a microscope with a graticule.)

Interval Scale Variables measured on an interval scale have values in which differences are uniform and meaningful but ratios will not be so. An oft quoted example is that of the Celsius scale of temperature. A difference between 5° and 10° is equivalent to a difference between 10° and 15°, but the ratio between 15° and 5° does not imply that the former is three times as warm as the latter.

Ratio Scale Variables on a ratio scale have a meaningful zero point. In keeping with the above example one might cite the Kelvin temperature scale. Because there is an absolute zero, it is true to say that 400°K is twice as warm as 200°K, though one should do so with tongue in cheek. A better day-to-day example would be to say that a 180 kg Sumo wrestler is three times heavier than his 60 kg wife.

Circular Scale When one measures annual dates, clock times and a few other forms of data, a circular scale is in use. It can happen that neither differences nor ratios of such variables are sensible derivatives, and special methods have to be employed for such data.

...... :)

2.2 Primary and Secondary Data

Data can be classified as either **primary** or **secondary**.

2.2.1 Primary Data

Primary data means original data that has been **collected** specially for the purpose in mind. It means when an authorized organization, investigator or an enumerator collects the data for the first time from the original source. Data collected this way is called primary data.

Research where one gathers this kind of data is referred to as ’field research.

For example: your own questionnaire.

2.2.2 Secondary Data

Secondary data is data that has been **collected** for another purpose. When we use Statistical Method with Primary Data from another purpose for our purpose we refer to it as Secondary Data. It means that one purpose’s Primary Data is another purpose’s Secondary Data. Secondary data is data that is being reused. Usually in a different context.

Research where one gathers this kind of data is referred to as ’desk research.

For example: data from a book.

2.2.3 Why Classify Data This Way?

Knowing how the data was collected allows critics of a study to search for bias in how it was conducted. A good study will welcome such scrutiny. Each type has its own weaknesses and strengths. Primary Data is gathered by people who can focus directly on the purpose in mind. This helps ensure that questions are meaningful to the purpose but can introduce bias in those same questions. Secondary data doesn’t have the privilege of this focus but is only susceptible to bias introduced in the choice of what data to reuse. Stated another way, those who gather Primary Data get to write the questions. Those who gather secondary data get to pick the questions.

<< Different Types of Data[1] | Statistics[2] | >> Qualitative and Quantitative[3]

Quantitative and qualitative data are two types of data.

2.3 Qualitative data

Qualitative data is a categorical measurement expressed not in terms of numbers, but rather by means of a natural language description. In statistics, it is often used interchangeably with "categorical" data.

```
For example: favorite color = "yellow"
             height = "tall"
```

[1] Chapter 2 on page 13

[2] http://en.wikibooks.org/wiki/Statistics

[3] Chapter 2.2.3 on page 15

Although we may have categories, the categories may have a structure to them. When there is not a natural ordering of the categories, we call these **nominal** categories. Examples might be gender, race, religion, or sport.

When the categories may be ordered, these are called **ordinal** variables. **Categorical variables** that judge size (small, medium, large, etc.) are ordinal variables. Attitudes (strongly disagree, disagree, neutral, agree, strongly agree) are also ordinal variables, however we may not know which value is the best or worst of these issues. Note that the distance between these categories is not something we can measure.

2.4 Quantitative data

Quantitative data is a numerical measurement expressed not by means of a natural language description, but rather in terms of numbers. However, not all numbers are continuous and measurable. For example, the social security number is a number, but not something that one can add or subtract.

```
For example: favorite color = "450 nm"
             height = "1.8 m"
```

Quantitative data always are associated with a scale measure.

Probably the most common scale type is the ratio-scale. Observations of this type are on a scale that has a meaningful zero value but also have an equidistant measure (i.e., the difference between 10 and 20 is the same as the difference between 100 and 110). For example, a 10 year-old girl is twice as old as a 5 year-old girl. Since you can measure zero years, time is a ratio-scale variable. Money is another common ratio-scale quantitative measure. Observations that you count are usually ratio-scale (e.g., number of widgets).

A more general quantitative measure is the interval scale. Interval scales also have a equidistant measure. However, the doubling principle breaks down in this scale. A temperature of 50 degrees Celsius is not "half as hot" as a temperature of 100, but a difference of 10 degrees indicates the same difference in temperature anywhere along the scale. The Kelvin temperature scale, however, constitutes a ratio scale because on the Kelvin scale zero indicates absolute zero in temperature, the complete absence of heat. So one can say, for example, that 200 degrees Kelvin is twice as hot as 100 degrees Kelvin.

<< Different Types of Data[4] | Statistics[5]

[4] Chapter 2.1.3 on page 14

[5] http://en.wikibooks.org/wiki/Statistics

3 Methods of Data Collection

The main portion of Statistics is the display of summarized data. Data is initially collected from a given source, whether they are experiments, surveys, or observation, and is presented in one of four methods:

Textular Method

The reader acquires information through reading the gathered data.

Tabular Method

Provides a more precise, systematic and orderly presentation of data in rows or columns.

Semi-tabular Method

Uses both textual and tabular methods.

Graphical Method

The utilization of graphs is most effective method of visually presenting statistical results or findings.

3.1 Experiments

Scientists try to identify cause-and-effect relationships because this kind of knowledge is especially powerful, for example, drug A cures disease B. Various methods exist for detecting cause-and-effect relationships. An experiment is a method that most clearly shows cause-and-effect because it isolates and manipulates a single variable, in order to clearly show its effect. Experiments almost always have two distinct variables: First, an independent variable (IV) is manipulated by an experimenter to exist in at least two levels (usually "none" and "some"). Then the experimenter measures the second variable, the dependent variable (DV).

A simple example:

Suppose the experimental hypothesis that concerns the scientist is that reading a Wiki will enhance knowledge. Notice that the hypothesis is really an attempt to state a causal relationship like, "if you read a Wiki, then you will have enhanced knowledge." The antecedent condition (reading a Wiki) causes the consequent condition (enhanced knowledge). Antecedent conditions are always IVs and consequent conditions are always DVs in experiments. So the experimenter would produce two levels of Wiki reading (none and some, for example) and record knowledge. If the subjects who got no Wiki exposure had less knowledge than those who were exposed to Wikis, it follows that the difference is caused by the IV.

So, the reason scientists utilize experiments is that it is the only way to determine causal relationships between variables. Experiments tend to be artificial because they try to make both groups identical with the single exception of the levels of the independent variable.

3.2 Sample Surveys

Sample surveys involve the selection and study of a sample of items from a population. A sample is just a set of members chosen from a population, but not the whole population. A survey of a whole population is called a **census**.

A sample from a population may not give accurate results but it helps in decision making.

3.2.1 Examples

Examples of sample surveys:

- Phoning the fifth person on every page of the local phonebook and asking them how long they have lived in the area. (Systematic Sample)
- Dropping a quad. in five different places on a field and counting the number of wild flowers inside the quad. (Cluster Sample)
- Selecting sub-populations in proportion to their incidence in the overall population. For instance, a researcher may have reason to select a sample consisting 30% females and 70% males in a population with those same gender proportions. (Stratified Sample)
- Selecting several cities in a country, several neighbourhoods in those cities and several streets in those neighbourhoods to recruit participants for a survey (Multi-stage sample)

The term random sample is used for a sample in which every item in the population is equally likely to be selected.

3.2.2 Bias

While sampling is a more cost effective method of determining a result, small samples or samples that depend on a certain selection method will result in a bias within the results.

The following are common sources of bias:

- Sampling bias or statistical bias, where some individuals are more likely to be selected than others (such as if you give equal chance of cities being selected rather than weighting them by size)
- Systemic bias, where external influences try to affect the outcome (e.g. funding organizations wanting to have a specific result)

3.3 Observational Studies

The most primitive method of understanding the laws of nature utilizes observational studies. Basically, a researcher goes out into the world and looks for variables that are associated with one another. Notice that, unlike experiments, observational research had no Independent Variables --- nothing is manipulated by the experimenter. Rather, observations (also called correlations, after the statistical techniques used to analyze the data) have the equivalent of two Dependent Variables.

Some of the foundations of modern scientific thought are based on observational research. Charles Darwin, for example, based his explanation of evolution entirely on observations he made. Case studies, where individuals are observed and questioned to determine possible causes of problems, are a form of observational research that continues to be popular today. In fact, every time you see a physician he or she is performing observational science.

There is a problem in observational science though --- it cannot ever identify causal relationships because even though two variables are related both might be caused by a third, unseen, variable. Since the underlying laws of nature are assumed to be causal laws, observational findings are generally regarded as less compelling than experimental findings.

The key way to identify experimental studies is that they involve an intervention such as the administration of a drug to one group of patients and a placebo to another group. Observational studies only collect data and make comparisons.

Medicine is an intensively studied discipline, and not all phenomenon can be studies by experimentation due to obvious ethical or logistical restrictions. Types of studies include:

Case series: These are purely observational, consisting of reports of a series of similar medical cases. For example, a series of patients might be reported to suffer from bone abnormalities as well as immunodeficiencies. This association may not be significant, occurring purely by chance. On the other hand, the association may point to a mutation in common pathway affecting both the skeletal system and the immune system.

Case-Control: This involves an observation of a disease state, compared to normal healthy controls. For example, patients with lung cancer could be compared with their otherwise healthy neighbors. Using neighbors limits bias introduced by demographic variation. The cancer patients and their neighbors (the control) might be asked about their exposure history (did they work in an industrial setting), or other risk factors such as smoking. Another example of a case-control study is the testing of a diagnostic procedure against the gold standard. The gold standard represents the control, while the new diagnostic procedure is the "case." This might seem to qualify as an "intervention" and thus an experiment.

Cross-sectional: Involves many variables collected all at the same time. Used in epidemiology to estimate prevalence, or conduct other surveys.

Cohort: A group of subjects followed over time, prospectively. Framingham study is classic example. By observing exposure and then tracking outcomes, cause and effect can be better isolated. However this type of study cannot conclusively isolate a cause and effect relationship.

Historic Cohort: This is the same as a cohort except that researchers use an historic medical record to track patients and outcomes.

4 Data Analysis

Data analysis is one of the more important stages in our research. Without performing exploratory analyses of our data, we set ourselves up for mistakes and loss of time.

Generally speaking, our goal here is to be able to "visualize" the data and get a sense of their values. We plot histograms and compute summary statistics to observe the trends and the distribution of our data.

4.1 Data Cleaning

’Cleaning’ refers to the process of removing invalid data points from a dataset.

Many statistical analyses try to find a pattern in a data series, based on a hypothesis or assumption about the nature of the data. ’Cleaning’ is the process of removing those data points which are either (a) Obviously disconnected with the effect or assumption which we are trying to isolate, due to some other factor which applies only to those particular data points. (b) Obviously erroneous, i.e. some external error is reflected in that particular data point, either due to a mistake during data collection, reporting etc.

In the process we ignore these particular data points, and conduct our analysis on the remaining data.

’Cleaning’ frequently involves human judgement to decide which points are valid and which are not, and there is a chance of valid data points caused by some effect not sufficiently accounted for in the hypothesis/assumption behind the analytical method applied.

The points to be cleaned are generally extreme outliers. ’Outliers’ are those points which stand out for not following a pattern which is generally visible in the data. One way of detecting outliers is to plot the data points (if possible) and visually inspect the resultant plot for points which lie far outside the general distribution. Another way is to run the analysis on the entire dataset, and then eliminating those points which do not meet mathematical ’control limits’ for variability from a trend, and then repeating the analysis on the remaining data.

Cleaning may also be done judgementally, for example in a sales forecast by ignoring historical data from an area/unit which has a tendency to misreport sales figures. To take another example, in a double blind medical test a doctor may disregard the results of a volunteer whom the doctor happens to know in a non-professional context.

’Cleaning’ may also sometimes be used to refer to various other judgemental/mathematical methods of validating data and removing suspect data.

The importance of having clean and reliable data in any statistical analysis cannot be stressed enough. Often, in real-world applications the analyst may get mesmerised by the complexity or beauty of the method being applied, while the data itself may be unreliable and lead to results which suggest courses of action without a sound basis. A good statistician/researcher (personal opinion) spends 90% of his/her time on collecting and cleaning data, and developing hypothesis which cover as many external explainable factors as possible, and only 10% on the actual mathematical manipulation of the data and deriving results.

5 Summary Statistics

5.1 Summary Statistics

The most simple example of statistics "in practice" is in the generation of summary statistics. Let us consider the example where we are interested in the weight of eighth graders in a school. (Maybe we're looking at the growing epidemic of child obesity in America!) Our school has 200 eighth graders, so we gather all their weights. What we have are 200 positive real numbers.

If an administrator asked you what the weight was of this eighth grade class, you wouldn't grab your list and start reading off all the individual weights; it's just too much information. That same administrator wouldn't learn anything except that she shouldn't ask you any questions in the future! What you want to do is to distill the information — these 200 numbers — into something concise.

What might we express about these 200 numbers that would be of interest? The most obvious thing to do is to calculate the average or *mean* value so we know how much the "typical eighth grader" in the school weighs. It would also be useful to express how much this number varies; after all, eighth graders come in a *wide variety* of shapes and sizes! In reality, we can probably reduce this set of 200 weights into at most four or five numbers that give us a firm comprehension of the data set.

5.2 Averages

An average is simply a number that is representative of data. More particularly, it is a measure of central tendency. There are several types of average. Averages are useful for comparing data, especially when sets of different size are being compared. It acts as a representative figure of the whole set of data.

Perhaps the simplest and commonly used average the **arithmetic mean** or more simply MEAN[1] which is explained in the next section.

Other common types of average are the **median, the mode, the geometric mean,** and the **harmonic mean,** each of which may be the most appropriate one to use under different circumstances.

STATISTICS[2] | SUMMARY STATISTICS[3] | >> MEAN, MEDIAN AND MODE[4]

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ASUMMARY%2FAVERAGES%2FMEAN%23MEAN
[2] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS
[3] Chapter 5 on page 23
[4] Chapter 5.2 on page 23

5.2.1 Mean, Median and Mode

Mean

The mean, or more precisely the arithmetic mean, is simply the arithmetic average of a group of numbers (or **data set**) and is shown using -bar symbol $\bar{}$. So the mean of the variable x is $\bar{x}$, pronounced "x-bar". It is calculated by adding up all of the values in a data set and dividing by the number of values in that data set **:**$\bar{x} = \frac{\sum x}{n}$**.**For example, take the following set of data: {1,2,3,4,5}. The mean of this data would be:

$$\bar{x} = \frac{\sum x}{n} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Here is a more complicated data set: {10,14,86,2,68,99,1}. The mean would be calculated like this:

$$\bar{x} = \frac{\sum x}{n} = \frac{10+14+86+2+68+99+1}{7} = \frac{280}{7} = 40$$

Median

The median is the "middle value" in a set. That is, the median is the number in the center of a data set that has been ordered sequentially.

For example, let's look at the data in our second data set from above: {10,14,86,2,68,99,1}. What is its median?

- First, we sort our data set sequentially: {1,2,10,14,68,85,99}
- Next, we determine the total number of points in our data set (in this case, 7.)
- Finally, we determine the central position of or data set (in this case, the 4th position), and the number in the central position is our median - {1,2,10,**14**,68,85,99}, making 14 our median.

Helpful Hint:
An easy way to determine the central position or positions for any ordered set is to take the total number of points, add 1, and then divide by 2. If the number you get is a whole number, then that is the central position. If the number you get is a fraction, take the two whole numbers on either side.

Because our data set had an odd number of points, determining the central position was easy - it will have the same number of points before it as after it. But what if our data set has an even number of points?

Let's take the same data set, but add a new number to it: {1,2,10,14,68,85,99,*100*} What is the median of this set?

When you have an even number of points, you must determine the *two* central positions of the data set. (See side box for instructions.) So for a set of 8 numbers, we get (8 + 1) / 2 = 9 / 2 = 4 1/2, which has 4 and 5 on either side.

Looking at our data set, we see that the 4th and 5th numbers are 14 and 68. From there, we return to our trusty friend the mean to determine the median. (14 + 68) / 2 = 82 / 2 = **41**. find the median of 2 , 4 , 6, 8 => firstly we must count the numbers to determine its odd or even as we see it is even so we can write : M=4+6/2=10/2=5 5 is the median of above sequentiall numbers.

Mode

The mode is the most common or "most frequent" value in a data set. Example: the mode of the following data set (1, 2, **5**, **5**, 6, 3) is 5 since it **appears twice**. This is the most common value of the data set. Data sets having one mode are said to be **unimodal**, with two are said to be **bimodal** and with more than two are said to be **multimodal** . An example of a unimodal dataset is {1, 2, 3, **4**, **4**, **4**, **5**, 6, 7, 8, 8, 9}. The mode for this data set is 4. An example of a bimodal data set is {1, **2**, **2**, **3**, **3**}. This is because both 2 and 3 are modes. **Please note**: If all points in a data set occur with equal frequency, it is equally accurate to describe the data set as having many modes or no mode.

Midrange

The midrange is the arithmetic mean strictly between the minimum and the maximum value in a data set.

Relationship of the Mean, Median, and Mode

The relationship of the mean, median, and mode to each other can provide some information about the relative shape of the data distribution. If the mean, median, and mode are approximately equal to each other, the distribution can be assumed to be approximately symmetrical. If the mean > median > mode, the distribution will be skewed to the left or positively skewed. If the mean < median < mode, the distribution will be skewed to the right or negatively skewed.

5.2.2 Questions

1. There is an old joke that states: "Using median size as a reference it's perfectly possible to fit four ping-pong balls and two blue whales in a rowboat." Explain why this statement is true.

STATISTICS[5] | MEAN[6]

[5] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS

[6] Chapter 5.2 on page 23

5.2.3 Geometric Mean

The Geometric Mean is calculated by taking the *n*th root of the product of a set of data.

$$\tilde{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

For example, if the set of data was:

1,2,3,4,5

The geometric mean would be calculated:

$$\sqrt[5]{1 \times 2 \times 3 \times 4 \times 5} = \sqrt[5]{120} = 2.61$$

Of course, with large n this can be difficult to calculate. Taking advantage of two properties of the logarithm:

$$\log(a \cdot b) = \log(a) + \log(b)$$

$$\log(a^n) = n \cdot \log(a)$$

We find that by taking the logarithmic transformation of the geometric mean, we get:

$$\log\left(\sqrt[n]{x_1 \times x_2 \times x_3 \cdots x_n}\right) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

Which leads us to the equation for the geometric mean:

$$\tilde{x} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right)$$

5.2.4 When to use the geometric mean

The arithmetic mean is relevant any time several quantities add together to produce a total. The arithmetic mean answers the question, "if all the quantities had the same value, what would that value have to be in order to achieve the same total?"

In the same way, the geometric mean is relevant any time several quantities multiply together to produce a product. The geometric mean answers the question, "if all the quantities had the same value, what would that value have to be in order to achieve the same product?"

For example, suppose you have an investment which returns 10% the first year, 50% the second year, and 30% the third year. What is its average rate of return? It is not the arithmetic mean, because what these numbers mean is that on the first year your investment was multiplied (not added to) by 1.10, on the second year it was multiplied by 1.50, and the third year it was multiplied by 1.30. The relevant quantity is the geometric mean of these three numbers.

It is known that the geometric mean is always less than or equal to the arithmetic mean (equality holding only when A=B). The proof of this is quite short and follows from the fact that $(\sqrt(A) - \sqrt(B))^2$ is always a non-negative number. This inequality can be surprisingly powerful though and comes up from time to time in the proofs of theorems in calculus. SOURCE[7].

5.2.5 Harmonic Mean

The arithmetic mean cannot be used when we want to average quantities such as speed.

Consider the example below:

Example 1: The distance from my house to town is 40 km. I drove to town at a speed of 40 km per hour and returned home at a speed of 80 km per hour. What was my average speed for the whole trip?.

Solution: If we just took the arithmetic mean of the two speeds I drove at, we would get 60 km per hour. This isn't the correct average speed, however: it ignores the fact that I drove at 40 km per hour for twice as long as I drove at 80 km per hour. To find the correct average speed, we must instead calcuate the harmonic mean.

For two quantities A and B, the harmonic mean is given by: $\frac{2}{\frac{1}{A}+\frac{1}{B}}$

This can be simplified by adding in the denominator and multiplying by the reciprocal: $\frac{2}{\frac{1}{A}+\frac{1}{B}} = \frac{2}{\frac{B+A}{AB}} = \frac{2AB}{A+B}$

For N quantities: A, B, C......

Harmonic mean = $\frac{N}{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\ldots}$

Let us try out the formula above on our example:

Harmonic mean = $\frac{2AB}{A+B}$

Our values are A = 40, B = 80. Therefore, harmonic mean $= \frac{2\times40\times80}{40+80} = \frac{6400}{120} \approx 53.333$

Is this result correct? We can verify it. In the example above, the distance between the two towns is 40 km. So the trip from A to B at a speed of 40 km will take 1 hour. The trip

[7] HTTP://WWW.MATH.TORONTO.EDU/MATHNET/QUESTIONCORNER/GEOMEAN.HTML

from B to A at a speed to 80 km will take 0.5 hours. The total time taken for the round distance (80 km) will be 1.5 hours. The average speed will then be $\frac{80}{1.5} \approx 53.33$ km/hour.

The harmonic mean also has physical significance.

5.2.6 Relationships among Arithmetic, Geometric and Harmonic Mean

The Means mentioned above are realizations of the generalized mean

$$\bar{x}(m) = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} |x_i|^m \right)^{1/m}$$

and ordered this way:

$Minimum = \bar{x}(-\infty)$

$< harmonicMean = \bar{x}(-1)$

$< geometricMean = \bar{x}(0)$

$< arithmeticMean = \bar{x}(1)$

$< Maximum = \bar{x}(\infty)$

5.3 Measures of dispersion

5.3.1 Range of Data

The **range** of a sample (set of data) is simply the maximum possible difference in the data, i.e. the difference between the maximum and the minimum values. A more exact term for it is "**range width**" and is usually denoted by the letter R or w. The two individual values (the max. and min.) are called the **"range limits"**. Often these terms are confused and students should be careful to use the correct terminology.

For example, in a sample with values 2 3 5 7 8 11 12, the range is 10 and the range limits are 2 and 12.

The range is the simplest and most easily understood measure of the dispersion (spread) of a set of data, and though it is very widely used in everyday life, it is too rough for serious statistical work. It is not a "robust" measure, because clearly the chance of finding the maximum and minimum values in a population depends greatly on the size of the sample we choose to take from it and so its value is likely to vary widely from one sample to another. Furthermore, it is not a satisfactory descriptor of the data because it depends on only two items in the sample and overlooks all the rest. A far better measure of dispersion is the standard deviation (s), which takes into account all the data. It is not only more robust and "efficient" than the range, but is also amenable to far greater statistical manipulation.

Nevertheless the range is still much used in simple descriptions of data and also in quality control charts.

The **mean range** of a set of data is however a quite efficient measure (statistic) and can be used as an easy way to calculate s. What we do in such cases is to subdivide the data into groups of a few members, calculate their average range, $\bar{R}$ and divide it by a factor (from tables), which depends on n. In chemical laboratories for example, it is very common to analyse samples in duplicate, and so they have a large source of ready data to calculate s.

$$s = \frac{\bar{R}}{k}$$

(The factor k to use is given under standard deviation.)

For example: If we have a sample of size 40, we can divide it into 10 sub-samples of n=4 each. If we then find their mean range to be, say, 3.1, the standard deviation of the parent sample of 40 items is appoximately 3.1/2.059 = 1.506.

With simple electronic calculators now available, which can calculate s directly at the touch of a key, there is no longer much need for such expedients, though students of statistics should be familiar with them.

5.3.2 Quartiles

The quartiles of a data set are formed by the two boundaries on either side of the median, which divide the set into four equal sections. The lowest 25% of the data being found below the first quartile value, also called the lower quartile (Q1). The median, or second quartile divides the set into two equal sections. The lowest 75% of the data set should be found below the third quartile, also called the upper quartile (Q3). These three numbers are measures of the dispersion of the data, while the mean, median and mode are measures of central tendency.

Examples

Given the set {1,3,5,8,9,12,24,25,28,30,41,50} we would find the first and third quartiles as follows:

There are 12 elements in the set, so 12/4 gives us three elements in each quarter of the set.

So the first or lowest quartile is: **5**, the second quartile is the median**12**, and the third or upper quartile is **28.**

However some people when faced with a set with an even number of elements (values) still want the true median (or middle value), with an equal number of data values on each side of the median (rather than 12 which has 5 values less than and 6 values greater than. This value is then the average of 12 and 24 resulting in 18 as the true median (which is closer to the mean of 19 2/3. The same process is then applied to the lower and upper quartiles, giving **6.5**, **18**, and **29**. This is only an issue if the data contains an even number of elements with an even number of equally divided sections, or an odd number of elements with an odd number of equally divided sections.

Inter-Quartile Range

The inter quartile range is a statistic which provides information about the spread of a data set, and is calculated by subtracting the first quartile from the third quartile), giving the range of the middle half of the data set, trimming off the lowest and highest quarters. Since the IQR is not affected at all by OUTLIERS[8] in the data, it is a more robust measure of dispersion than the RANGE[9]

IQR = Q3 - Q1

Another useful quantile is the **quintiles** which subdivide the data into five equal sections. The advantage of quintiles is that there is a central one with boundaries on either side of the median which can serve as an average group. In a Normal distribution the boundaries of the quintiles have boundaries ±0.253*s and ±0.842*s on either side of the mean (or median),where s is the sample standard deviation. Note that in a Normal distribution the mean, median and mode coincide.

Other frequently used quantiles are the **deciles** (10 equal sections) and the **percentiles** (100 equal sections)

[8] HTTP://EN.WIKIPEDIA.ORG/WIKI/OUTLIER%20

[9] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ASUMMARY%2FRANGE%20

5.3.3 Variance and Standard Deviation

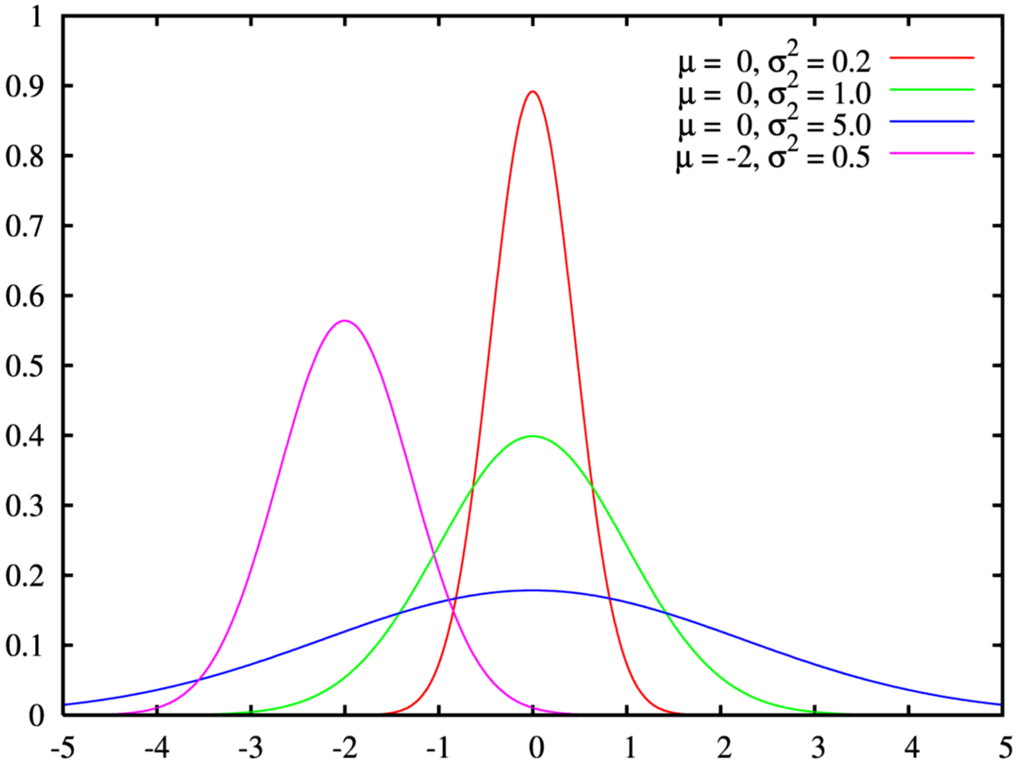


Figure 1: Probability density function for the normal distribution. The green line is the standard normal distribution.

Measure of Scale

When describing data it is helpful (and in some cases necessary) to determine the *spread* of a distribution. One way of measuring this spread is by calculating the variance or the standard deviation of the data.

In describing a complete population, the data represents all the elements of the population. As a measure of the "spread" in the population one wants to know a measure of the possible distances between the data and the population mean. There are several options to do so. One is to measure the average absolute value of the deviations. Another, called the variance, measures the average square of these deviations.

A clear distinction should be made between dealing with the population or with a sample from it. When dealing with the complete population the (population) variance is a constant, a parameter which helps to describe the population. When dealing with a sample from the population the (sample) variance is actually a random variable, whose value differs from sample to sample. Its value is only of interest as an estimate for the population variance.

Population variance and standard deviation

Let the population consist of the N elements $x_1,...,x_N$. The (population) mean is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

.

The *(population) variance* σ^2 is the average of the squared deviations from the mean or (x_i - μ)^2 - the square of the value's distance from the distribution's mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

.

Because of the squaring the variance is not directly comparable with the mean and the data themselves. The square root of the variance is called the Standard Deviation σ. Note that σ is the root mean squared of differences between the data points and the average.

Sample variance and standard deviation

Let the sample consist of the n elements x_1,...,x_n, taken from the population. The (sample) mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

.

The sample mean serves as an estimate for the population mean μ.

The *(sample) variance* s^2 is a kind of average of the squared deviations from the (sample) mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

.

Also for the sample we take the square root to obtain the (sample) standard deviation s

A common question at this point is "why do we square the numerator?" One answer is: to get rid of the negative signs. Numbers are going to fall above and below the mean and, since the variance is looking for distance, it would be counterproductive if those distances factored each other out.

Example

When rolling a fair die, the population consists of the 6 possible outcomes 1 to 6. A sample may consist instead of the outcomes of 1000 rolls of the die.

The population mean is:

$$\mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

,

and the population variance:

$$\sigma^2 = \frac{1}{6}\sum_{i=1}^{n}(i-3.5)^2 = \frac{1}{6}(6.25+2.25+0.25+0.25+2.25+6.25) = \frac{35}{12} \approx 2.917$$

The population standard deviation is:

$$\sigma = \sqrt{\frac{35}{12}} \approx 1.708$$

.

Notice how this standard deviation is somewhere in between the possible deviations.

So if we were working with one six-sided die: $X = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$, then $\sigma^2 = 2.917$. We will talk more about why this is different later on, but for the moment assume that you should use the equation for the sample variance unless you see something that would indicate otherwise.

Note that none of the above formulae are ideal when calculating the estimate and they all introduce rounding errors. Specialized statistical software packages use more complicated LOGARITHMS THAT TAKE A SECOND PASS[10] of the data in order to correct for these errors. Therefore, if it matters that your estimate of standard deviation is accurate, specialized software should be used. If you are using non-specialized software, such as some popular spreadsheet packages, you should find out how the software does the calculations and not just assume that a sophisticated algorithm has been implemented.

For Normal Distributions

The empirical rule states that approximately 68 percent of the data in a normally distributed dataset is contained within one standard deviation of the mean, approximately 95 percent

[10] HTTP://EN.WIKIBOOKS.ORG/WIKI/HANDBOOK_OF_DESCRIPTIVE_STATISTICS/MEASURES_OF_STATISTICAL_VARIABILITY/VARIANCE

of the data is contained within 2 standard deviations, and approximately 99.7 percent of the data falls within 3 standard deviations.

As an example, the verbal or math portion of the SAT has a mean of 500 and a standard deviation of 100. This means that 68% of test-takers scored between 400 and 600, 95% of test takers scored between 300 and 700, and 99.7% of test-takers scored between 200 and 800 assuming a completely normal distribution (which isn't quite the case, but it makes a good approximation).

Robust Estimators

For a normal distribution the relationship between the standard deviation and the interquartile range is roughly: SD = IQR/1.35.

For data that are non-normal, the standard deviation can be a terrible estimator of scale. For example, in the presence of a single outlier, the standard deviation can grossly overestimate the variability of the data. The result is that confidence intervals are too wide and hypothesis tests lack power. In some (or most) fields, it is uncommon for data to be normally distributed and outliers are common.

One robust estimator of scale is the "average absolute deviation", or *aad*. As the name implies, the mean of the absolute deviations about some estimate of location is used. This method of estimation of scale has the advantage that the contribution of outliers is not squared, as it is in the standard deviation, and therefore outliers contribute less to the estimate. This method has the disadvantage that a single large outlier can completely overwhelm the estimate of scale and give a misleading description of the spread of the data.

Another robust estimator of scale is the "median absolute deviation", or *mad*. As the name implies, the estimate is calculated as the median of the absolute deviation from an estimate of location. Often, the median of the data is used as the estimate of location, but it is not necessary that this be so. Note that if the data are non-normal, the mean is unlikely to be a good estimate of location.

It is necessary to scale both of these estimators in order for them to be comparable with the standard deviation when the data are normally distributed. It is typical for the terms *aad* and *mad* to be used to refer to the scaled version. The unscaled versions are rarely used.

External links

w:Variance[11] w:Standard deviation[12]

[11] http://en.wikipedia.org/wiki/Variance

[12] http://en.wikipedia.org/wiki/Standard%20deviation

5.4 Other summaries

5.4.1 Moving Average

A moving average is used when you want to get a general picture of the trends contained in a data set. The data set of concern is typically a so-called "time series", i.e a set of observations ordered in time. Given such a data set $\mathbf{X}$, with individual data points x_i, a 2n+1 point moving average is defined as $\bar{x_i} = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} x_k$, and is thus given by taking the average of the 2n points around x_i. Doing this on all data points in the set (except the points too close to the edges) generates a new time series that is somewhat smoothed, revealing only the general tendencies of the first time series.

The moving average for many time-based observations is often lagged. That is, we take the 10 -day moving average by looking at the average of the last 10 days. We can make this more exciting (who knew statistics was exciting?) by considering different weights on the 10 days. Perhaps the most recent day should be the most important in our estimate and the value from 10 days ago would be the least important. As long as we have a set of weights that sums to 1, this is an acceptable moving-average. Sometimes the weights are chosen along an exponential curve to make the exponential moving-average.

6 Displaying Data

A single statistic tells only part of a dataset's story. The mean is one perspective; the median yet another. And when we explore relationships between multiple variables, even more statistics arise. The coefficient estimates in a regression model, the Cochran-Maentel-Haenszel test statistic in partial contingency tables; a multitude of statistics are available to summarize and test data.

But our ultimate goal in statistics is not to summarize the data, it is to fully understand their complex relationships. A well designed statistical graphic helps us explore, and perhaps understand, these relationships.

This section will help you let the data speak, so that the world may know its story.

Statistics[1] | >> Bar Charts[2]

6.1 External Links

- "The Visual Display of Quantitative Information"[3] is the seminal work on statistical graphics. It is a must read.
- http://search.barnesandnoble.com/booksearch/isbnInquiry.asp?z=y&isbn=0970601999&it "Show me the Numbers" by Stephen Few has a less technical approach to creating graphics. You might want to scan through this book if you are building a library on making graphs.

[1] http://en.wikibooks.org/wiki/Statistics

[2] Chapter 7 on page 39

[3] http://www.edwardtufte.com/tufte/books_vdqi

[4] http://search.barnesandnoble.com/booksearch/isbnInquiry.asp?z=y&isbn=0970601999&itm=1

7 Bar Charts

The Bar Chart (or Bar Graph) is one of the most common ways of displaying catagorical/qualitative data. Bar Graphs consist of 2 variables, one response (sometimes called "dependent") and one predictor (sometimes called "independent"), arranged on the horizontal and vertical axis of a graph. The relationship of the predictor and response variables is shown by a mark of some sort (usually a rectangular box) from one variable's value to the other's.

To demonstrate we will use the following data(tbl. 3.1.1) representing a hypothetical relationship between a qualitative predictor variable, "Graph Type", and a quantitative response variable, "Votes".

tbl. 3.1.1 - Favourite Graphs

Graph Type	**Votes**
Bar Charts	10
Pie Graphs	2
Histograms	3
Pictograms	8
Comp. Pie Graphs	4
Line Graphs	9
Frequency Polygon	1
Scatter Graphs	5

From this data we can now construct an appropriate graphical representation which, in this case will be a Bar Chart. The graph may be orientated in several ways, of which the vertical chart (fig. 3.1.1) is most common, with the horizontal chart(fig. 3.1.2) also being used often

fig. 3.1.1 - vertical chart

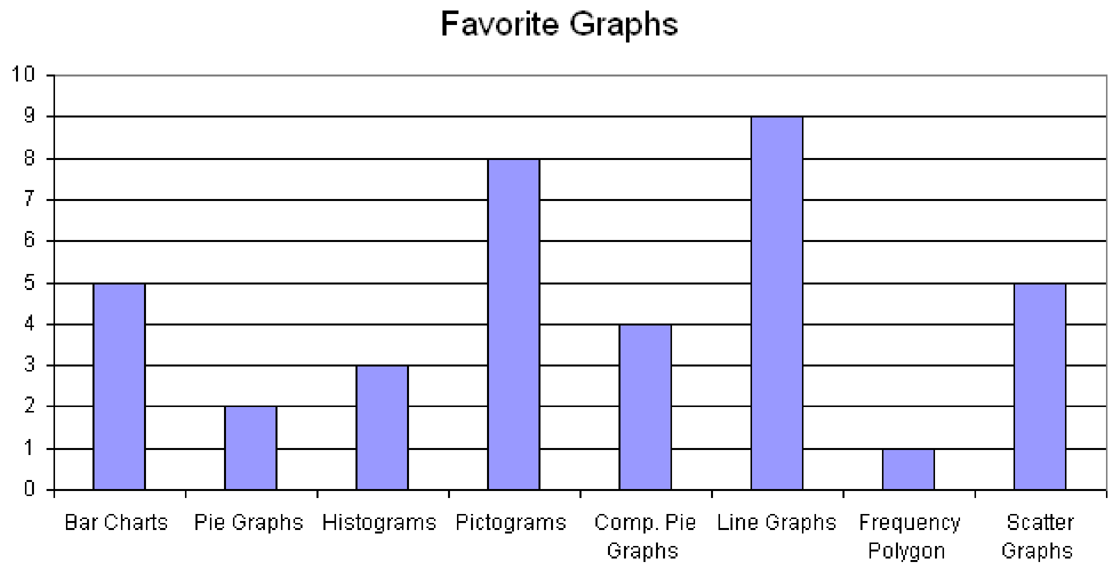


Figure 2: Vertical Bar Chart Example

fig. 3.1.2 - horizontal chart

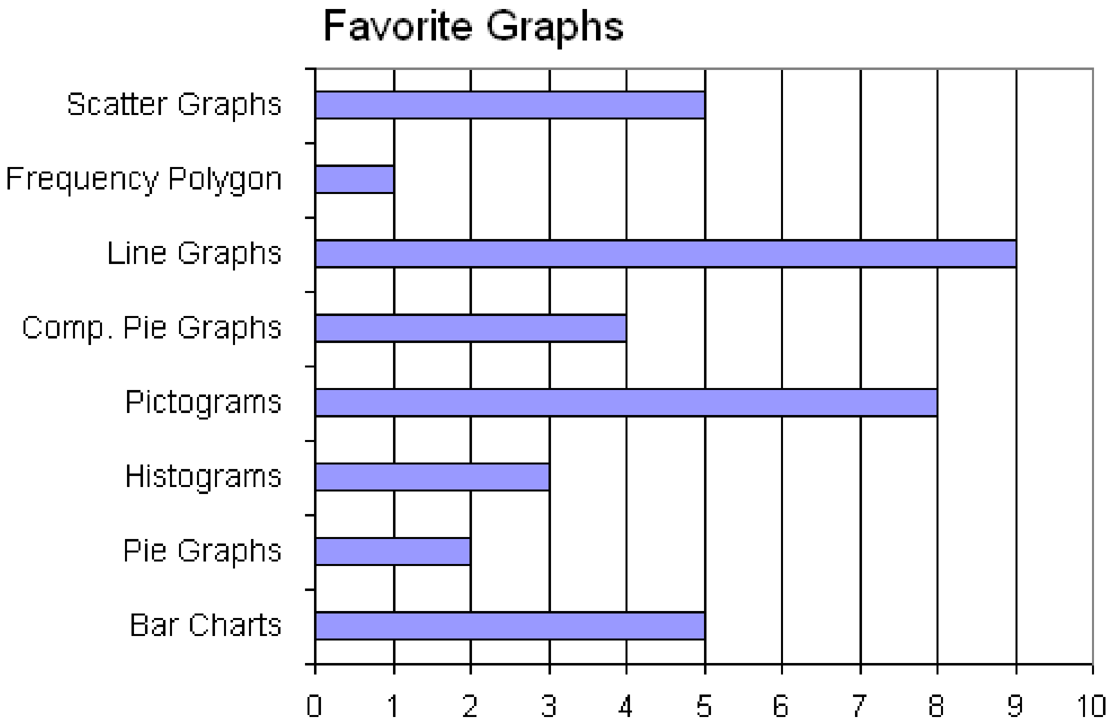


Figure 3: Horizontal Bar Chart Example

Take note that the height and width of the bars, in the vertical and horizontal Charts, respectfully, are equal to the response variable's corresponding value - "Bar Chart" bar equals the number of votes that the Bar Chart type received in tbl. 3.1.1

Also take note that there is a pronounced amount of space between the individual bars in each of the graphs, this is important in that it help differentiate the Bar Chart graph type from the Histogram graph type discussed in a later section.

7.1 External Links

- Interactive Java-based Bar-Chart Applet[1]

[1] http://socr.ucla.edu/htmls/chart/BoxAndWhiskersChartDemo3_Chart.html

8 Histograms

8.1 Histograms

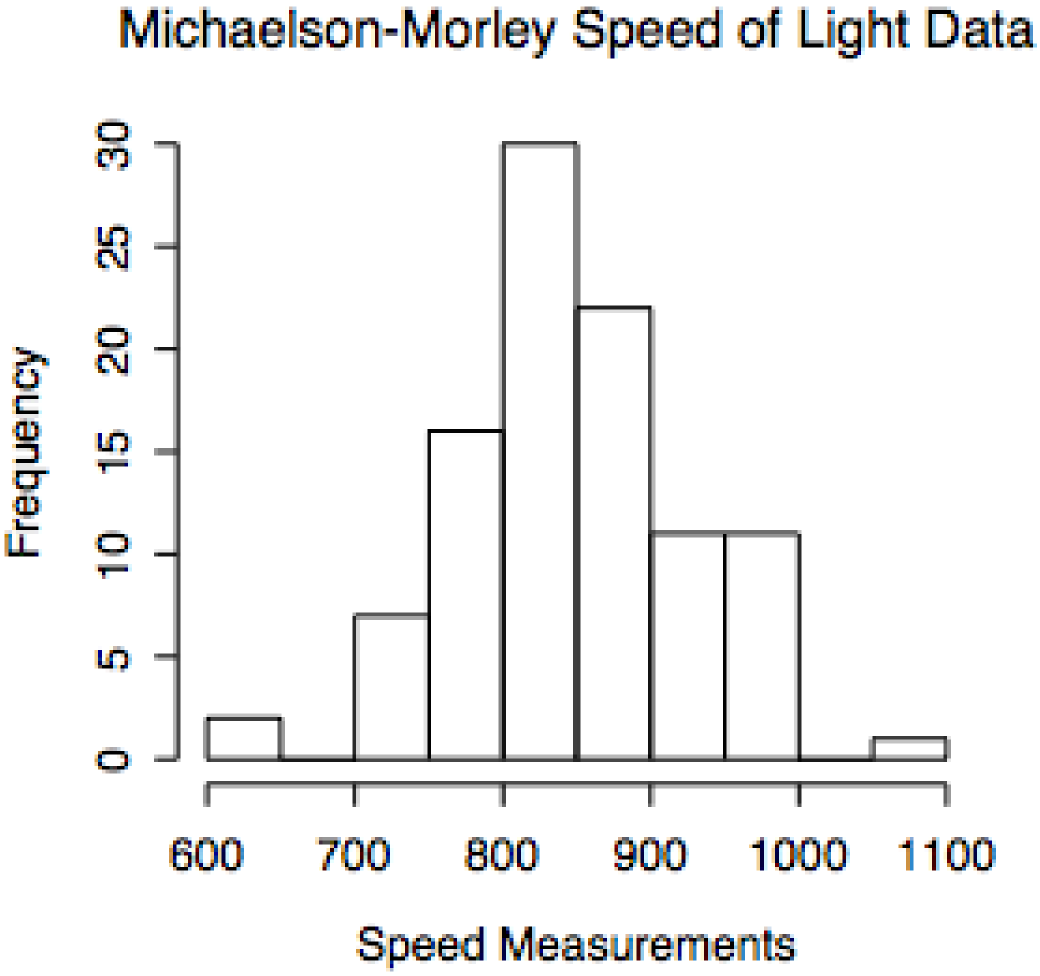


Figure 4

It is often useful to look at the distribution of the data, or the frequency with which certain values fall between pre-set bins of specified sizes. The selection of these bins is up to you, but remember that they should be selected in order to *illuminate* your data, not *obfuscate* it.

To produce a histogram:

- **Select a minimum, a maximum, and a bin size.** All three of these are up to you. In the Histogram data used above the minimum is 1, the maximum is 110, and the bin size is 10.
- **Calculate your bins and how many values fall into each of them.** For the Histogram data the bins are:
  - $1 \leq x < 10$, 16 values.
  - $10 \leq x < 20$, 4 values.
  - $20 \leq x < 30$, 4 values.
  - $30 \leq x < 40$, 2 values.
  - $40 \leq x < 50$, 2 values.
  - $50 \leq x < 60$, 1 values.
  - $60 \leq x < 70$, 0 values.
  - $70 \leq x < 80$, 0 values.
  - $80 \leq x < 90$, 0 values.
  - $90 \leq x < 100$, 0 value.
  - $100 \leq x < 110$, 0 value.
  - $110 \leq x < 120$, 1 value.
- **Plot the counts you figured out above.** Do this using a standard BAR PLOT[1].

There! You are done. Now let's do an example.

8.1.1 Worked Problem

Let's say you are an avid roleplayer who loves to play Mechwarrior, a d6 (6 sided die) based game. You have just purchased a new 6 sided die and would like to see whether it is biased (in combination with you when you roll it).

What We Expect

So before we look at what we get from rolling the die, let's look at what we would expect. First, if a die is unbiased it means that the odds of rolling a six are exactly the same as the odds of rolling a 1--there wouldn't be any favoritism towards certain values. Using the standard equation for the ARITHMETIC MEAN[2] find that $\mu = 3.5$. We would also expect the histogram to be roughly even all of the way across--though it will almost never be perfect simply because we are dealing with an element of random chance.

What We Get

Here are the numbers that you collect:

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ADISPLAYING_DATA%2FBAR_CHARTS

[2] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ASUMMARY%2FAVERAGES%2FMEAN%23MEAN

1	5	6	4	1	3	5	5	6	4	1	5	6	6	4	5	1	4	3	6
1	3	6	4	2	4	1	6	4	2	2	4	3	4	1	1	6	3	5	5
4	3	5	3	4	2	2	5	6	5	4	3	5	3	3	1	5	4	4	5
1	2	5	1	6	5	4	3	2	4	2	1	3	3	3	4	6	1	1	3
6	6	1	4	6	6	6	5	3	1	5	6	3	4	5	5	5	2	4	4

Analysis

$$\bar{X} = 3.71$$

Referring back to what we would expect for an unbiased die, this is pretty close to what we would expect. So let's create a histogram to see if there is any significant difference in the distribution.

The only logical way to divide up dice rolls into bins is by what's showing on the die face:

1	**2**	**3**	**4**	**5**	**6**
16	9	17	21	20	17

If we are good at visualizing information, we can simple use a table, such as in the one above, to see what might be happening. Often, however, it is useful to have a visual representation. As the amount of variety of data we want to display increases, the need for graphs instead of a simple table increases.

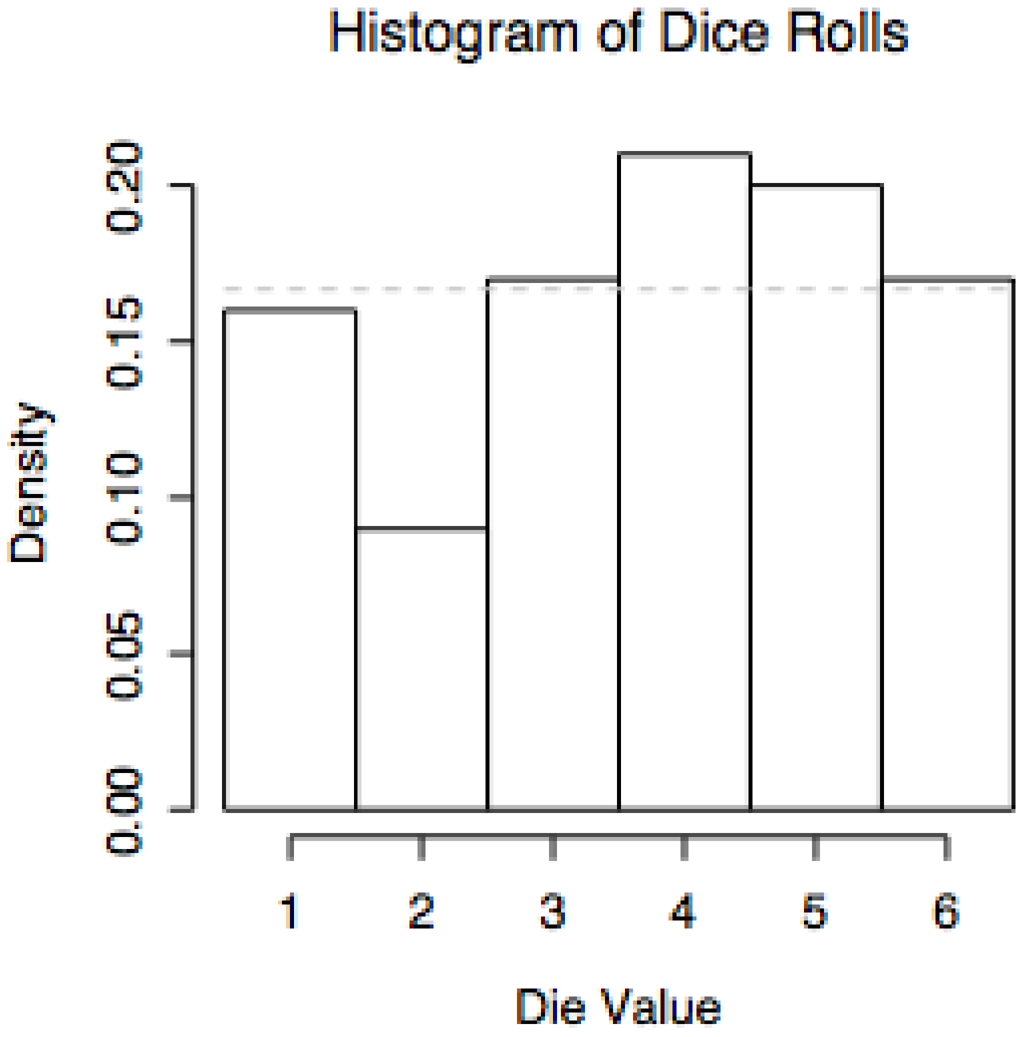


Figure 5

Looking at the above figure, we clearly see that sides 1, 3, and 6 are almost exactly what we would expect by chance. Sides 4 and 5 are slightly greater, but not too much so, and side 2 is a lot less. This could be the result of chance, or it could represent an actual anomaly in the data and it is something to take note of keep in mind. We'll address this issue again in later chapters.

8.1.2 Frequency Density

Another way of drawing a histogram is to work out the Frequency Density.

Frequency Density

The *Frequency Density* is the frequency divided by the class width.

The advantage of using frequency density in a histogram is that doesn't matter if there isn't an obvious standard width to use. For all the groups, you would work out the frequency divided by the class width for all of the groups.

8.2 External Links

- Interactive Java-based Bar-Chart Applet[3]

Statistics[4]

[3] http://socr.ucla.edu/htmls/chart/HistogramChartDemo1_Chart.html

[4] http://en.wikibooks.org/wiki/Statistics

9 Scatter Plots

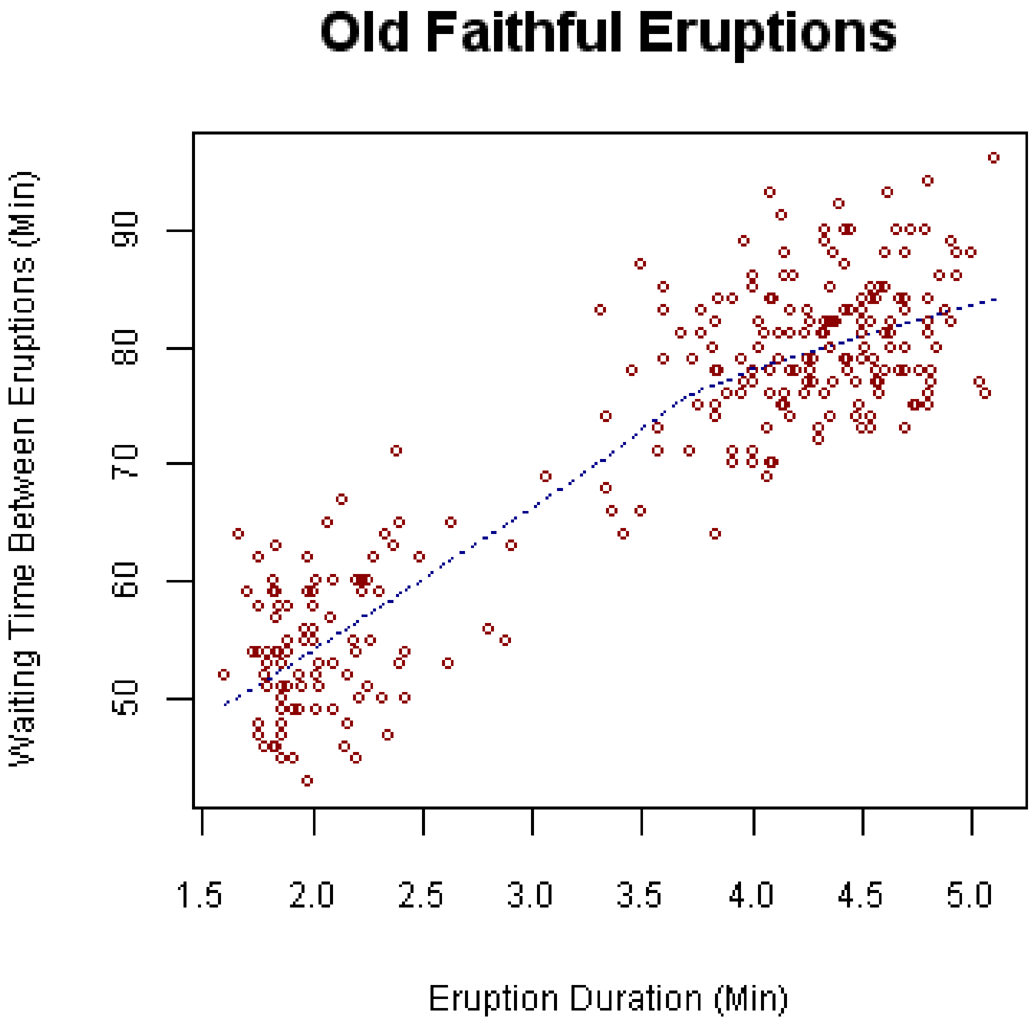


Figure 6

Scatter Plot is used to show the relationship between 2 numeric variables. It is not useful when comparing discrete variables versus numeric variables. A scatter plot matrix is a collection of pairwise scatter plots of numeric variables.

9.1 External Links

- Interactive Java-based Bar-Chart Applet[1]

[1] http://socr.ucla.edu/htmls/chart/ScatterChartDemo1_Chart.html

10 Box Plots

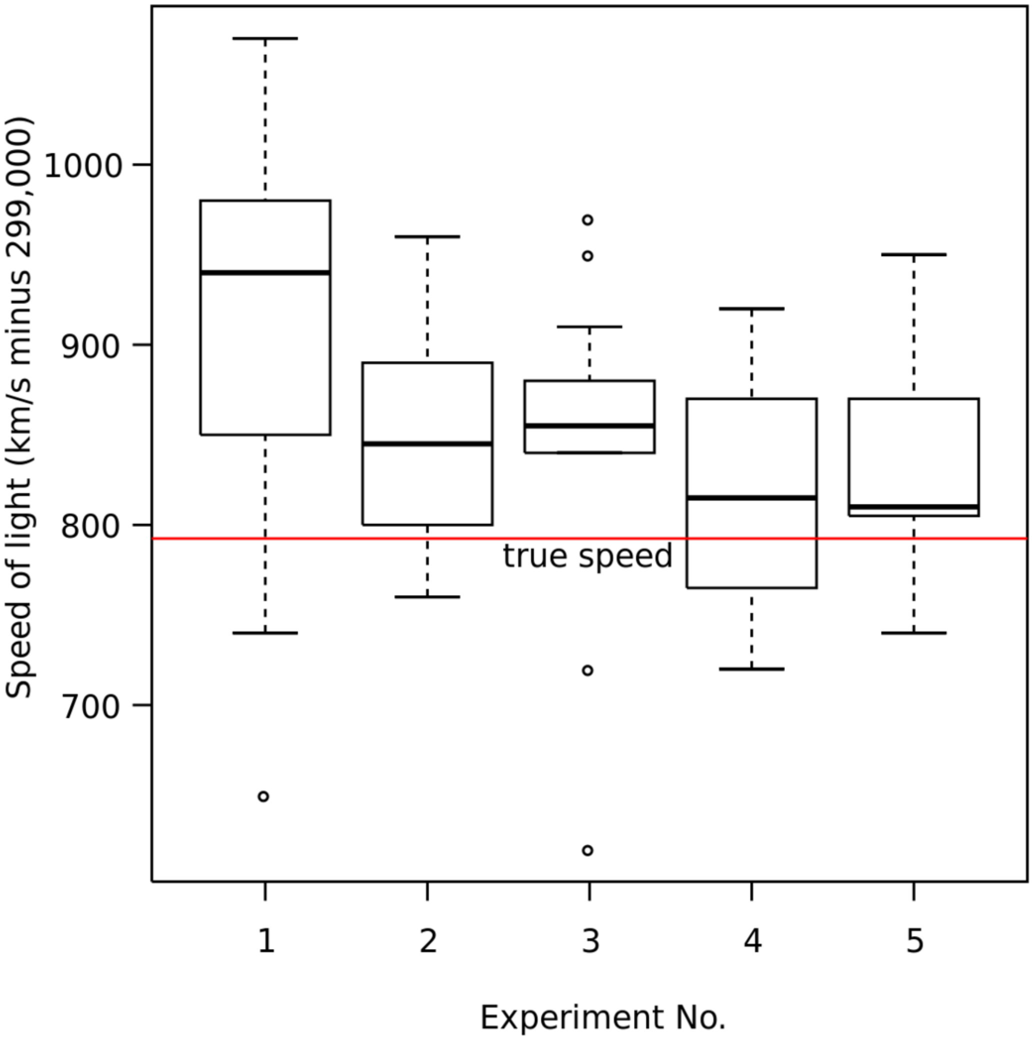


Figure 7: Figure 1. Box plot of data from the Michelson-Morley Experiment

A **box plot** (also called a box and whisker diagram) is a simple visual representation of key features of a univariate sample.

The box lies on a vertical axis in the range of the sample. Typically, a top to the box is placed at the 1st quartile, the bottom at the third quartile. The width of the box is arbitrary, as there is no x-axis (though see Violin Plots, below).

In between the top and bottom of the box is some representation of central tendency. A common version is to place a horizontal line at the median, dividing the box into two. Additionally, a star or asterisk is placed at the mean value, centered in the box in the horizontal direction.

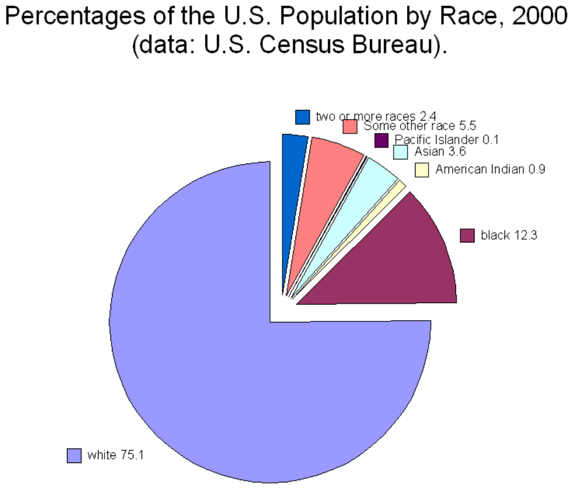
Another common extension is to the 'box-and-whisker' plot. This adds vertical lines extending from the top and bottom of the plot to for example, the maximum and minimum values, The farthest value within 2 standard deviations above and below the mean. Alternatively, the whiskers could extend to the 2.5 and 97.5 percentiles. Finally, it is common in the box-and-whisker plot to show OUTLIERS[1] (however defined) with asterisks at the individual values beyond the ends of the whiskers.

Violin Plots are an extension to box plots using the horizontal information to present more data. They show some estimate of the CDF[2] instead of a box, though the quantiles of the distribution are still shown.

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/OUTLIERS

[2] HTTP://EN.WIKIBOOKS.ORG/WIKI/CDF

11 Pie Charts



Percentages of the U.S. Population by Race, 2000
(data: U.S. Census Bureau).

Figure 8: A pie chart showing the racial make-up of the US in 2000.

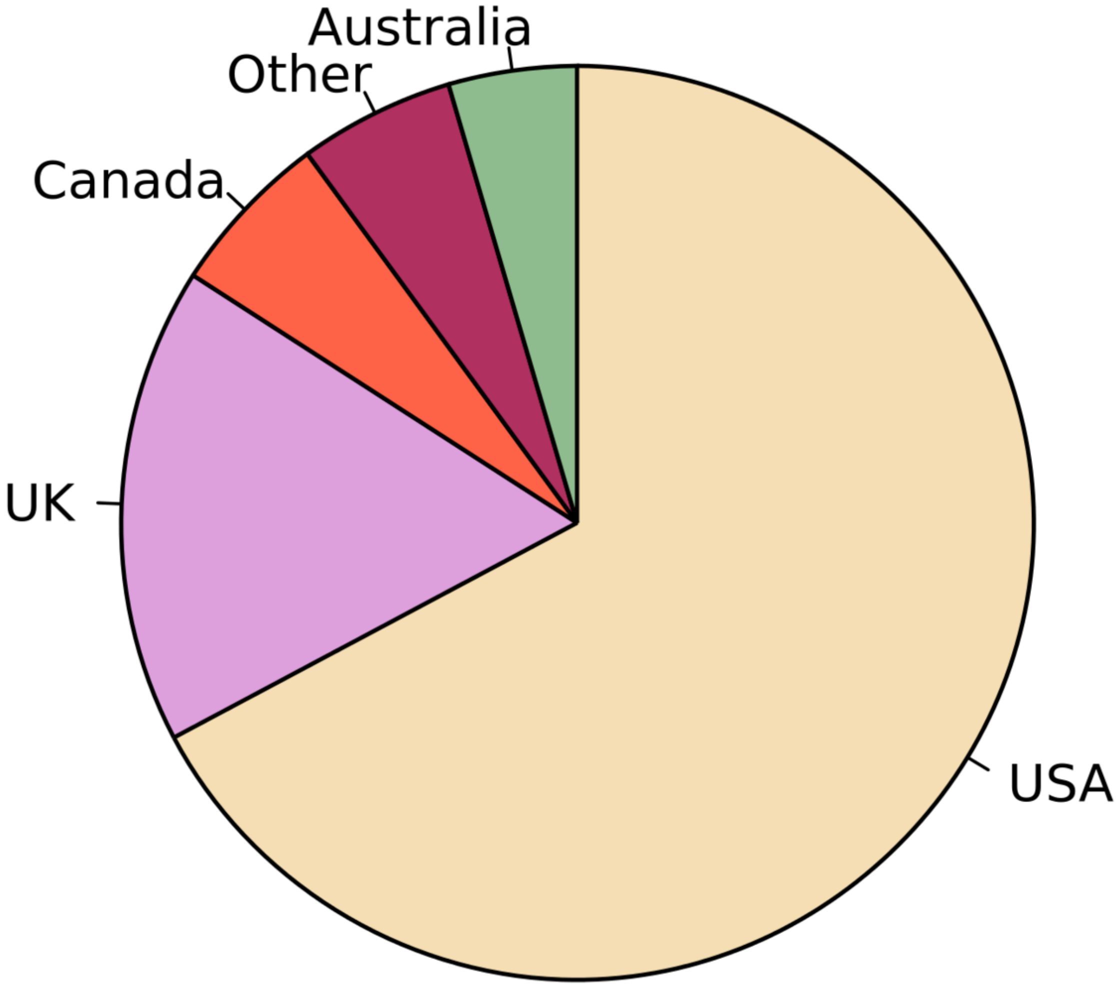


Figure 9: Pie chart of populations of English language-speaking people

A Pie-Chart/Diagram is a graphical device - a circular shape broken into sub-divisions. The sub-divisions are called "*sectors*", whose areas are proportional to the various parts into which the whole quantity is divided. The sectors may be coloured differently to show the relationship of parts to the whole. A pie diagram is an alternative of the sub-divided bar diagram.

To construct a pie-chart, first we draw a circle of any suitable radius then the whole quantity which is to be divided is equated to 360 degrees. The different parts of the circle in terms of angles are calculated by the following formula.

```
Component Value / Whole Quantity * 360
```

The component parts i.e. sectors have been cut beginning from top in clockwise order.

Note that the percentages in a list may not add up to exactly 100% due to rounding. For example if a person spends a third of their time on each of three activities: 33%, 33% and 33% sums to 99%.

Warning: Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

11.1 External Links

- Interactive Java-based Pie-Chart Applet[1]

[1] http://socr.ucla.edu/htmls/chart/PieChartDemo1_Chart.html

12 Comparative Pie Charts

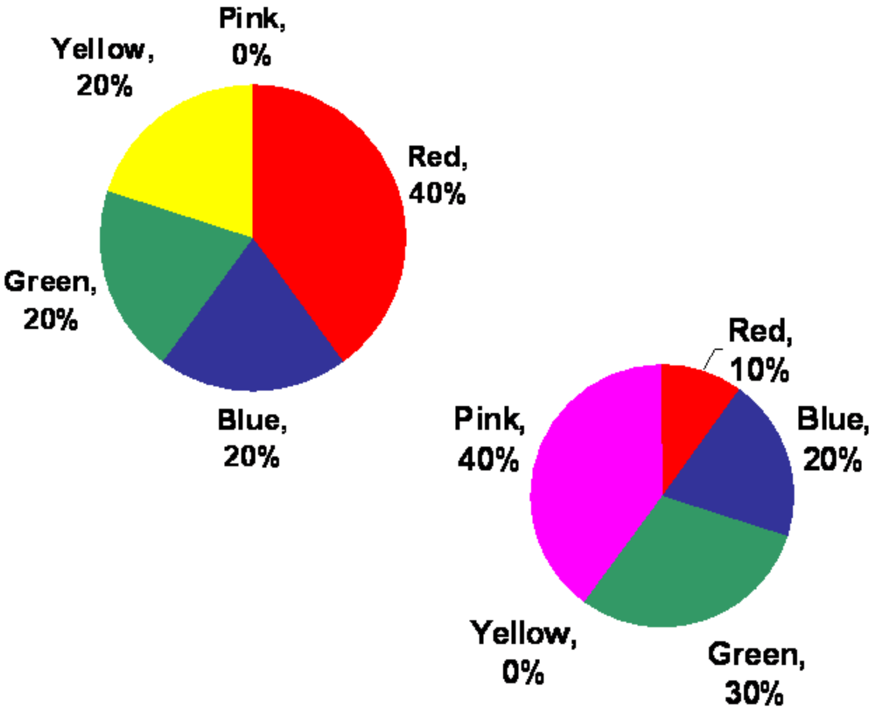


Figure 10: A pie chart showing preference of colors by two groups.

The comparative pie charts are very difficult to read and compare if the ratio of the pie chart is not given.

Examine our example of color preference for two different groups. How much work does it take to see that the Blue preference for both groups is the same? First, we have to find blue on each pie, and then remember how many degrees it has. If we did not include the share for blue in the label, then we would probably be approximating the comparison. So, if we use multiple pie charts, we have to expect that comparisions between charts would only be approximate.

What is the most popular color in the left graph? Red. But note, that you have to look at all of the colors and read the label to see which it might be. Also, this author was kind when creating these two graphs because I used the same color for the same object. Imagine the confusion if one had made the most important color get Red in the right-hand chart?

If two shares of data should not be compared via the comparative pie chart, what kind of graph would be preferred? The stacked bar chart is probably the most appropriate for sharing of the total comparisons. Again, exact comparisons cannot be done with graphs and therefore a table may supplement the graph with detailed information.

13 Pictograms

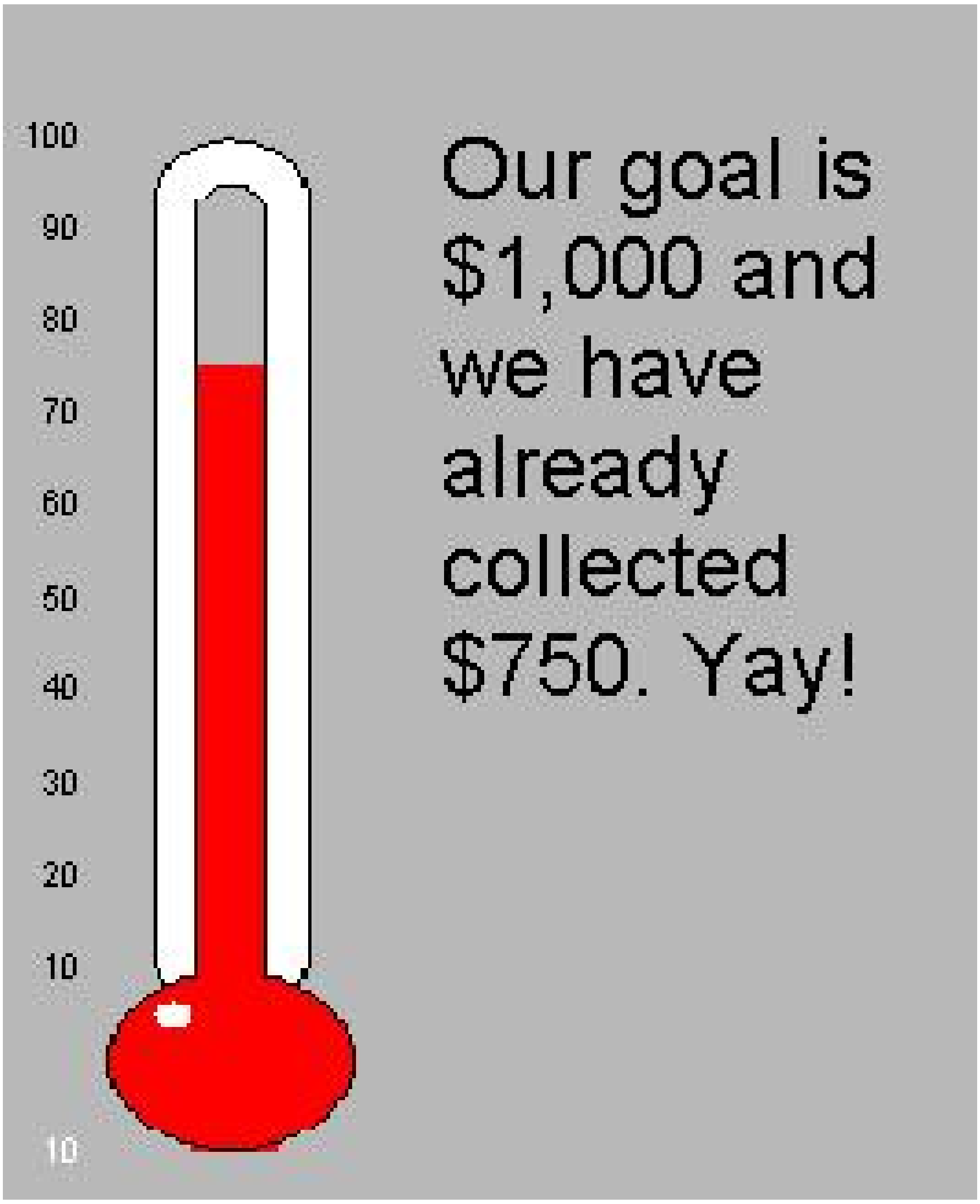


Figure 11

A pictogram is simply a picture that conveys some statistical information. A very common example is the thermometer graph so common in fund drives. The entire thermometer is the goal (number of dollars that the fund raisers wish to collect. The red stripe (the "mercury") represents the proportion of the goal that has already been collected.

Another example is a picture that represents the gender constitution of a group. Each small picture of a male figure might represent 1,000 men and each small picture of a female figure would, then, represent 1,000 women. A picture consisting of 3 male figures and 4 female figures would indicate that the group is made up of 3,000 men and 4,000 women.

An interesting pictograph is the Chernoff Faces. It is useful for displaying information on cases for which several variables have been recorded. In this kind of plot, each case is represented by a separate picture of a face. The sizes of the various features of each face are used to present the value of each variable. For instance, if blood pressure, high density cholesterol, low density cholesterol, body temperature, height, and weight are recorded for 25 individuals, 25 faces would be displayed. The size of the nose on each face would represent the level of that person's blood pressure. The size of the left eye may represent the level of low density cholesterol while the size of the right eye might represent the level of high density cholesterol. The length of the mouth could represent the person's temperature. The length of the left ear might indicate the person's height and that of the right ear might represent their weight. Of course, a legend would be provided to help the viewer determine what feature relates to which variable. Where it would be difficult to represent the relationship of all 6 variables on a single (6-dimensional) graph, the Chernoff Faces would give a relatively easy to interpret 6-dimensional representation.

14 Line Graphs

Basically, a **line graph** can be, for example, a picture of what happened by/to something (a variable) during a specific time period (also a variable).

On the left side of such a graph usually is as an indication of that "something" in the form of a scale, and at the bottom is an indication of the specific time involved.

Usually a **line graph** is plotted after a table has been provided showing the relationship between the two variables in the form of pairs. Just as in (x,y) graphs, each of the pairs results in a specific point on the graph, and being a LINE graph these points are connected to one another by a LINE.

Many other line graphs exist; they all CONNECT the points by LINEs, not necessarily straight lines. Sometimes polynomials, for example, are used to describe approximately the basic relationship between the given pairs of variables, and between these points. The higher the degree of the polynomial, the more accurate is the "picture" of that relationship, but the degree of that polynomial must never be higher than *n-1*, where *n* is the number of the given points.

14.1 See also

Graph theory[1]

Curve fitting[2]

From Wikipedia: Line graph[3] and Curve fitting[4]

14.2 External Links

- Interactive Java-based Line Graph Applet[5]

[1] http://en.wikibooks.org/wiki/Discrete%20Mathematics%2FGraph%20theory

[2] http://en.wikibooks.org/wiki/..%2F..%2FCurve%20fitting

[3] http://en.wikipedia.org/wiki/Line%20graph

[4] http://en.wikipedia.org/wiki/Curve%20fitting

[5] http://socr.ucla.edu/htmls/chart/LineChartDemo1_Chart.html

15 Frequency Polygon

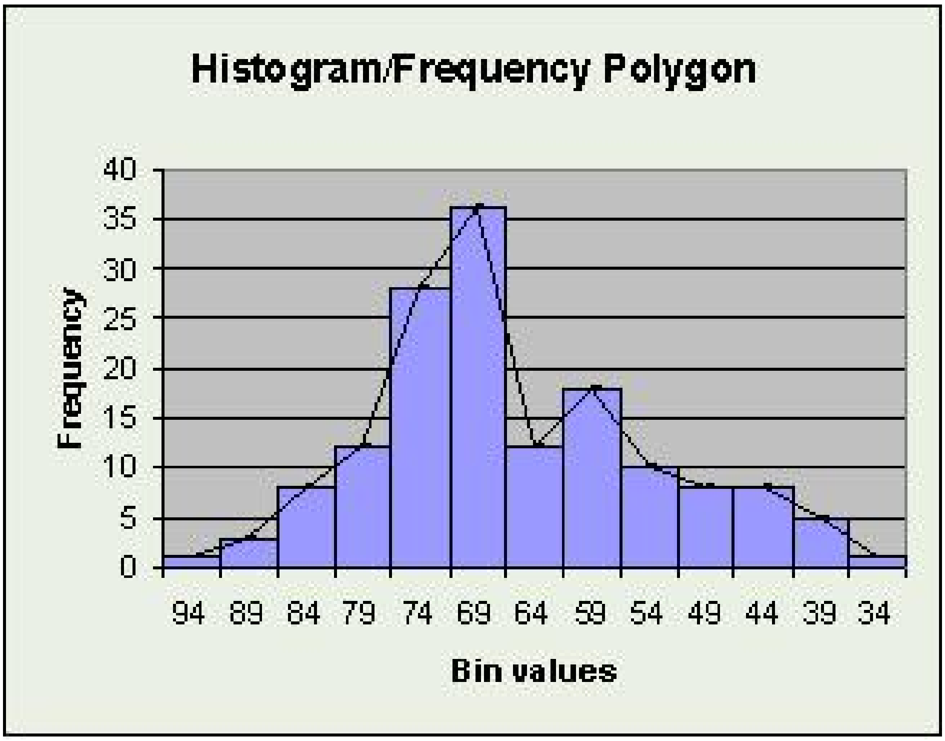


Figure 12: This is a histogram with an overlaid frequency polygon.

Midpoints of the interval of corresponding rectangle in a histogram are joined together by straight lines. It gives a polygon i.e. a figure with many angles. it is used when two or more sets of data are to be illustrated on the same diagram such as death rates in smokers and non smokers, birth and death rates of a population etc

One way to form a frequency polygon is to connect the midpoints at the top of the bars of a histogram with line segments (or a smooth curve). Of course the midpoints themselves could easily be plotted without the histogram and be joined by line segments. Sometimes it is beneficial to show the histogram and frequency polygon together.

Unlike histograms, frequency polygons can be superimposed so as to compare several frequency distributions.

16 Introduction to Probability

Figure 13: When throwing two dice, what is the probability that their sum equals seven?

16.1 Introduction to probability

Please note that this page is just a stub, more will be added later.

16.1.1 Why have probability in a statistics textbook?

Very little in mathematics is truly self contained. Many branches of mathematics touch and interact with one another, and the fields of probability and statistics are no different. A basic understanding of probability is vital in grasping basic statistics, and probability is largely abstract without statistics to determine the "real world" probabilities.

This section is not meant to give a comprehensive lecture in probability, but rather simply touch on the basics that are needed for this class, covering the basics of Bayesian Analysis for those students who are looking for something a little more interesting. This knowledge will be invaluable in attempting to understand the mathematics involved in various DISTRIBUTIONS[1] that come later.

16.1.2 Set notion

A set is a collection of objects. We usually use capital letters to denote sets, for e.g., A is the set of females in this room.

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ADISTRIBUTIONS

- The members of a set A are called the elements of A. For e.g., Patricia is an element of A (Patricia $\in$ A) Patrick is not an element of A (Patrick $\notin$ A).
- The universal set, U, is the set of all objects under consideration. For e.g., U is the set of all people in this room.
- The null set or empty set, $\varnothing$, has no elements. For e.g., the set of males above 2.8m tall in this room is an empty set.
- The complement A^c of a set A is the set of elements in U outside A. I.e. x $\in$ A^c iff x $\notin$ A.
- Let A and B be 2 sets. A is a subset of B if each element of A is also an element of B. Write A $\subset$ B. For e.g., The set of females wearing metal frame glasses in this room $\subset$ the set of females wearing glasses in this room $\subset$ the set of females in this room.
- The intersection A $\cap$ B of two sets A and B is the set of the common elements. I.e. x $\in$ A $\cap$ B iff x $\in$ A and x $\in$ B.
- The union A $\cup$ B of two sets A and B is the set of all elements from A or B. I.e. x $\in$ A $\cup$ B iff x $\in$ A or x $\in$ B.

16.1.3 Venn diagrams and notation

A Venn diagram visually models defined events. Each event is expressed with a circle. Events that have outcomes in common will overlap with what is known as the intersection of the events.

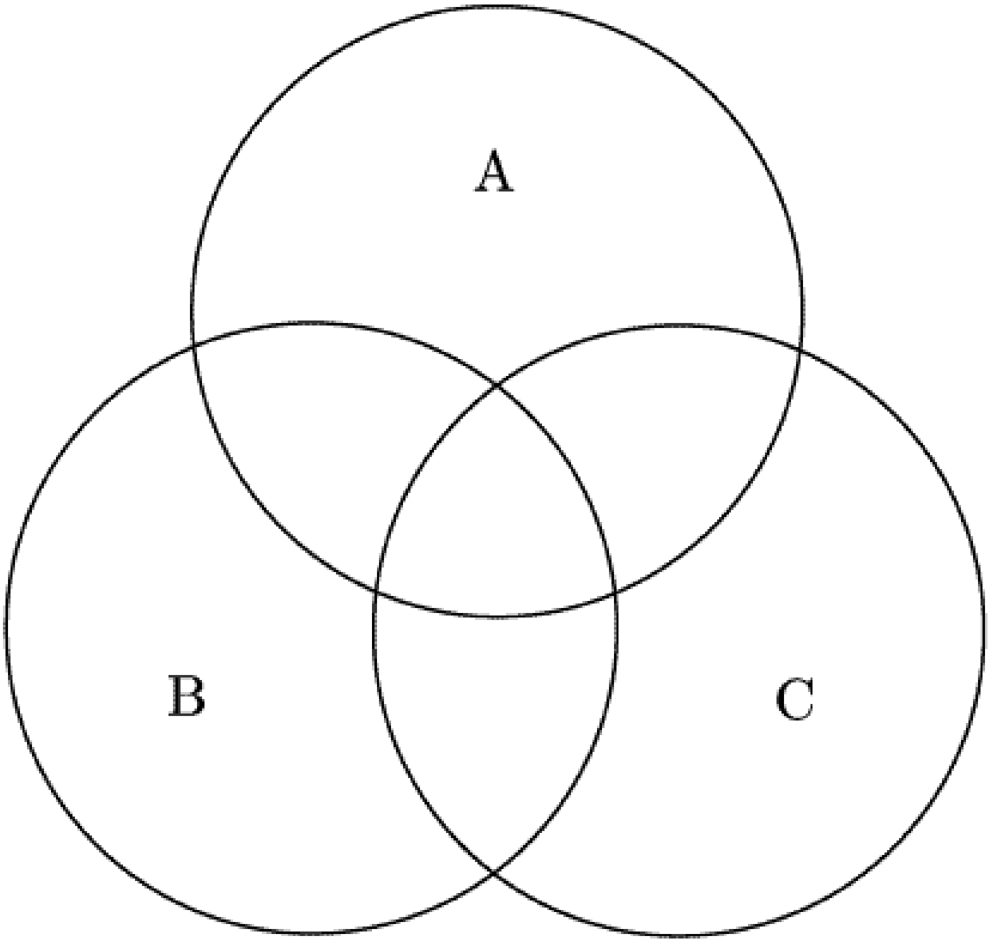


Figure 14: A Venn diagram.

16.2 Probability

Probability is connected with some unpredictability. We know what outcomes may occur, but not exactly which one. The set of possible outcomes plays a basic role. We call it the *sample space* and indicate it by S. Elements of S are called *outcomes*. In rolling a dice the sample space is S = {1,2,3,4,5,6}. Not only do we speak of the outcomes, but also about *events*, sets of outcomes. E.g. in rolling a dice we can ask whether the outcome was an even number, which means asking after the event "even" = E = {2,4,6}. In simple situations with a finite number of outcomes, we assign to each outcome s (∈ S) its *probability* (of occurrence) p(s) (written with a small p), a number between 0 and 1. It is a quite simple function, called the probability function, with the only further property that the total of all the probabilities sum up to 1. Also for events A do we speak of their probability P(A) (written with a capital P), which is simply the total of the probabilities of the outcomes in A. For a fair dice p(s) = 1/6 for each outcome s and P("even") = P(E) = 1/6+1/6+1/6 = 1/2.

The general concept of probability for non-finite sample spaces is a little more complex, although it rests on the same ideas.

16.2.1 Negation

Negation is a way of saying "not A", hence saying that the complement of A has occurred. *Note: The complement of an event A can be expressed as A' or* A^c

For example: "What is the probability that a six-sided die will **not** land on a one?" (five out of six, or $p = 0.833$)

$$P[X'] = 1 - P[X]$$

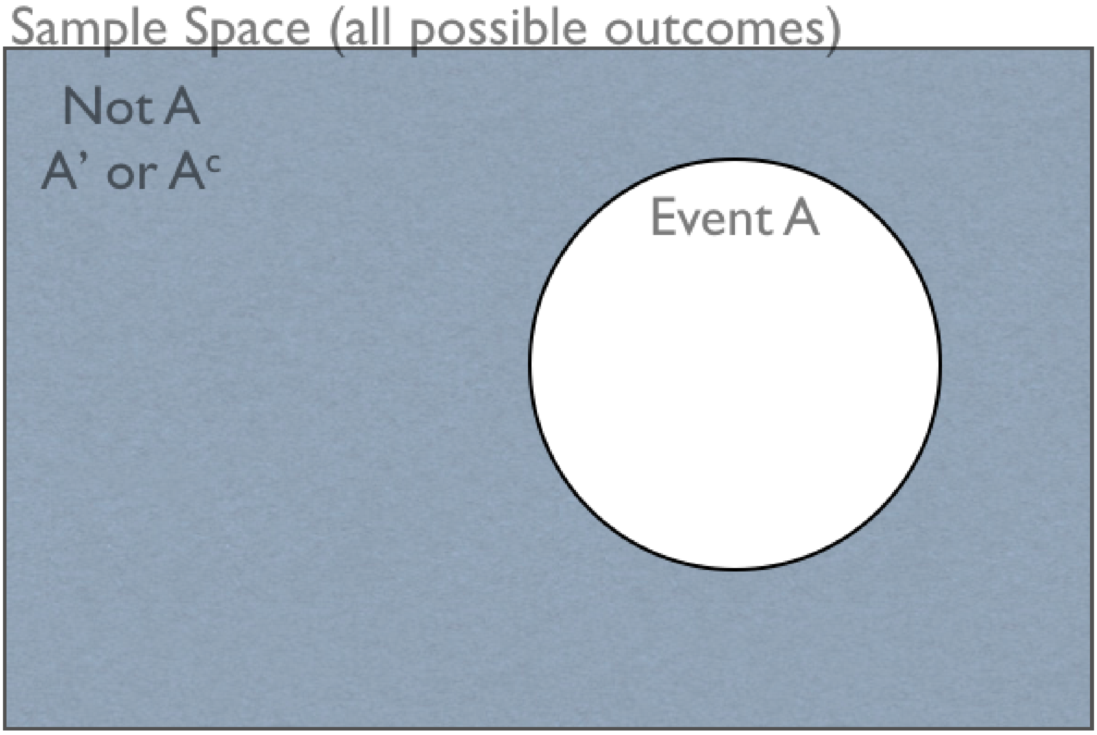


Figure 15: Complement of an Event

Or, more colloquially, "the probability of 'not X' together with the probability of 'X' equals one or 100%."

16.2.2 Calculating Probability

Relative frequency describes the number of successes over the total number of outcomes. For example if a coin is flipped and out of 50 flips 29 are heads then the relative frequency is $\frac{29}{50}$

The Union of two events is when you want to know Event A OR Event B.<Br>

This is different than "And." "And" is the intersection, "OR" is the union of the events (both events put together).

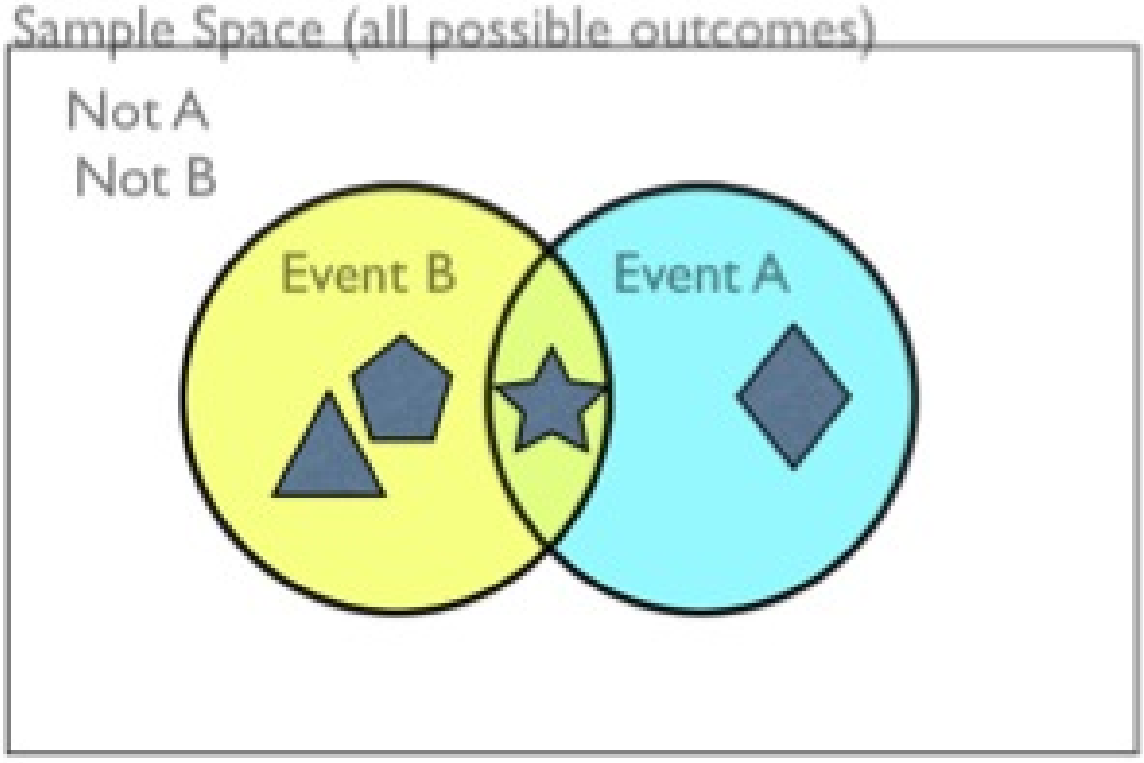


Figure 16

In the above example of events you will notice that...<Br>

Event A is a STAR and a DIAMOND.

Event B is a TRIANGLE and a PENTAGON and a STAR

(A ∩ B) = (A and B) = A intersect B is only the STAR

But (A ∪ B) = (A or B) = A Union B is EVERYTHING. The TRIANGLE, PENTAGON, STAR, and DIAMOND

Notice that both event A and Event B have the STAR in common. However, when you list the Union of the events you only list the STAR one time!

Event A = STAR, DIAMOND EVENT B = TRIANGLE, PENTAGON, STAR

When you combine them together you get (STAR + DIAMOND) + (TRIANGLE + PENTAGON + STAR) BUT WAIT!!! STAR is listed two times, so one will need to SUBTRACT the extra STAR from the list.

You should notice that it is the INTERSECTION that is listed TWICE, so you have to subtract the duplicate intersection.

Formula for the Union of Events: P(A ∪ B) = P(A) + P(B) - P(A ∩ B)

Example: Let P(A) = 0.3 and P(B) = 0.2 and P(A ∩ B) = 0.15. Find P(A ∪ B).

P(A ∪ B) = (0.3) + (0.2) - (0.15) = 0.35

Example: Let P(A) = 0.3 and P(B) = 0.2 and P(A ∩ B) = . Find P(A ∪ B).

Note: Since the intersection of the events is the null set, then you know the events are DISJOINT or MUTUALLY EXCLUSIVE.

P(A ∪ B) = (0.3) + (0.2) - (0) = 0.5

16.2.3 Conjunction

16.2.4 Disjunction

16.2.5 Law of total probability

Generalized case

16.2.6 Conclusion: putting it all together

16.2.7 Examples

17 Bernoulli Trials

A lot of experiments just have two possible outcomes, generally referred to as "success" and "failure". If such an experiment is independently repeated we call them (a series of) **Bernoulli trials**. Usually the probability of success is called p. The repetition may be done in several ways:

- a fixed number of times (n); as a consequence the observed number of successes is stochastic;
- until a fixed number of successes (m) is observed; as a consequence the number of experiments is stochastic;

In the first case the number of successes is Binomial distributed with parameters n and p. For n=1 the distribution is also called the Bernoulli distribution. In the second case the number of experiments is Negative Binomial distributed with parameters m and p. For m=1 the distribution is also called the Geometric distribution.

18 Introductory Bayesian Analysis

Bayesian analysis is the branch of statistics based on the idea that we have some knowledge in advance about the probabilities that we are interested in, so called *a priori* probabilities. This might be your degree of belief in a particular event, the results from previous studies, or a general agreed-upon starting value for a probability. The terminology "Bayesian" comes from the Bayesian rule or law, a law about conditional probabilities. The opposite of "Bayesian" is sometimes referred to as "Classical Statistics."

18.0.8 Example

Consider a box with 3 coins, with probabilities of showing heads respectively 1/4, 1/2 and 3/4. We choose arbitrarily one of the coins. Hence we take 1/3 as the a priori probability $P(C_1)$ of having chosen coin number 1. After 5 throws, in which X=4 times heads came up, it seems less likely that the coin is coin number 1. We calculate the a posteriori probability that the coin is coin number 1, as:

$$P(C_1|X=4) = \frac{P(X=4|C_1)P(C_1)}{P(X=4)} = \frac{P(X=4|C_1)P(C_1)}{P(X=4|C_1)+P(X=4|C_2)+P(X=4|C_3)} = \frac{\binom{5}{4}(\frac{1}{4})^4\frac{3}{4}\frac{1}{3}}{\binom{5}{4}(\frac{1}{4})^4\frac{3}{4}\frac{1}{3}+\binom{5}{4}(\frac{1}{2})^4\frac{1}{2}\frac{1}{3}+}$$

In words:

> The probability that the Coin is the first Coin, given that we know heads came up 4 times... Is equal to the probability that heads came up 4 times given we know it's the first coin, times the probability that the coin is the first coin. All divided by the probability that heads comes up 4 times (ignoring which of the three Coins is chosen). The binomial coefficients cancel out as well as all denominators when expanding 1/2 to 2/4. This results in
>
> $$\frac{3}{3+32+81} = \frac{3}{116}$$

In the same way we find:

$$P(C_2|X=4) = \frac{32}{3+32+81} = \frac{32}{116}$$

and

$$P(C_3|X=4) = \frac{81}{3+32+81} = \frac{81}{116}$$

.

This shows us that after examining the outcome of the five throws, it is most likely we did choose coin number 3.

Actually for a given result the denominator does not matter, only the relative Probabilities $p(C_i) = P(C_i|X=4)/P(X=4)$ When the result is 3 times heads the Probabilities change in favor of Coin 2 and further as the following table shows:

Heads	$p(C_1)$	$p(C_2)$	$p(C_3)$
5	1	32	243
4	3	32	81
3	9	32	27
2	27	32	9
1	81	32	3
0	243	32	1

19 Distributions

How are the results of the latest SAT test? What is the average height of females under 21 in Zambia? How does beer consumption among college students at engineering college compare to college students in liberal arts colleges?

To answer these questions, we would collect data and put them in a form that is easy to summarize, visualize, and discuss. Loosely speaking, the collection and aggregation of data result in a distribution. Distributions are most often in the form of a histogram or a table. That way, we can "see" the data immediately and begin our scientific inquiry.

For example, if we want to know more about students' latest performance on the SAT, we would collect SAT scores from ETS, compile them in a way that is pertinent to us, and then form a distribution of these scores. The result may be a data table or it may be a plot. Regardless, once we "see" the data, we can begin asking more interesting research questions about our data.

The distributions we create often parallel distributions that are mathematically generated. For example, if we obtain the heights of all high school students and plot this data, the graph may resemble a normal distribution, which is generated mathematically. Then, instead of painstakingly collecting heights of all high school students, we could simply use a normal distribution to approximate the heights without sacrificing too much accuracy.

In the study of statistics, we focus on mathematical distributions for the sake of simplicity and relevance to the real-world. Understanding these distributions will enable us to visualize the data easier and build models quicker. However, they cannot and do not replace the work of manual data collection and generating the actual data distribution.

What percentage lie within a certain range? Distributions show what percentage of the data lies within a certain range. So, given a distribution, and a set of values, we can determine the probability that the data will lie within a certain range.

The same data may lead to different conclusions if it is interposed on different distributions. So, it is vital in all statistical analysis for data to be put onto the correct distribution.

19.0.9 Distributions

1. Discrete Distributions[1]
    a) Uniform Distribution[2]
    b) Bernoulli Distribution[3]

[1] Chapter 20 on page 77

[2] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FDiscrete%20Uniform

[3] Chapter 21 on page 79

c) Binomial Distribution[4]
d) Poisson Distribution[5]
e) Geometric Distribution[6]
f) Negative Binomial Distribution[7]
g) Hypergeometric Distribution[8]

2. Continuous Distributions[9]

   a) Uniform Distribution[10]
   b) Exponential Distribution[11]
   c) Gamma Distribution[12]
   d) Normal Distribution[13]
   e) Chi-Square Distribution[14]
   f) Student-t Distribution[15]
   g) F Distribution[16]
   h) Beta Distribution[17]
   i) Weibull Distribution[18]
   j) Gumbel Distribution[19]

[4] Chapter 22 on page 81

[5] Chapter 23 on page 87

[6] Chapter 24 on page 91

[7] Chapter 25 on page 95

[8] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FHypergeometric

[9] Chapter 26 on page 99

[10] Chapter 27 on page 101

[11] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FExponential

[12] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FGamma

[13] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FNormal%20%28Gaussian%29

[14] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FChi-square

[15] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FStudent-t

[16] Chapter 29 on page 105

[17] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FBeta

[18] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FWeibull

[19] http://en.wikibooks.org/wiki/Statistics%2FDistributions%2FGumbel

20 Discrete Distributions

'Discrete' data are data that assume certain discrete and quantized values. For example, true-false answers are discrete, because there are only two possible choices. Valve settings such as 'high/medium/low' can be considered as discrete values. As a general rule, if data can be counted in a practical manner, then they can be considered to be discrete.

To demonstrate this, let us consider the population of the world. It is a discrete number because the number of civilians is theoretically countable. But since this is not practicable, statisticians often treat this data as continuous. That is, we think of population as within a range of numbers rather than a single point.

For the curious, the world population is 6,533,596,139 as of August 9, 2006. Please note that statisticians did not arrive at this figure by counting individual residents. They used much smaller samples of the population to estimate the whole. Going back to Chapter 1, this is a great reason to learn statistics - we need only a smaller sample of data to make intelligent descriptions of the entire population!

Discrete distributions result from plotting the frequency distribution of data which is discrete in nature.

20.1 Cumulative Distribution Function

A discrete random variable has a cumulative distribution function that describes the probability that the random variable is below the point. The cumulative distribution must increase towards 1. Depending on the random variable, it may reach one at a finite number, or it may not. The cdf is represented by a capital F.

20.2 Probability Mass Function

A discrete random variable has a probability mass function that describes how likely the random variable is to be at a certain point. The probability mass function must have a total of 1, and sums to the cdf. The pmf is represented by the lowercase f.

20.3 Special Values

The expected value of a discrete variable is $\sum_{n_{min}}^{n_{max}} x_i f(x_i)$

The expected value of any function of a discrete variable g(X) is $\sum_{n_{min}}^{n_{max}} g(x_i) f(x_i)$

The variance is equal to $E((X-E(X))^2)$

20.4 External Links

Simulating binomial, hypergeometric, and the Poisson distribution: Discrete Distributions[1]

[1] HTTP://WWW.VIAS.ORG/SIMULATIONS/SIMUSOFT_DISCRETEDISTRIS.HTML

21 Bernoulli Distribution

21.1 Bernoulli Distribution: The coin toss

There is no more basic random event than the flipping of a coin. Heads or tails. It's as simple as you can get! The "BERNOULLI TRIAL[1]" refers to a single event which can have one of two possible outcomes with a fixed probability of each occurring. You can describe these events as "yes or no" questions. For example:

- Will the coin land *heads*?
- Will the newborn child be a girl?
- Are a random person's eyes green?
- Will a mosquito die after the area was sprayed with insecticide?
- Will a potential customer decide to buy my product?
- Will a citizen vote for a specific candidate?
- Is an employee going to vote pro-union?
- Will this person be abducted by aliens in their lifetime?

The Bernoulli Distribution has one controlling parameter: the probability of success. A "fair coin" or an experiment where success and failure are equally likely will have a probability of 0.5 (50%). Typically the variable p is used to represent this parameter.

If a random variable X is distributed with a Bernoulli Distribution with a parameter p we write its PROBABILITY MASS FUNCTION[2] as:

$$f(x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \end{cases} \quad 0 \le p \le 1$$

Where the event *X=1* represents the "yes."

This distribution may seem trivial, but it is still a very important building block in probability. The Binomial distribution extends the Bernoulli distribution to encompass multiple "yes" or "no" cases with a fixed probability. Take a close look at the examples cited above. Some similar questions will be presented in the next section which might give an understanding of how these distributions are related.

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/BERNOULLI%20TRIAL

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/PROBABILITY%20MASS%20FUNCTION

21.1.1 Mean

The mean (E[X]) can be derived:

$$\mathrm{E}[X] = \sum_i f(x_i) \cdot x_i$$

$$\mathrm{E}[X] = p \cdot 1 + (1-p) \cdot 0$$

$$\mathrm{E}[X] = p$$

21.1.2 Variance

$$\operatorname{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sum_i f(x_i) \cdot (x_i - \mathrm{E}[X])^2$$

$$\operatorname{Var}(X) = p \cdot (1-p)^2 + (1-p) \cdot (0-p)^2$$

$$\operatorname{Var}(X) = [p(1-p) + p^2](1-p)$$

$$\operatorname{Var}(X) = p(1-p)$$

21.2 External links

- INTERACTIVE BERNOULLI DISTRIBUTION WEB APPLET (JAVA)[3]

[3] HTTP://SOCR.UCLA.EDU/HTMLS/DIST/BERNOULLI_DISTRIBUTION.HTML

22 Binomial Distribution

22.1 Binomial Distribution

Where the Bernoulli Distribution[1] asks the question of "Will this single event succeed?" the Binomial is associated with the question "Out of a given number of trials, how many will succeed?" Some example questions that are modeled with a Binomial distribution are:

- Out of ten tosses, how many times will this coin land *heads*?
- From the children born in a given hospital on a given day, how many of them will be girls?
- How many students in a given classroom will have green eyes?
- How many mosquitos, out of a swarm, will die when sprayed with insecticide?

The relation between the Bernoulli and Binomial distributions is intuitive: The Binomial distribution is composed of multiple Bernoulli trials. We conduct n repeated experiments where the probability of success is given by the parameter p and add up the number of successes. This number of successes is represented by the random variable X. The value of X is then between 0 and n.

When a random variable X has a Binomial Distribution with parameters p and n we write it as X ˜ Bin(n,p) or X ˜ B(n,p) and the probability mass function is given by the equation:

$$P\left[X=k\right]=\begin{cases}\binom{n}{k}p^k\left(1-p\right)^{n-k} & 0\leq k\leq n\\ 0 & \text{otherwise}\end{cases}\quad 0\leq p\leq 1,\quad n\in\mathbb{N}$$

where $\binom{n}{k}=\frac{n!}{k!(n-k)!}$

For a refresher on factorials (n!), go back to the Refresher Course[2] earlier in this wiki book.

22.1.1 An example

Let's walk through a simple example of the Binomial distribution. We're going to use some pretty small numbers because factorials can be hard to compute. (Few basic calculators even feature them!) We are going to ask five random people if they believe there is life on other planets. We are going to assume in this example that we know 30% of people believe

[1] Chapter 21 on page 79

[2] Chapter 1.4.2 on page 9

this to be true. We want to ask the question: "How many people will say they believe in extraterrestrial life?" Actually, we want to be more specific than that: **"What is the probability that exactly 2 people will say they believe in extraterrestrial life?"**

We know all the values that we need to plug into the equation. The number of people asked, n=5. The probability of any given person answering "yes", p=0.3. (Remember, I said that 30% of people believe in life on other planets!) Finally, we're asking for the probability that exactly 2 people answer "yes" so k=2. This yields the equation:

$$P[X=2] = \binom{5}{2} \cdot 0.3^2 \cdot (1-0.3)^3 = 10 \cdot 0.3^2 \cdot (1-0.3)^3 = 0.3087$$

since

$$\binom{5}{2} = \frac{5!}{2! \cdot 3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot (3 \cdot 2 \cdot 1)} = \frac{120}{12} = 10$$

Here are the probabilities for all the possible values of X. You can get these values by replacing the k=2 in the above equation with all values from 0 to 5.

Value for k	Probability f(k)
0	0.16807
1	0.36015
2	0.30870
3	0.13230
4	0.02835
5	0.00243

What can we learn from these results? Well, first of all we'll see that it's just a little more likely that only one person will confess to believing in life on other planets. There's a distinct chance (about 17%) that nobody will believe it, and there's only a 0.24% (a little over 2 in 1000) that all five people will be believers.

22.1.2 Explanation of the equation

Take the above example. Let's consider each of the five people one by one.

The probability that any one person believes in extraterrestrial life is 30%, or 0.3. So the probability that any two people both believe in extraterrestrial life is 0.3 squared. Similarly, the probability that any one person does not believe in extraterrestrial life is 70%, or 0.7, so the probability that any three people do not believe in extraterrestrial life is 0.7 cubed.

Now, for two out of five people to believe in extraterrestrial life, two conditions must be satisfied: two people believe in extraterrestrial life, and three do not. The probability of two out of five people believing in extraterrestrial life would thus appear to be 0.3 squared (two believers) times 0.7 cubed (three non-believers), or 0.03087.

However, in doing this, we are only considering the case whereby the first two selected people are believers. How do we consider cases such as that in which the third and fifth people are believers, which would also mean a total of two believers out of five?

The answer lies in combinatorics. Bearing in mind that the probability that the first two out of five people believe in extraterrestrial life is 0.03087, we note that there are C(5,2), or 10, ways of selecting a set of two people from out of a set of five, i.e. there are ten ways of considering two people out of the five to be the "first two". This is why we multiply by C(n,k). The probability of having any two of the five people be believers is ten times 0.03087, or 0.3087.

22.1.3 Mean

The mean can be derived as follow.

$$\mathrm{E}[X] = \sum_i f(x_i) \cdot x_i = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \cdot x$$

$$\mathrm{E}[X] = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} x$$

$$\mathrm{E}[X] = \frac{n!}{0!(n-0)!} p^0 (1-p)^{n-0} \cdot 0 + \sum_{x=1}^{n} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} x$$

$$\mathrm{E}[X] = 0 + \sum_{x=1}^{n} \frac{n(n-1)!}{x(x-1)!(n-x)!} p \cdot p^{x-1} (1-p)^{n-x} x$$

$$\mathrm{E}[X] = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$$

Now let *w=x-1* and *m=n-1*. We see that *m-w=n-x*. We can now rewrite the summation as

$$\mathrm{E}[X] = np \left[\sum_{w=0}^{m} \frac{m!}{w!(m-w)!} p^w (1-p)^{m-w} \right]$$

We now see that the summation is the sum over the complete pmf of a binomial random variable distributed Bin(m, p). This is equal to 1 (and can be easily verified using the BINOMIAL THEOREM[3]). Therefore, we have

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/BINOMIAL%20THEOREM

$$\mathrm{E}[X] = np\,[1] = np$$

22.1.4 Variance

We derive the variance using the following formula:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

We have already calculated E[X] above, so now we will calculate E[X^2] and then return to this variance formula:

$$\mathrm{E}[X^2] = \sum_i f(x_i)\cdot x^2 = \sum_{x=0}^{n} x^2 \cdot \binom{n}{x} p^x (1-p)^{n-x}.$$

We can use our experience gained above in deriving the mean. We use the same definitions of m and w.

$$\mathrm{E}[X^2] = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x \, (1-p)^{n-x} \, x^2$$

$$\mathrm{E}[X^2] = 0 + \sum_{x=1}^{n} \frac{n!}{x!(n-x)!} p^x \, (1-p)^{n-x} \, x^2$$

$$\mathrm{E}[X^2] = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} \, (1-p)^{n-x} \, x$$

$$\mathrm{E}[X^2] = np \sum_{w=0}^{m} \binom{m}{w} p^w \, (1-p)^{m-w} \, (w+1)$$

$$\mathrm{E}[X^2] = np \left[\sum_{w=0}^{m} \binom{m}{w} p^w \, (1-p)^{m-w} \, w + \sum_{w=0}^{m} \binom{m}{w} p^w \, (1-p)^{m-w} \right]$$

The first sum is identical in form to the one we calculated in the Mean (above). It sums to mp. The second sum is 1.

$$\mathrm{E}[X^2] = np \cdot (mp+1) = np((n-1)p+1) = np(np-p+1).$$

Using this result in the expression for the variance, along with the Mean ($\mathrm{E}(X) = np$), we get

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = np(np-p+1) - (np)^2 = np(1-p).$$

22.2 External links

- Interactive Binomial Distribution Web Applet (Java)[4]

[4] http://socr.ucla.edu/htmls/dist/Binomial_Distribution.html

23 Poisson Distribution

23.1 Poisson Distribution

Any French speaker will notice that "Poisson" means "fish", but really there's nothing fishy about this distribution. It's actually pretty straightforward. The name comes from the mathematician SIMÉON-DENIS POISSON[1] (1781-1840).

The Poisson Distribution is *very similar* to the BINOMIAL DISTRIBUTION[2]. We are examining the number of times an event happens. The difference is subtle. Whereas the Binomial Distribution looks at how many times we register a success over a fixed total number of trials, the Poisson Distribution measures how many times a discrete event occurs, over a period of continuous space or time. There isn't a "total" value n. As with the previous sections, let's examine a couple of experiments or questions that might have an underlying Poisson nature.

- How many pennies will I encounter on my walk home?
- How many children will be delivered at the hospital today?
- How many mosquito bites did you get today after having sprayed with insecticide?
- How many angry phone calls did I get after airing a particularly distasteful political ad?
- How many products will I sell after airing a new television commercial?
- How many people, per hour, will cross a picket line into my store?
- How many alien abduction reports will be filed this year?
- How many defects will there be per 100 metres of rope sold?

What's a little different about this distribution is that the random variable X which counts the number of events can take on *any non-negative integer* value. In other words, I could walk home and find no pennies on the street. I could also find one penny. It's also possible (although unlikely, short of an armored-car exploding nearby) that I would find 10 or 100 or 10,000 pennies.

Instead of having a parameter p that represents a component probability like in the Bernoulli and Binomial distributions, this time we have the parameter "lambda" or λ which represents the "average or expected" number of events to happen within our experiment. The probability mass function of the Poisson is given by

$$P(N=k)=\frac{e^{-\lambda}\lambda^k}{k!}$$

.

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/SIMEON_POISSON

[2] Chapter 22 on page 81

23.1.1 An example

We run a restaurant and our signature dish (which is very expensive) gets ordered *on average* 4 times per day. What is the probability of having this dish ordered exactly 3 times tomorrow? If we only have the ingredients to prepare 3 of these dishes, what is the probability that it will get sold out and we'll have to turn some orders away?

The probability of having the dish ordered 3 times exactly is given if we set k=3 in the above equation. Remember that we've already determined that we sell *on average* 4 dishes per day, so λ=4.

$$P(N=k)=\frac{e^{-\lambda}\lambda^k}{k!}=\frac{e^{-4}4^3}{3!}=0.195$$

Here's a table of the probabilities for all values from k=0..6:

Value for k	Probability f(k)
0	0.0183
1	0.0733
2	0.1465
3	0.1954
4	0.1954
5	0.1563
6	0.1042

Now for the big question: Will we run out of food by the end of the day tomorrow? In other words, we're asking if the random variable X>3. In order to compute this we would have to add the probabilities that X=4, X=5, X=6,... all the way to infinity! But wait, there's a better way!

The probability that we run out of food P(X>3) is the same as 1 minus the probability that we *don't* run out of food, or 1-P(X≤3). So if we total the probability that we sell zero, one, two and three dishes and subtract that from 1, we'll have our answer. So,

1 - P(X≤3) = 1 - (P(X=0) + P(X=1) + P(X=2) + P(X=3)) = 1 - 0.4335 = 0.5665

In other words, we have a 56.65% chance of selling out of our wonderful signature dish. I guess crossing our fingers is in order!

DE:MATHEMATIK: STATISTIK: POISSONVERTEILUNG[3]

23.1.2 Mean

We calculate the mean as follows:

[3] HTTP://DE.WIKIBOOKS.ORG/WIKI/MATHEMATIK%3A%20STATISTIK%3A%20POISSONVERTEILUNG

$$\mathrm{E}[X] = \sum_i f(x_i) \cdot x_i = \sum_{x=0} \frac{e^{-\lambda}\lambda^x}{x!} x$$

$$\mathrm{E}[X] = \frac{e^{-\lambda}\lambda^0}{0!} \cdot 0 + \sum_{x=1} \frac{e^{-\lambda}\lambda^x}{x!} x$$

$$\mathrm{E}[X] = 0 + e^{-\lambda} \sum_{x=1} \frac{\lambda\lambda^{x-1}}{(x-1)!}$$

$$\mathrm{E}[X] = \lambda e^{-\lambda} \sum_{x=1} \frac{\lambda^{x-1}}{(x-1)!}$$

$$\mathrm{E}[X] = \lambda e^{-\lambda} \sum_{x=0} \frac{\lambda^x}{x!}$$

REMEMBER[4] that $\mathrm{e}^{\lambda} = \sum_{x=0} \frac{\lambda^x}{x!}$

$$\mathrm{E}[X] = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

23.1.3 Variance

We derive the variance using the following formula:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

We have already calculated E[X] above, so now we will calculate E[X^2] and then return to this variance formula:

$$\mathrm{E}[X^2] = \sum_i f(x_i) \cdot x^2$$

$$\mathrm{E}[X^2] = \sum_{x=0} \frac{e^{-\lambda}\lambda^x}{x!} x^2$$

[4] HTTP://EN.WIKIPEDIA.ORG/WIKI/TAYLOR_SERIES%23LIST_OF_MACLAURIN_SERIES_OF_SOME_COMMON_FUNCTIONS

$$\mathrm{E}[X^2] = 0 + \sum_{x=1} \frac{e^{-\lambda}\lambda\lambda^{x-1}}{(x-1)!}x$$

$$\mathrm{E}[X^2] = \lambda \sum_{x=0} \frac{e^{-\lambda}\lambda^{x}}{x!}(x+1)$$

$$\mathrm{E}[X^2] = \lambda \left[\sum_{x=0} \frac{e^{-\lambda}\lambda^{x}}{x!}x + \sum_{x=0} \frac{e^{-\lambda}\lambda^{x}}{x!} \right]$$

The first sum is E[X]=λ and the second we also calculated above to be 1.

$$\mathrm{E}[X^2] = \lambda[\lambda+1] = \lambda^2 + \lambda$$

Returning to the variance formula we find that

$$\mathrm{Var}[X] = (\lambda^2+\lambda) - (\lambda)^2 = \lambda$$

23.2 External links

- Interactive Poisson Distribution Web Applet (Java)[5]

[5] http://socr.ucla.edu/htmls/dist/Poisson_Distribution.html

24 Geometric Distribution

24.1 Geometric distribution

There are two similar distributions with the name "Geometric Distribution".

- The probability distribution of the number X of BERNOULLI TRIAL[1]s needed to get one success, supported on the set { 1, 2, 3, ...}
- The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set { 0, 1, 2, 3, ... }

These two different geometric distributions should not be confused with each other. Often, the name *shifted* geometric distribution is adopted for the former one. We will use X and Y to refer to distinguish the two.

24.1.1 Shifted

The shifted Geometric Distribution refers to the probability of the number of times needed to do something until getting a desired result. For example:

- How many times will I throw a coin until it lands on *heads*?
- How many children will I have until I get a girl?
- How many cards will I draw from a pack until I get a Joker?

Just like the BERNOULLI DISTRIBUTION[2], the Geometric distribution has one controlling parameter: The probability of success in any independent test.

If a random variable X is distributed with a Geometric Distribution with a parameter p we write its PROBABILITY MASS FUNCTION[3] as:

$P(X = i) = p(1-p)^{i-1}$

With a Geometric Distribution it is also pretty easy to calculate the probability of a "more than n times" case. The probability of failing to achieve the wanted result is $(1-p)^k$.

Example: a student comes home from a party in the forest, in which INTERESTING SUBSTANCES[4] were consumed. The student is trying to find the key to his front door, out of a keychain with 10 different keys. What is the probability of the student succeeding in finding the right key in the 4th attempt?

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/BERNOULLI%20TRIAL
[2] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ADISTRIBUTIONS%2FBERNOULLI
[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/PROBABILITY%20MASS%20FUNCTION
[4] HTTP://EN.WIKIPEDIA.ORG/WIKI/CANNABIS

$P(X=4)=\frac{1}{10}\left(1-\frac{1}{10}\right)^{4-1}=\frac{1}{10}\left(\frac{9}{10}\right)^{3}=0.0729$

24.1.2 Unshifted

The probability mass function is defined as:

$$f(x)=p(1-p)^x$$

for

$$x\in\{0,1,2,\}$$

Mean

$$\mathrm{E}[X]=\sum_i f(x_i)x_i=\sum_0 p(1-p)^x x$$

Let *q=1-p*

$$\mathrm{E}[X]=\sum_0 (1-q)q^x x$$

$$\mathrm{E}[X]=\sum_0 (1-q)qq^{x-1}x$$

$$\mathrm{E}[X]=(1-q)q\sum_0 q^{x-1}x$$

$$\mathrm{E}[X]=(1-q)q\sum_0 \frac{d}{dq}q^x$$

We can now interchange the derivative and the sum.

$$\mathrm{E}[X]=(1-q)q\frac{d}{dq}\sum_0 q^x$$

$$\mathrm{E}[X]=(1-q)q\frac{d}{dq}\frac{1}{1-q}$$

$$\mathrm{E}[X] = (1-q)q\frac{1}{(1-q)^2}$$

$$\mathrm{E}[X] = q\frac{1}{(1-q)}$$

$$\mathrm{E}[X] = \frac{(1-p)}{p}$$

Variance

We derive the variance using the following formula:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

We have already calculated E[X] above, so now we will calculate E[X^2] and then return to this variance formula:

$$\mathrm{E}[X^2] = \sum_i f(x_i)\cdot x^2$$

$$\mathrm{E}[X^2] = \sum_0 p(1-p)^x x^2$$

Let *q=1-p*

$$\mathrm{E}[X^2] = \sum_0 (1-q)q^x x^2$$

We now manipulate x^2 so that we get forms that are easy to handle by the technique used when deriving the mean.

$$\mathrm{E}[X^2] = (1-q)\sum_0 q^x[(x^2-x)+x]$$

$$\mathrm{E}[X^2] = (1-q)\left[\sum_0 q^x(x^2-x)+\sum_0 q^x x\right]$$

$$\mathrm{E}[X^2] = (1-q)\left[q^2 \sum_0 q^{x-2}x(x-1) + q\sum_0 q^{x-1}x\right]$$

$$\mathrm{E}[X^2] = (1-q)q\left[q\sum_0 \frac{d^2}{(dq)^2}q^x + \sum_0 \frac{d}{dq}q^x\right]$$

$$\mathrm{E}[X^2] = (1-q)q\left[q\frac{d^2}{(dq)^2}\sum_0 q^x + \frac{d}{dq}\sum_0 q^x\right]$$

$$\mathrm{E}[X^2] = (1-q)q\left[q\frac{d^2}{(dq)^2}\frac{1}{1-q} + \frac{d}{dq}\frac{1}{1-q}\right]$$

$$\mathrm{E}[X^2] = (1-q)q\left[q\frac{2}{(1-q)^3} + \frac{1}{(1-q)^2}\right]$$

$$\mathrm{E}[X^2] = \frac{2q^2}{(1-q)^2} + \frac{q}{(1-q)}$$

$$\mathrm{E}[X^2] = \frac{2q^2 + q(1-q)}{(1-q)^2}$$

$$\mathrm{E}[X^2] = \frac{q(q+1)}{(1-q)^2}$$

$$\mathrm{E}[X^2] = \frac{(1-p)(2-p)}{p^2}$$

We then return to the variance formula

$$\mathrm{Var}[X] = \left[\frac{(1-p)(2-p)}{p^2}\right] - \left(\frac{1-p}{p}\right)^2$$

$$\mathrm{Var}[X] = \frac{(1-p)}{p^2}$$

24.2 External links

- Interactive Geometric Distribution Web Applet (Java)[5]

[5] http://socr.ucla.edu/htmls/dist/Geoemtric_Distribution.html

25 Negative Binomial Distribution

25.1 Negative Binomial Distribution

Just as the Bernoulli and the Binomial distribution are related in counting the number of successes in 1 or more trials, the Geometric and the Negative Binomial distribution are related in the number of trials needed to get 1 or more successes.

The Negative Binomial distribution refers to the probability of the number of times needed to do something until achieving a fixed number of desired results. For example:

- How many times will I throw a coin until it lands on *heads* for the 10th time?
- How many children will I have when I get my third daughter?
- How many cards will I have to draw from a pack until I get the second Joker?

Just like the BINOMIAL DISTRIBUTION[1], the Negative Binomial distribution has two controlling parameters: the probability of success p in any independent test and the desired number of successes m. If a random variable X has Negative Binomial distribution with parameters p and m, its PROBABILITY MASS FUNCTION[2] is:

$$P(X=n) = \binom{n-1}{m-1} p^m (1-p)^{n-m}, \text{ for } n \geq m$$

.

25.1.1 Example

A travelling salesman goes home if he has sold 3 encyclopedias that day. Some days he sells them quickly. Other days he's out till late in the evening. If on the average he sells an encyclopedia at one out of ten houses he approaches, what is the probability of returning home after having visited only 10 houses?

Answer:

The number of trials X is Negative Binomial distributed with parameters p=0.1 and m=3, hence:

[1] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ADISTRIBUTIONS%2FBINOMIAL

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/PROBABILITY%20MASS%20FUNCTION

$$P(X=10)=\binom{9}{2}0.1^3 0.9^7=0.0172186884$$

.

25.1.2 Mean

The mean can be derived as follows.

$$\mathrm{E}[X]=\sum_i f(x_i)\cdot x_i=\sum_{x=0}^{\binom{x+r-1}{r-1}} p^x(1-p)^r\cdot x$$

$$\mathrm{E}[X]=\binom{0+r-1}{r-1}p^0\,(1-p)^r\cdot 0+\sum_{x=1}^{\binom{x+r-1}{r-1}} p^x(1-p)^r\cdot x$$

$$\mathrm{E}[X]=0+\sum_{x=1}^{\frac{(x+r-1)!}{(r-1)!x!}} p^x(1-p)^r\cdot x$$

$$\mathrm{E}[X]=\frac{rp}{1-p}\sum_{x=1}^{\frac{(x+r-1)!}{r!(x-1)!}} p^{x-1}(1-p)^{r+1}$$

Now let s = $r+1$ and $w=x-1$ inside the summation.

$$\mathrm{E}[X]=\frac{rp}{1-p}\sum_{w=0}^{\frac{(w+s-1)!}{(s-1)!w!}} p^w(1-p)^s$$

$$\mathrm{E}[X]=\frac{rp}{1-p}\sum_{w=0}^{\binom{w+s-1}{s-1}} p^w(1-p)^s$$

We see that the summation is the sum over a the complete pmf of a negative binomial random variable distributed NB(s,p), which is 1 (and can be verified by applying NEWTON'S GENERALIZED BINOMIAL THEOREM[3]).

$$\mathrm{E}[X]=\frac{rp}{1-p}$$

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/BINOMIAL_THEOREM%23NEWTON.27S_GENERALIZED_BINOMIAL_THEOREM

25.1.3 Variance

We derive the variance using the following formula:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

We have already calculated E[X] above, so now we will calculate E[X^2] and then return to this variance formula:

$$\mathrm{E}[X^2] = \sum_i f(x_i) \cdot x^2 = \sum_{x=0}^{\binom{x+r-1}{r-1}} p^x (1-p)^r \cdot x^2$$

$$\mathrm{E}[X^2] = 0 + \sum_{x=1}^{\binom{x+r-1}{r-1}} p^x (1-p)^r x^2$$

$$\mathrm{E}[X^2] = \sum_{x=1}^{\frac{(x+r-1)!}{(r-1)!x!}} p^x (1-p)^r x^2$$

$$\mathrm{E}[X^2] = \frac{rp}{1-p} \sum_{x=1}^{\frac{(x+r-1)!}{r!(x-1)!}} p^{x-1} (1-p)^{r+1} x$$

Again, let let $s = r{+}1$ and $w{=}x{-}1$.

$$\mathrm{E}[X^2] = \frac{rp}{1-p} \sum_{w=0}^{\frac{(w+s-1)!}{(s-1)!w!}} p^w (1-p)^s (w+1)$$

$$\mathrm{E}[X^2] = \frac{rp}{1-p} \sum_{w=0}^{\binom{w+s-1}{s-1}} p^w (1-p)^s (w+1)$$

$$\mathrm{E}[X^2] = \frac{rp}{1-p} \left[\sum_{w=0}^{\binom{w+s-1}{s-1}} p^w (1-p)^s w + \sum_{w=0}^{\binom{w+s-1}{s-1}} p^w (1-p)^s \right]$$

The first summation is the mean of a negative binomial random variable distributed NB(s,p) and the second summation is the complete sum of that variable's pmf.

$$\mathrm{E}[X^2] = \frac{rp}{1-p}\left[\frac{sp}{1-p}+1\right]$$

$$\mathrm{E}[X^2] = \frac{rp(1+rp)}{(1-p)^2}$$

We now insert values into the original variance formula.

$$\mathrm{Var}[X] = \frac{rp(1+rp)}{(1-p)^2} - \left(\frac{rp}{1-p}\right)^2$$

$$\mathrm{Var}[X] = \frac{rp}{(1-p)^2}$$

25.2 External links

- Interactive Negative Binomial Distribution Web Applet (Java)[4]

[4] http://socr.ucla.edu/htmls/dist/Negative_Binomial_Distribution.html

26 Continuous Distributions

A continuous statistic is a random variable that does not have any points at which there is any distinct probability that the variable will be the corresponding number.

26.1 Cumulative Distribution Function

A continuous random variable, like a discrete random variable, has a cumulative distribution function. Like the one for a discrete random variable, it also increases towards 1. Depending on the random variable, it may reach one at a finite number, or it may not. The cdf is represented by a capital F.

26.2 Probability Distribution Function

Unlike a discrete random variable, a continuous random variable has a probability density function instead of a probability mass function. The difference is that the former must integrate to 1, while the latter must have a total value of 1. The two are very similar, otherwise. The pdf is represented by a lowercase f.

26.3 Special Values

The expected value for a continuous variable is defined as $\int_{-\infty}^{\infty} x f(x)\, dx$

The expected value of any function of a continuous variable g(x) is defined as $\int_{-\infty}^{\infty} g(x) f(x)\, dx$

The mean of a continuous or discrete distribution is defined as E[X]

The variance of a continuous or discrete distribution is defined as E[(X-E[X]^{2})]

Expectations can also be derived by producing the Moment Generating Function for the distribution in question. This is done by finding the expected value E[e$^{\mathrm{tX}}$]. Once the Moment Generating Function has been created, each derivative of the function gives a different piece of information about the distribution function.

d^{1}x/d^{1}y = mean

d^{2}x/d^{2}y = variance

d^3x/d^3y = skewness

d^4x/d^4y = kurtosis

27 Uniform Distribution

27.1 Continuous Uniform Distribution

The (continuous) uniform distribution, as its name suggests, is a distribution with probability densities that are the same at each point in an interval. In casual terms, the uniform distribution shapes like a rectangle.

Mathematically speaking, the probability density function of the uniform distribution is defined as

$f(x) = \begin{cases} \frac{1}{b-a} \ \forall \ real \ x \ \in [a,b] \end{cases}$

And the cumulative distribution function is:

$F(x) = \begin{cases} 0, & \text{if } x \leq a \\ \frac{x-a}{b-a}, & \text{if } a < x < b \\ 1, & \text{if } x \geq b \end{cases}$

27.1.1 Mean

We derive the mean as follows.

$$\mathrm{E}[X] = \int^{-f(x)\cdot xdx}$$

As the uniform distribution is 0 everywhere but $[a,\ b]$ we can restrict ourselves that interval

$$\mathrm{E}[X] = \int_a^b \frac{1}{b-a} xdx$$

$$\mathrm{E}[X] = \frac{1}{(b-a)} \frac{1}{2} x^2 \bigg|_a^b$$

$$\mathrm{E}[X] = \frac{1}{2(b-a)} \left[b^2 - a^2 \right]$$

$$\mathrm{E}[X] = \frac{b+a}{2}$$

27.1.2 Variance

We use the following formula for the variance.

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

$$\mathrm{Var}(X) = \left[\int^{-f(x)\cdot x^2 dx}\right] - \left(\frac{b+a}{2}\right)^2$$

$$\mathrm{Var}(X) = \left[\int_a^b \frac{1}{b-a} x^2 dx\right] - \frac{(b+a)^2}{4}$$

$$\mathrm{Var}(X) = \frac{1}{b-a}\frac{1}{3}x^3\Big|_a^b - \frac{(b+a)^2}{4}$$

$$\mathrm{Var}(X) = \frac{1}{3(b-a)}[b^3 - a^3] - \frac{(b+a)^2}{4}$$

$$\mathrm{Var}(X) = \frac{4(b^3-a^3) - 3(b+a)^2(b-a)}{12(b-a)}$$

$$\mathrm{Var}(X) = \frac{(b-a)^3}{12(b-a)}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

27.2 External links

- Interactive Uniform Distribution Web Applet (Java)[1]

[1] http://socr.ucla.edu/htmls/dist/ContinuousUniform_Distribution.html

28 Normal Distribution

The Normal Probability Distribution is one of the most useful and more important distributions in statistics. It is a continuous variable distribution. Although the mathematics of this distribution can be quite off putting for students of a first course in statistics it can nevertheless be usefully applied with out over complication.

The Normal distribution is used frequently in statistics for many reasons:

1) The Normal distribution has many convenient mathematical properties.

2) Many natural phenomena have distributions which when studied have been shown to be close to that of the Normal Distribution.

3) The Central Limit Theorem shows that the Normal Distribution is a suitable model for large samples regardless of the actual distribution.

28.1 Mathematical Characteristics of the Normal Distribution

A continuous random variable , X, is normally distributed with a probability density function :

$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

29 F Distribution

Named after Sir Ronald Fisher, who developed the F distribution for use in determining ANOVA critical values. The cutoff values in an F table are found using three variables- ANOVA numerator degrees of freedom, ANOVA denominator degrees of freedom, and significance level.

ANOVA is an abbreviation of analysis of variance. It compares the size of the variance between two different samples. This is done by dividing the larger variance over the smaller variance. The formula of the F statistic is:

$F(r_1,r_2) = \frac{\chi^2_{r1}/r_1}{\chi^2_{r2}/r_2}$

where χ^2_{r1} and χ^2_{r2} are the chi-square statistics of sample one and two respectively, and $r1$and $r2$ are their degrees of freedom, i.e. the number of observations.

One example could be if you want to compare apples that look alike but are from different trees and have different sizes. You want to investigate whether they have the same variance of the weight on average.

There are three apples from the first tree that weigh 110, 121 and 143 grams respectively, and four from the other which weigh 88, 93, 105 and 124 grams respectively. The mean and variance of the first sample are 124.67 and 16.80 respectively, and of the second sample 102.50 and 16.01. The chi-square statistic of the first sample is

$\frac{110-124.67}{16.80^2} + \frac{121-124.67}{16.80^2} + \frac{143-124.67}{16.80^2} = 2.00,$

and for the second sample

$\frac{88-102.50}{16.01^2} + \frac{93-102.50}{16.01^2} + \frac{105-102.50}{16.01^2} + \frac{124-102.50}{16.01^2} = 3.00.$

The F statistic is now F = $\frac{3/4}{2/3} = 1.125$. The Chi-square statistic divided by degrees of freedom appears on the nominator for the second sample because it was larger than that of the first sample.

The critical value of the F distribution for 4 degrees of freedom. in the nominator and 3 degrees of freedom in the denominator, i.e. F(f1=4, f2=3) is 9.12 at a 5% level of confidence. Since the test statistic 1.125 is smaller than the critical value, we cannot reject the null hypothesis that they have the same variance. The conclusion is that they have the same variance.

29.1 External links

- Interactive F Distribution Web Applet (Java)[1]

[1] http://socr.ucla.edu/htmls/dist/Fisher_Distribution.html

30 Testing Statistical Hypothesis

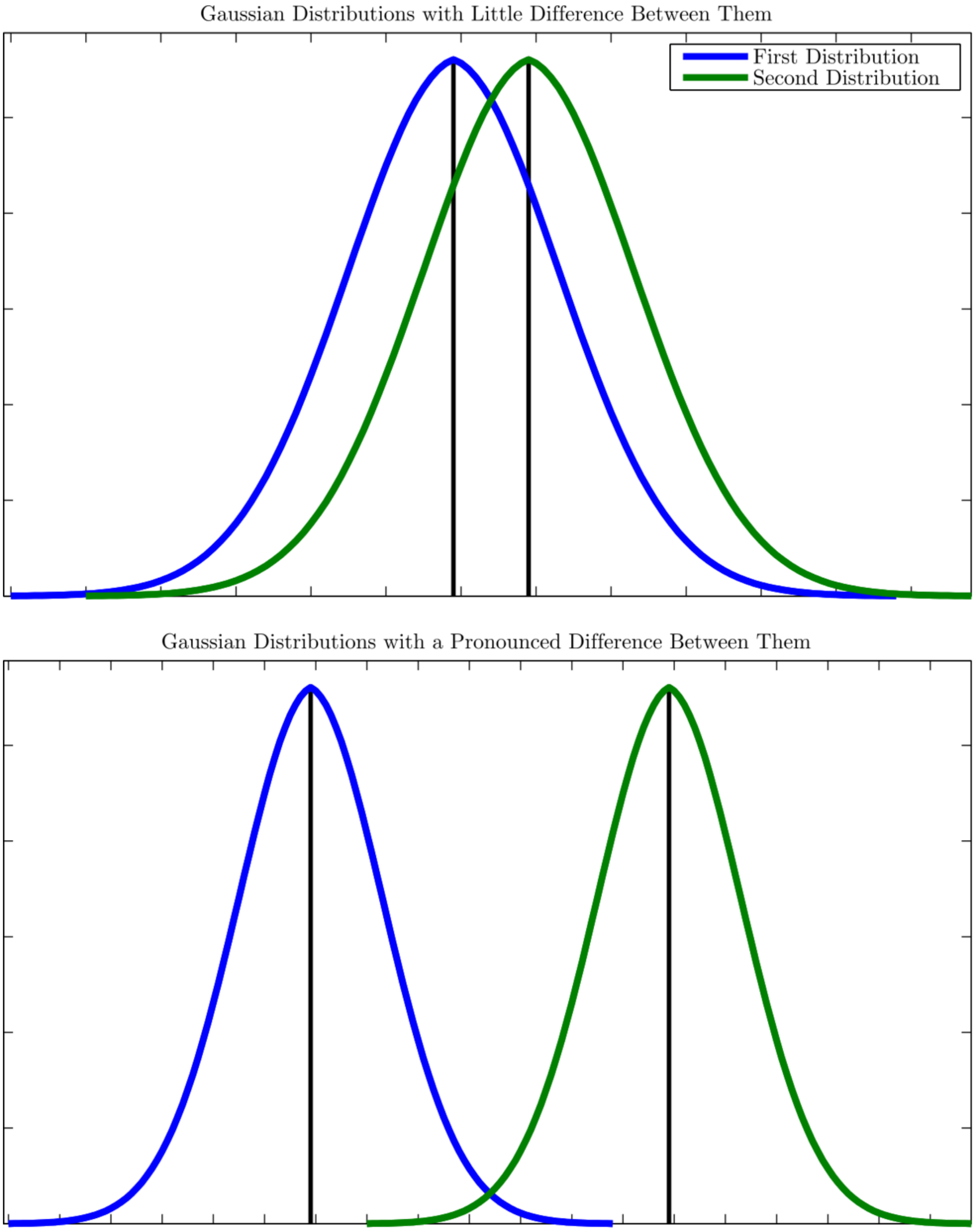


Figure 17: Two examples of how the means of two distributions may be different, leading to two different statistical hypotheses

There are many different tests for the many different kinds of data. A way to get started is to understand what kind of data you have. Are the variables quantitative or qualitative? Certain tests are for certain types of data depending on the size, distribution or scale. Also, it is important to understand how samples of data can differ. The 3 primary characteristics of quantitative data are: central tendency, spread, and shape.

When most people "test" quantitative data, they tend to do tests for central tendency. Why? Well, let's say you had 2 sets of data and you wanted to see if they were different from each other. One way to test this would be to test to see if their central tendency (their means for example) differ.

Imagine two symmetric, bell shaped curves with a vertical line drawn directly in the middle of each, as shown here. If one sample was a lot different than another (a lot higher in values,etc.) then the means would be different typically. So when testing to see if two samples are different, usually two means are compared.

Two medians (another measure of central tendency) can be compared also. Or perhaps one wishes to test two samples to see if they have the same spread or variation. Because statistics of central tendency, spread, etc. follow different distributions - different testing procedures must be followed and utilized.

In the end, most folks summarize the result of a hypothesis test into one particular value - the p-value. If the p-value is smaller than the level of significance (usually $\alpha = 5\%$, but even lower in other fields of science i.e. Medicine) then the zero-hypothesis rejected and the alternative hypothesis accepted. The p-value is actually the probability of making a statistical error. If the p-value is higher than the level of significance you accept the zero-hypothesis and reject the alternative hypothesis, however that does not necessarily mean that the zero-hypothesis is correct.

31 Purpose of Statistical Tests

31.1 Purpose of Statistical Tests

In general, the purpose of statistical tests is to determine whether some hypothesis is extremely unlikely given observed data.

There are two common philosophical approaches to such tests, *significance testing* (due to Fisher) and *hypothesis testing* (due to Neyman and Pearson).

Significance testing aims to quantify evidence against a particular hypothesis being true. We can think of it as testing to guide research. We believe a certain statement may be true and want to work out whether it is worth investing time investigating it. Therefore, we look at the opposite of this statement. If it is quite likely then further study would seem to not make sense. However if it is extremely unlikely then further study would make sense.

A concrete example of this might be in drugs testing. We have a number of drugs that we want to test and only limited time, so we look at the hypothesis that an individual drug has no positive effect whatsoever, and only look further if this is unlikley.

Hypothesis testing rather looks at evidence for a particular hypothesis being true. We can think of this as a guide to making a decision. We need to make a decision soon, and suspect that a given statement is true. Thus we see how unlikely we are to be wrong, and if we are sufficiently unlikely to be wrong we can assume that this statement is true. Often this decision is final and cannot be changed.

Statisticians often overlook these differences and incorrectly treat the terms "significance test" and "hypothesis test" as though they are interchangeable.

A data analyst frequently wants to know whether there is a difference between two sets of data, and whether that difference is likely to occur due to random fluctuations, or is instead unusual enough that random fluctuations rarely cause such differences.

In particular, frequently we wish to know something about the average (or mean), or about the variability (as measured by variance or standard deviation).

Statistical tests are carried out by first making some assumption, called the Null Hypothesis, and then determining whether the data observed is unlikely to occur given that assumption. If the probability of seeing the observed data is small enough under the assumed Null Hypothesis, then the Null Hypothesis is rejected.

A simple example might help. We wish to determine if men and women are the same height on average. We select and measure 20 women and 20 men. We assume the Null Hypothesis that there is no difference between the average value of heights for men vs. women. We can then test using the T-TEST[1] to determine whether our sample of 40 heights would be unlikely to occur given this assumption. The basic idea is to assume heights are normally distributed, and to assume that the means and standard deviations are the same for women and for men. Then we calculate the average of our 20 men, and of our 20 women, we also calculate the sample standard deviation for each. Then using the t-test of two means with 40-2 = 38 degrees of freedom we can determine whether the difference in heights between the sample of men and the sample of women is sufficiently large to make it unlikely that they both came from the same normal population.

[1] Chapter 36 on page 127

32 Different Types of Tests

A statistical test is always about one or more parameters of the concerned population (distribution). The appropiate test depends on the type of null and alternative hypothesis about this (these) parameter(s) and the available information from the sample.

32.1 Example

It is conjectured that British children gain more weight lately. Hence the population mean μ of the weight X of children of let's say 12 years of age is the parameter at stake. In the recent past the mean weight of this group of children turned out to be 45 kg. Hence the null hypothesis (of no change) is:

$$H_0 : \mu = 45$$

.

As we suspect a gain in weight, the alternative hypothesis is:

$$H_1 : \mu > 45$$

.

A random sample of 100 children shows an average weight of 47 kg with a standard deviation of 8 kg.

Because it is reasonable to assume that the weights are normally distributed, the appropriate test will be a t-test, with test statistic:

$$T = \frac{\bar{X} - 45}{S}\sqrt{100}$$

.

Under the null hypothesis T will be Student distributed with 99 degrees of freedom, which means approximately standard normally distributed.

The null hypothesis will be rejected for large values of T. For this sample the value t of T is:

$$t = \frac{47 - 45}{8}\sqrt{100} = 2.5$$

.

Is this a large value? That depends partly on our demands. The so called p-value of the observed value t is:

$$p = P(T \geq t; H_0) = P(T \geq 2.5; H_0) \approx P(Z \geq 2.5) < 0.01$$

,

in which Z stands for a standard normally distributed random variable.

If we are not too critical this is small enough, so reason to reject the null hypothesis and to assume our conjecture to be true.

Now suppose we have lost the individual data, but still know that the maximum weight in the sample was 68 kg. It is not possible then to use the t-test, and instead we have to use a test based on the statistic max(X).

It might also be the case that our assumption on the distribution of the weight is questionable. To avoid discussion we may use a distribution free test instead of a t-test.

A statistical test begins with a hypothesis; the form of that hypothesis determines the type(s) of test(s) that can be used. In some cases, only one is appropriate; in others, one may have some choice.

For example: if the hypothesis concerns the value of a single population mean (μ), then a one sample test for mean is indicated. Whether the z-test or t-test should be used depends on other factors (each test has its own requirements).

A complete listing of the conditions under which each type of test is indicated is probably beyond the scope of this work; refer to the sections for the various types of tests for more information about the indications and requirements for each test.

33 z Test for a Single Mean

The Null Hypothesis should be an assumption concerning the value of the population mean. The data should consist of a single sample of quantitative data from the population.

33.1 Requirements

The sample should be drawn from a population from which the Standard Deviation (or Variance) is known. Also, the measured variable (typically listed as $x - \bar{x}$ is the *sample statistic*) should have a Normal Distribution.

Note that if the distribution of the variable in the population is non-normal (or unknown), the z-test can still be used for approximate results, provided the sample size is sufficiently large. Historically, sample sizes of at least 30 have been considered sufficiently large; reality is (of course) much more complicated, but this rule of thumb is still in use in many textbooks.

If the population Standard Deviation is unknown, then a z-test is typically not appropriate. However, when the sample size is large, the sample standard deviation can be used as an estimate of the population standard deviation, and a z-test can provide approximate results.

33.2 Definitions of Terms

$$\mu;$$

= Population Mean

$$\sigma_x$$

= Population Standard Deviation

$$\bar{x}$$

= Sample Mean

$$\sigma_{\bar{x}}$$

= Sample Standard Deviation

$$N$$

= Sample Population

33.3 Procedure

- The Null Hypothesis:

This is a statement of *no change* or *no effect;* often, we are looking for evidence that this statement is *no longer true.*

$\mathrm{H}_0 : \mu = \mu_0$

- The Alternate Hypothesis:

This is a statement of inequality; we are looking for evidence that this statement *is* true.

$\mathrm{H}_1 : \mu < \mu_0$ or

$\mathrm{H}_1 : \mu > \mu_0$ or

$\mathrm{H}_1 : \mu \neq \mu_0$

- The Test Statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- The Significance (p-value)

Calculate the probability of observing a value of z (from a Standard Normal Distribution) using the Alternate Hypothesis to indicate the direction in which the area under the Probability Density Function is to be calculated. This is the Attained Significance, or p-value.

Note that some (older) methods first chose a Level Of Significance, which was then translated into a value of z. This made more sense (and was easier!) in the days before computers and graphics calculators.

- Decision

The Attained Significance represents the probability of obtaining a test statistic as extreme, or more extreme, than ours - if the null hypothesis is true.

If the Attained Significance (p-value) is sufficiently low, then this indicates that our test statistic is unusual (rare) - we usually take this as evidence that the null hypothesis is in error. In this case, we *reject the null hypothesis.*

If the p-value is large, then this indicates that the test statistic is usual (common) - we take this as a lack of evidence against the null hypothesis. In this case, we *fail to reject the null hypothesis.*

It is common to use 5% as the dividing line between the common and the unusual; again, reality is more complicated. Sometimes a lower level of uncertainty must be chosen should the consequences of error results in a decision that can injure or kill people or do great economic harm. We would more likely tolerate a drug that kills 5% of patients with a terminal cancer but cures 95% of all patients, but we would hardly tolerate a cosmetic that disfigures 5% of those who use it.

33.4 Worked Examples

33.4.1 Are The Kids Above Average?

Scores on a certain test of mathematical aptitude have mean $\mu = 50$ and standard deviation $\sigma = 10$. An amateur researcher believes that the students in his area are brighter than average, and wants to test his theory.

The researcher has obtained a random sample of 45 scores for students in his area. The mean score for this sample is 52.

Does the researcher have evidence to support his belief?

The null hypothesis is that *there is no difference,* and that the students in his area are no different than those in the general population; thus,

$$\mathrm{H}_0 : \mu = 50$$

(where μ represents the mean score for students in his area)

He is looking for evidence that the students in his area are above average; thus, the alternate hypothesis is

$$\mathrm{H}_1 : \mu > 50$$

Since the hypothesis concerns a single population mean, a z-test is indicated. The sample size is fairly large (greater than 30), and the standard deviation is known, so a z-test is appropriate.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{52 - 50}{10/\sqrt{45}} = 1.3416$$

We now find the area under the Normal Distribution to the right of z = 1.3416 (*to the right, since the alternate hypothesis is to the right*). This can be done with a table of values, or software- I get a value of 0.0899.

If the null hypothesis is true (and these students are no better than the general population), then the probability of obtaining a sample mean of 52 or higher is 8.99%. This occurs fairly frequently (using the 5% rule), so it does not seem unusual. I fail to reject the null hypothesis (at the 5% level).

It appears that the evidence *does not* support the researcher's belief.

33.4.2 Is The Machine Working Correctly?

Sue is in charge of Quality Control at a bottling facility. Currently, she is checking the operation of a machine that is supposed to deliver 355 mL of liquid into an aluminum can. If the machine delivers too little, then the local Regulatory Agency may fine the company. If the machine delivers too much, then the company may lose money. For these reasons, Sue is looking for any evidence that the amount delivered by the machine is different from 355 mL.

During her investigation, Sue obtains a random sample of 10 cans, and measures the following volumes:

355.02 355.47 353.01 355.93 356.66 355.98 353.74 354.96 353.81 355.79

The machine's specifications claim that the amount of liquid delivered varies according to a normal distribution, with mean $\mu = 355$ mL and standard deviation $\sigma = 0.05$ mL.

Do the data suggest that the machine is operating correctly?

The null hypothesis is that the machine is operating according to its specifications; thus

$H_0 : \mu = 355$

(where μ is the mean volume delivered by the machine)

Sue is looking for evidence of *any* difference; thus, the alternate hypothesis is

$H_1 : \mu \neq 355$

Since the hypothesis concerns a single population mean, a z-test is indicated. The population follows a normal distribution, and the standard deviation is known, so a z-test is appropriate.

In order to calculate the test statistic (z), we must first find the sample mean from the data. Use a calculator or computer to find that $\bar{x} = 355.037$.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{355.037 - 355}{0.05/\sqrt{10}} = 2.34$$

The calculation of the p-value will be a little different. If we only find the area under the normal curve *above* z = 2.34, then we have found the probability of obtaining a sample mean of 355.037 or *higher*—what about the probability of obtaining a low value?

In the case that the alternate hypothesis uses $\neq$, the p-value is found by *doubling the tail area*—in this case, we double the area above z = 2.34.

The area above z = 2.34 is 0.0096; thus, the p-value for this test is 0.0192.

If the machine is delivering 355 mL, then the probability of obtaining a sample mean this far (0.037 mL) or farther from 355 mL is 0.0096, or 0.96%. This is pretty rare; I'll reject the null hypothesis.

It appears that the machine is not working correctly.

N.B.: since the alternate hypothesis is $\neq$, we cannot conclude that the machine is delivering *more* than 355 mL—we can only say that the amount is *different* from 355 mL.

34 z Test for Two Means

34.1 Indications

The Null Hypothesis should be an assumption about the difference in the population means for two populations (note that the same quantitative variable must have been measured in each population). The data should consist of two samples of quantitative data (one from each population). The samples must be obtained independently from each other.

34.2 Requirements

The samples must be drawn from populations which have known Standard Deviations (or Variances). Also, the measured variable in each population (generically denoted x_1 and x_2) should have a Normal Distribution.

Note that if the distributions of the variables in the populations are non-normal (or unknown), the two-sample z-test can still be used for approximate results, provided the combined sample size (sum of sample sizes) is sufficiently large. Historically, a combined sample size of at least 30 has been considered sufficiently large; reality is (of course) much more complicated, but this rule of thumb is still in use in many textbooks.

34.3 Procedure

- The Null Hypothesis:

$$H_0 : \mu_1 - \mu_2 = \delta$$

in which δ is the supposed difference in the expected values under the null hypothesis.

- The Alternate Hypothesis:

$$H_0 : \mu_1 - \mu_2 < \delta$$

$$H_0 : \mu_1 - \mu_2 > \delta$$

$$H_0 : \mu_1 - \mu_2 \neq \delta$$

For more information about the Null and Alternate Hypotheses, see the page on the z test for a single mean.

- The Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Usually, the null hypothesis is that the population means are equal; in this case, the formula reduces to

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In the past, the calculations were simpler if the Variances (and thus the Standard Deviations) of the two populations could be assumed equal. This process is called Pooling, and many textbooks still use it, though it is falling out of practice (since computers and calculators have all but removed any computational problems).

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The Significance (p-value)

Calculate the probability of observing a value of z (from a Standard Normal Distribution) using the Alternate Hypothesis to indicate the direction in which the area under the Probability Density Function is to be calculated. This is the Attained Significance, or p-value.

Note that some (older) methods first chose a Level Of Significance, which was then translated into a value of z. This made more sense (and was easier!) in the days before computers and graphics calculators.

- Decision

The Attained Significance represents the probability of obtaining a test statistic as extreme, or more extreme, than ours—if the null hypothesis is true.

If the Attained Significance (p-value) is sufficiently low, then this indicates that our test statistic is unusual (rare)—we usually take this as evidence that the null hypothesis is in error. In this case, we *reject the null hypothesis.*

If the p-value is large, then this indicates that the test statistic is usual (common)—we take this as a lack of evidence against the null hypothesis. In this case, we *fail to reject the null hypothesis.*

It is common to use 5% as the dividing line between the common and the unusual; again, reality is more complicated.

34.4 Worked Examples

34.4.1 Do Professors Make More Money at Larger Universities?

Universities and colleges in the United States of America are categorized by the highest degree offered. Type IIA institutions offer a Master's Degree, and type IIB institutions offer a Baccalaureate degree. A professor, looking for a new position, wonders if the salary difference between type IIA and IIB institutions is really significant.

He finds that a random sample of 200 IIA institutions has a mean salary (for full professors) of $54,218.00, with standard deviation $8,450. A random sample of 200 IIB institutions has a mean salary (for full professors) of $46,550.00, with standard deviation $9,500 (assume that the sample standard deviations are in fact the population standard deviations).

Do these data indicate a significantly higher salary at IIA institutions?

The null hypothesis is that there is no difference; thus

$$\mathrm{H}_0 : \mu_\mathrm{A} = \mu_\mathrm{B}$$

(where μ_A is the true mean full professor salary at IIA institutions, and μ_B is the mean at IIB institutions)

He is looking for evidence that IIA institutions have a higher mean salary; thus the alternate hypothesis is

$$\mathrm{H}_1 : \mu_\mathrm{A} > \mu_\mathrm{B}$$

Since the hypotheses concern means from independent samples (we'll assume that these are independent samples), a two sample test is indicated. The samples are large, and the standard deviations are known (assumed?), so a two sample z-test is appropriate.

$$z = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{54218 - 46550}{\sqrt{\frac{8450^2}{200} + \frac{9500^2}{200}}} = 8.5292$$

Now we find the area to the right of z = 8.5292 in the Standard Normal Distribution. This can be done with a table of values or software—I get 0.

If the null hypothesis is true, and there is no difference in the salaries between the two types of institutions, then the probability of obtaining samples where the mean for IIA institutions is at least $7,668 higher than the mean for IIB institutions is essentially zero.

This occurs far too rarely to attribute to chance variation; it seems quite unusual. I reject the null hypothesis (at any reasonable level of significance!).

It appears that IIA schools have a significantly higher salary than IIB schools.

34.4.2 Example 2

35 t Test for a single mean

The t- test is the most powerful parametric test for calculating the significance of a small sample mean.

A one sample t-test has the following null hypothesis:

$H_0: \quad \mu = c$

where the Greek letter μ (mu) represents the population mean and c represents its assumed (hypothesized) value. In statistics it is usual to employ Greek letters for population parameters and Roman letters for sample statistics. The t-test is the small sample analog of the z test which is suitable for large samples. A small sample is generally regarded as one of size n<30.

A t-test is necessary for small samples because their distributions are not normal. If the sample is large (n>=30) then statistical theory says that the sample mean is normally distributed and a z test for a single mean can be used. This is a result of a famous statistical theorem, the Central limit theorem.

A t-test, however, can still be applied to larger samples and as the sample size n grows larger and larger, the results of a t-test and z-test become closer and closer. In the limit, with infinite degrees of freedom, the results of t and z tests become identical.

In order to perform a t-test, one first has to calculate the "degrees of freedom." This quantity takes into account the sample size and the number of parameters that are being estimated. Here, the population parameter, mu is being estimated by the sample statistic x-bar, the mean of the sample data. For a t-test the degrees of freedom of the single mean is n-1. This is because only one population parameter (the population mean)is being estimated by a sample statistic (the sample mean).

```
degrees of freedom (df)=n-1
```

For example, for a sample size n=15, the df=14.

35.0.3 Example

A college professor wants to compare her students' scores with the national average. She chooses an SRS of 20 students, who score an average of 50.2 on a standardized test. Their scores have a standard deviation of 2.5. The national average on the test is a 60. She wants to know if her students scored 'significantly**lower than the national average.**

Significance tests follow a procedure in several steps.

Step 1

First, state the problem in terms of a distribution and identify the parameters of interest. Mention the sample. We will assume that the scores (X) of the students in the professor's class are approximately normally distributed with unknown parameters μ and σ

Step 2

State the hypotheses in symbols and words.

$H_O: \quad \mu = 60$

The null hypothesis is that her students scored on par with the national average.

$H_A: \quad \mu < 60$

The alternative hypothesis is that her students scored lower than the national average.

Step 3

Secondly, identify the test to be used. Since we have an SRS of small size and do not know the standard deviation of the population, we will use a one-sample t-test.

The formula for the t-statistic T for a one-sample test is as follows:

$$T = \frac{\overline{X} - 60}{S/\sqrt{20}}$$

where $\overline{X}$ is the sample mean and S is the sample standard deviation.

A quite common mistake is to say that the formula for the t-test statistic is:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

This is not a statistic, because μ is unknown, which is the crucial point in such a problem. Most people even don't notice it. Another problem with this formula is the use of x and s. They are to be considered the sample statistics and not their values.

The right general formula is:

$$T = \frac{\overline{X} - c}{S/\sqrt{n}}$$

in which c is the hypothetical value for μ specified by the null hypothesis.

(The standard deviation of the sample divided by the square root of the sample size is known as the "standard error" of the sample.)

Step 4

State the distribution of the test statistic under the null hypothesis. Under H_0 the statistic T will follow a Student's distribution with 19 degrees of freedom: $T \sim \tau \cdot (20-1)$.

Step 5

Compute the observed value t of the test statistic T, by entering the values, as follows:

$$t = \frac{\overline{x}-60}{s/\sqrt{20}} = \frac{50.2-60.0}{2.5/\sqrt{20}} = \frac{-9.8}{2.5/4.47} = \frac{-9.8}{0.559} = -17.5$$

Step 6

Determine the so-called p-value of the value t of the test statistic T. We will reject the null hypothesis for too small values of T, so we compute the left p-value:

p-value

$$= P(T \leq t; H_0) = P(T(19) \leq -17.5) \approx 0$$

The Student's distribution gives $T(19) = 1.729$ at probabilities 0.95 and degrees of freedom 19. The p-value is approximated at 1.777e-13.

Step 7

Lastly, interpret the results in the context of the problem. The p-value indicates that the results almost certainly did not happen by chance and we have sufficient evidence to **reject the null hypothesis.** *The professor's students did score significantly lower than the national average.*

35.0.4 See also

- w:Errors and residuals in statistics[1]

[1] http://en.wikipedia.org/wiki/Errors%20and%20residuals%20in%20statistics

36 t Test for Two Means

In both the one- and two-tailed versions of the small two-sample t-test, we assume that the means of the two populations are equal. To use a t-test for small (independent) samples, the following conditions must be met:

1. The samples must be selected randomly.
2. The samples must be independent.
3. Each population must have a normal distribution.

A small two sample t-test is used to test the difference between two population means m1 and m2 when the sample size for at least one population is less than 30.The standardized test statistic is:

37 One-Way ANOVA F Test

The one-way ANOVA F-test is used to identify if there are differences between subject effects. For instance, to investigate the effect of a certain new drug on the number of white blood cells, in an experiment the drug is given to three different groups, one of healthy people, one with people with a light form of the considered disease and one with a severe form of the disease. Generally the **an**alysis **o**f **va**riance identifies whether there is a significant difference in effect of the drug on the number of white blood cells between the groups. Significant refers to the fact that there will always be difference between the groups and also within the groups, but the purpose is to investigate whether the difference between the groups are large compared to the differences within the groups. To set up such an experiment three assumptions must be validated before calculating an F statistic: independent samples, homogeneity of variance, and normality. The first assumption suggests that there is no relation between the measurements for different subjects. Homogeneity of variance refers to equal variances among the different groups in the experiment (e.g., drug vs. placebo). Furthermore, the assumption of normality suggests that the distribution of each of these groups should be approximately normally distributed.

37.1 Model

The situation is modelled in the following way. The measurement of the j-th test person in group i is indicated by:

$$X_{ij} = \mu + \alpha_i + U_{ij}$$

.

This reads: the outcome of the measurement for j in group i is due to a general effect indicated by μ, an effect due to the group, α_i and an individual contribution U_{ij}.

The individual, or random, contributions U_{ij}, often referred to as *disturbances*, are considered to be independently, normally distributed, all with expected value 0 and standard deviation σ.

To make the model unambiguous the group effects are restrained by the condition:

$$\sum_i \alpha_i = 0$$

.

Now. a notational note: it is common practice to indicate averages over one or more indices by writing a dot in the place of the index or indices. So for instance

$$X_{i.} = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$$

The analysis of variance now divides the total "variance" in the form of the total "sum of squares" in two parts, one due to the variation within the groups and one due to the variation between the groups:

$$SST = \sum_{ij}(X_{ij} - X..)^2 = \sum_{ij}(X_{ij} - X_{i.} + X_{i.} - X..)^2 = \sum_{ij}(X_{ij} - X_{i.})^2 + \sum_{ij}(X_{i.} - X..)^2$$

.

We see the term sum of squares of error:

$$SSE = \sum_{ij}(X_{ij} - X_{i.})^2$$

of the total squared differences of the individual measurements from their group averages, as an indication of the variation within the groups, and the term sum of square of the factor

$$SSA = \sum_{ij}(X_{i.} - X..)^2$$

of the total squared differences of the group means from the overall mean, as an indication of the variation between the groups.

Under the null hypothesis of no effect:

$$H_0 : \forall_i \ \alpha_i = 0$$

we find:

$$SSE/\sigma^2$$

is chi-square distributed with a(m-1) degrees of freedom, and

$$SSA/\sigma^2$$

is chi-square distributed with a-1 degrees of freedom,

where a is the number of groups and m is the number of persons in each group.

Hence the quotient of the so-called mean sum of squares:

$$MSA = \frac{SSA}{a-1}$$

and

$$MSE = \frac{SSE}{a(m-1)}$$

may be used as a test statistic

$$F = \frac{MSA}{MSE}$$

which under the null hypothesis is F-distributed with $a-1$ degrees of freedom in the nominator and $a(m-1)$ in the denominator, because the unknown parameter σ does not play a role since it is cancelled out in the quotient.

38 Testing whether Proportion A Is Greater than Proportion B in Microsoft Excel

A running example from the 2004 American Presidential Race follows. It should be clear that the choice of poll and who is leading is irrelevant to the presentation of the concepts. According to an October 2nd Poll by NEWSWEEK[1] (LINK[2]), 47% of 1,013 registered VOTERS[3] would vote for JOHN KERRY[4]/JOHN EDWARDS[5] if the election were held today. 45% would vote for GEORGE BUSH[6]/DICK CHENEY[7], and 2% would vote for RALPH NADER[8]/PETER CAMEJO[9].

- Open a new Blank Workbook in the program MICROSOFT EXCEL[10].
- Enter Kerry's reported percentage *p* in cell A1 (0.47).
- Enter Bush's reported percentage *q* in cell B1 (0.45).
- Enter the number of respondents *N* in cell C1 (1013). This can be found in most responsible reports on polls.
- In cell A2, copy and paste the next line of text in its entirety and press Enter. This is the Microsoft Excel expression of the standard error of the difference as shown ABOVE[11].

```
=sqrt(A1*(1-A1)/C1+B1*(1-B1)/C1+2*A1*B1/C1)
```

- In cell A3, copy and paste the next line of text in its entirety and press Enter. This is the Microsoft Excel expression of the probability that Kerry is leading based on the NORMAL DISTRIBUTION[12] given the logic HERE[13].

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/NEWSWEEK

[2] HTTP://WWW.MSNBC.MSN.COM/ID/6159637/SITE/NEWSWEEK/

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/VOTERS

[4] HTTP://EN.WIKIPEDIA.ORG/WIKI/JOHN%20KERRY

[5] HTTP://EN.WIKIPEDIA.ORG/WIKI/JOHN%20EDWARDS

[6] HTTP://EN.WIKIPEDIA.ORG/WIKI/GEORGE%20BUSH

[7] HTTP://EN.WIKIPEDIA.ORG/WIKI/DICK%20CHENEY

[8] HTTP://EN.WIKIPEDIA.ORG/WIKI/RALPH%20NADER

[9] HTTP://EN.WIKIPEDIA.ORG/WIKI/PETER%20CAMEJO

[10] HTTP://EN.WIKIPEDIA.ORG/WIKI/MICROSOFT%20EXCEL

[11] HTTP://EN.WIKIPEDIA.ORG/WIKI/MARGIN%20OF%20ERROR%23COMPARING%20PERCENTAGES%3A%20THE%20PROBABILITY%20OF%20LEADING

[12] HTTP://EN.WIKIPEDIA.ORG/WIKI/NORMAL%20DISTRIBUTION

[13] HTTP://EN.WIKIPEDIA.ORG/WIKI/MARGIN%20OF%20ERROR%23COMPARING%20PERCENTAGES%3A%20THE%20PROBABILITY%20OF%20LEADING

=normdist((A1-B1),0,A2,1)

- Don't forget that the percentages will be in decimal form. The percentage will be 0.5, or 50% if A1 and B1 are the same, of course.

The above text might be enough to do the necessary calculation, it doesn't contribute to the understanding of the statistical test involved. Much too often people think statistics is a matter of calculation with complex formulas.

So here is the problem: Let p be the population fraction of the registered voters who vote for Kerry and q likewise for Bush. In a poll n = 1013 respondents are asked to state their choice. A number of K respondents says to choose Kerry, a number B says to vote for Bush. K and B are random variables. The observed values for K and B are resp. k and b (numbers). So k/n is an estimate of p and b/n an estimate of q. The random variables K and B follow a trinomial distribution with parameters n, p, q and 1-p-q. Will Kerry be ahead of Bush? That is to say: wiil p > q? To investigate this we perform a statistical test, with null hypothesis:

$$H_0 : p = q$$

against the alternative

$$H_1 : p > q$$

.

What is an appropriate test statistic T? We take:

$$T = K - B$$

.

(In the above calculation $T = \frac{K}{n} - \frac{B}{n} = \frac{K-B}{n}$ is taken, which leads to the same calculation.)

We have to state the distribution of T under the null hypothesis. We may assume T is approximately normally distributed.

It is quite obvious that its expectation under H$_0$ is:

$$E_0 T = 0$$

.

Its variance under H$_0$ is not as obvious.

$$var_0(T) = var(K-B) = var(K) + var(B) - 2cov(K,B) = np(1-p) + nq(1-q) + 2npq$$

.

We approximate the variance by using the sample fractions instead of the population fractions:

$$var_0(T) \approx 1013 \times 0.47(1-0,46) + 1013 \times 0.45(1-0.45) + 2 \times 1013 \times 0,47 \times 0.45 \approx 931$$

.

The standard deviation s will approximately be:

$$s = \sqrt{var_0(T)} \approx \sqrt{931} = 30.5$$

.

In the sample we have found a value t = k - b = (0.47-0.45)1013 = 20.26 for T. We will reject the null hypothesis in favour of the alternative for large values of T. So the question is: is 20.26 to be considered a large value for T? The criterion will be the so called p-value of this outcome:

$$p - value = P(T \geq t; H_0) = P(T \geq 20.26; H_0) = P(Z \geq \frac{20.26}{30.5}) = 1 - \Phi(0.67) = 0.25$$

.

This is a very large p-value, so there is no reason whatsoever to reject the null hypothesis.

39 Chi-Squared Tests

39.1 General idea

Assume you have observed absolute frequencies o_i and expected absolute frequencies e_i under the Null hypothesis of your test then it holds

$V = \sum_i \frac{(o_i - e_i)^2}{e_i} \approx \chi^2_f.$

i might denote a simple index running from $1, ..., I$ or even a multiindex $(i_1, ..., i_p)$ running from $(1, ..., 1)$ to $(I_1, ..., I_p)$.

The test statistics V is approximately χ^2 distributed, if

1. for all absolute expected frequencies e_i holds $e_i \geq 1$ and
2. for at least 80% of the absolute expected frequencies e_i holds $e_i \geq 5$.

Note: In different books you might find different approximation conditions, please feel free to add further ones.

The degrees of freedom can be computed by the numbers of absolute observed frequencies which can be chosen freely. We know that the sum of absolute expected frequencies is

$\sum_i o_i = n$

which means that the maximum number of degrees of freedom is $I - 1$. We might have to subtract from the number of degrees of freedom the number of parameters we need to estimate from the sample, since this implies further relationships between the observed frequencies.

39.2 Derivation of the distribution of the test statistic

Following Boero, Smith and Wallis (2002) we need knowledge about multivariate statistics to understand the derivation.

The random variable O describing the absolute observed frequencies $(o_1, ..., o_k)$ in a sample has a multinomial distribution $O \sim M(n; p_1, ..., p_k)$ with n the number of observations in the sample, p_i the unknown true probabilities. With certain approximation conditions (central limit theorem) it holds that

$O \sim M(n; p_1, ..., p_k) \approx N_k(\mu; \Sigma)$

with N_k the multivariate k dimensional normal distribution, $\mu = (np_1, ..., np_k)$ and

$$\Sigma = (\sigma_{ij})_{i,j=1,\dots,k} = \begin{cases} -np_ip_j, & \text{if } i \neq j \\ np_i(1-p_i) & \text{otherwise} \end{cases}.$$

The covariance matrix Σ has only rank $k-1$, since $p_1 + \dots + p_k = 1$.

If we considered the generalized inverse Σ^- then it holds that

$(O-\mu)^T\Sigma^-(O-\mu) = \sum_i \frac{(o_i-e_i)^2}{e_i} \sim \chi^2_{k-1}$

distributed (for a proof see Pringle and Rayner, 1971).

Since the multinomial distribution is approximately multivariate normal distributed, the term is

$\sum_i \frac{(o_i-e_i)^2}{e_i} \approx \chi^2_{k-1}$

distributed. If further relations between the observed probabilities are there then the rank of Σ will decrease further.

A common situation is that parameters on which the expected probabilities depend needs to be estimated from the observed data. As said above, usually is stated that the degrees of freedom for the chi square distribution is $k-1-r$ with r the number of estimated parameters. In case of parameter estimation with the maximum-likelihood method this is only true if the estimator is efficient (Chernoff and Lehmann, 1954). In general it holds that degrees of freedom are somewhere between $k-1-r$ and $k-1$.

39.3 Examples

The most famous examples will be handled in detail at further sections: χ^2 test for independence, χ^2 test for homogeneity and χ^2 test for distributions.

The χ^2 test can be used to generate "quick and dirty" test, e.g.

H_0 : The random variable X is symmetrically distributed versus

H_1 : the random variable X is not symmetrically distributed.

We know that in case of a symmetrical distribution the arithmetic mean $\bar{x}$ and median should be nearly the same. So a simple way to test this hypothesis would be to count how many observations are less than the mean (n_-)and how many observations are larger than the arithmetic mean (n_+). If mean and median are the same than 50% of the observation should smaller than the mean and 50% should be larger than the mean. It holds

$V = \frac{(n_- - n/2)^2}{n/2} + \frac{(n_+ - n/2)^2}{n/2} \approx \chi^2_1.$

39.4 References

- Boero, G., Smith, J., Wallis, K.F. (2002). *The properties of some goodness-of-fit test*, University of Warwick, Department of Economics, The Warwick Economics Research Paper Series 653, http://www2.warwick.ac.uk/fac/soc/economics/research/papers/twerp653.pdf

- Chernoff H, Lehmann E.L. (1952). *The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit.* The Annals of Mathematical Statistics; 25:576-586.
- Pringle, R.M., Rayner, A.A. (1971). Generalized Inverse Matrices with Applications to Statistics. London: Charles Griffin.
- Wikipedia, Pearson's chi-square test: http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test

40 Distributions Problems

A normal distribution has $\mu = 100$ and $\sigma = 15$. What percent of the distribution is greater than 120?

41 Numerical Methods

Often the solution of statistical problems and/or methods involve the use of tools from numerical mathematics. An example might be MAXIMUM-LIKELIHOOD ESTIMATION[1] of $\widehat{\Theta}$which involves the maximization of the LIKELIHOOD FUNCTION[2] L:

$\widehat{\Theta} = \max_\theta \; L(\theta|x_1,...,x_n)$.

The maximization here requires the use of optimization routines. Other numerical methods and their application in statistics are described in this section.

Contents of this section:

- BASIC LINEAR ALGEBRA AND GRAM-SCHMIDT ORTHOGONALIZATION[3]

This section is dedicated to the *Gram-Schmidt Orthogonalization* which occurs frequently in the solution of statistical problems. Additionally some results of algebra theory which are necessary to understand the *Gram-Schmidt Orthogonalization* are provided. The *Gram-Schmidt Orthogonalization* is an algorithm which generates from a set of linear dependent vectors a new set of linear independent vectors which span the same space. Computation based on linear independent vectors is simpler than computation based on linear dependent vectors.

- UNCONSTRAINED OPTIMIZATION[4]

Numerical Optimization occurs in all kind of problem - a prominent example being the Maximum-Likelihood estimation as described above. Hence this section describes one important class of optimization algorithms, namely the so-called *Gradient Methods.* After describing the theory and developing an intuition about the general procedure, three specific algorithms (the *Method of Steepest Descent*, the *Newtonian Method*, the class of *Variable Metric Methods*) are described in more detail. Especially we provide an (graphical) evaluation of the performance of these three algorithms for specific criterion functions (the Himmelblau function and the Rosenbrock function). Furthermore we come back to Maximum-Likelihood estimation and give a concrete example how to tackle this problem with the methods developed in this section.

- QUANTILE REGRESSION[5]

In *OLS*, one has the primary goal of determining the conditional mean of random variable Y, given some explanatory variable x_i , $E[Y|x_i]$. *Quantile Regression* goes beyond this and

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/MAXIMUM_LIKELIHOOD

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/LIKELIHOOD

[3] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ANUMERICAL%20METHODS%2FBASIC%20LINEAR%20ALGEBRA%20AND%20GRAM-SCHMIDT%20ORTHOGONALIZATION

[4] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ANUMERICAL%20METHODS%2FOPTIMIZATION

[5] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS%3ANUMERICAL%20METHODS%2FQUANTILE%20REGRESSION

enables us to pose such a question at any quantile of the conditional distribution function. It thereby focuses on the interrelationship between a dependent variable and its explanatory variables for a given quantile.

- Numerical Comparison of Statistical Software[6]

Statistical calculations require an extra accuracy and are open to some errors such as truncation or cancellation error etc. These errors occur due to binary representation and finite precision and may cause inaccurate results. In this work we are going to discuss the accuracy of the statistical software, different tests and methods available for measuring the accuracy and the comparison of different packages.

- Numerics in Excel[7]

The purpose of this paper is to evaluate the accuracy of MS Excel in terms of statistical procedures and to conclude whether the MS Excel should be used for (statistical) scientific purposes or not. The evaluation is made for MS Excel versions 97, 2000, XP and 2003.

- Random Number Generation[8]

[6] http://en.wikibooks.org/wiki/Statistics%3ANumerical%20Methods%2FNumerical%20Comparison%20of%20Statistical%20Software

[7] http://en.wikibooks.org/wiki/Statistics%3ANumerical%20Methods%2FNumerics%20in%20Excel

[8] http://en.wikibooks.org/wiki/Statistics%3ANumerical%20Methods%2FRandom%20Number%20Generation

42 Basic Linear Algebra and Gram-Schmidt Orthogonalization

42.1 Introduction

Basically, all the sections found here can be also found in a linear algebra book. However, the Gram-Schmidt Orthogonalization is used in statistical algorithm and in the solution of statistical problems. Therefore, we briefly jump into the linear algebra theory which is necessary to understand Gram-Schmidt Orthogonalization.

The following subsections also contain examples. It is very important for further understanding that the concepts presented here are not only valid for typical vectors as tuple of real numbers, but also functions that can be considered vectors.

42.2 Fields

42.2.1 Definition

A set R with two operations $+$ and $*$ on its elements is called a *field* (or short $(R,+,*)$), if the following conditions hold:

1. For all $\alpha,\beta \in R$ holds $\alpha+\beta \in R$
2. For all $\alpha,\beta \in R$ holds $\alpha+\beta = \beta+\alpha$ (commutativity)
3. For all $\alpha,\beta,\gamma \in R$ holds $\alpha+(\beta+\gamma) = (\alpha+\beta)+\gamma$ (associativity)
4. It exist a unique element 0, called *zero*, such that for all $\alpha \in R$ holds $\alpha+0 = \alpha$
5. For all $\alpha \in R$ a unique element $-\alpha$, such that holds $\alpha+(-\alpha) = 0$
6. For all $\alpha,\beta \in R$ holds $\alpha*\beta \in R$
7. For all $\alpha,\beta \in R$ holds $\alpha*\beta = \beta*\alpha$ (commutativity)
8. For all $\alpha,\beta,\gamma \in R$ holds $\alpha*(\beta*\gamma) = (\alpha*\beta)*\gamma$ (associativity)
9. It exist a unique element 1, called *one*, such that for all $\alpha \in R$ holds $\alpha*1 = \alpha$
10. For all non-zero $\alpha \in R$ a unique element α^{-1}, such that holds $\alpha*\alpha^{-1} = 1$
11. For all $\alpha,\beta,\gamma \in R$ holds $\alpha*(\beta+\gamma) = \alpha*\beta+\alpha*\gamma$ (distributivity)

The elements of R are also called *scalars*.

42.2.2 Examples

It can easily be proven that real numbers with the well known addition and multiplication $(IR,+,*)$ are a field. The same holds for complex numbers with the addition and multiplication. Actually, there are not many more sets with two operations which fulfill all of these conditions.

For statistics, only the real and complex numbers with the addition and multiplication are important.

42.3 Vector spaces

42.3.1 Definition

A set V with two operations $+$ and $*$ on its elements is called a *vector space over R*, if the following conditions hold:

1. For all $x,y \in V$ holds $x+y \in V$
2. For all $x,y \in V$ holds $x+y = y+x$ (commutativity)
3. For all $x,y,z \in V$ holds $x+(y+z) = (x+y)+z$ (associativity)
4. It exist a unique element $\mathbb{O}$, called *origin*, such that for all $x \in V$ holds $x+\mathbb{O} = x$
5. For all $x \in V$ exists a unique element $-v$, such that holds $x+(-x) = \mathbb{O}$
6. For all $\alpha \in R$ and $x \in V$ holds $\alpha * x \in V$
7. For all $\alpha,\beta \in R$ and $x \in V$ holds $\alpha*(\beta*x) = (\alpha*\beta)*x$ (associativity)
8. For all $x \in V$ and $1 \in R$ holds $1*x = x$
9. For all $\alpha \in R$ and for all $x,y \in V$holds $\alpha*(x+y) = \alpha*x+\alpha*y$ (distributivity wrt. vector addition)
10. For all $\alpha,\beta \in R$ and for all $x \in V$holds $(\alpha+\beta)*x = \alpha*x+\beta*x$ (distributivity wrt. scalar addition)

Note that we used the same symbols $+$ and $*$ for different operations in R and V. The elements of V are also called *vectors*.

Examples:

1. The set IR^p with the real-valued vectors $(x_1,...,x_p)$ with elementwise addition $x+y = (x_1+y_1,...,x_p+y_p)$ and the elementwise multiplication $\alpha*x = (\alpha x_1,...,\alpha x_p)$ is a vector space over IR.
2. The set of polynomials of degree p, $P(x) = b_0+b_1x+b_2x^2+...+b_px^p$, with usual addition and multiplication is a vector space over IR.

42.3.2 Linear combinations

A vector x can be written as a linear combination of vectors $x_1,...x_n$, if

$x = \sum_{i=1}^n \alpha_i x_i$

with $\alpha_i \in R$.

Examples:

- $(1,2,3)$ is a linear combination of $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ since $(1,2,3) = 1*(1,0,0)+2*(0,1,0)+3*(0,0,1)$

- $1+2*x+3*x^2$ is a linear combination of $1+x+x^2$, $x+x^2$, x^2 since $1+2*x+3*x^2 = 1*(1+x+x^2)+1*(x+x^2)+1*(x^2)$

42.3.3 Basis of a vector space

A set of vectors $x_1,...,x_n$ is called a *basis* of the vector space V, if

1. for each vector $xinV$ exist scalars $\alpha_1,...,\alpha_n \in R$ such that $x = \sum_i \alpha_i x_i$ 2. there is no subset of $\{x_1,...,x_n\}$ such that 1. is fulfilled.

Note, that a vector space can have several bases.

Examples:

- Each vector $(\alpha_1,\alpha_2,\alpha_3) \in IR^3$ can be written as $\alpha_1*(1,0,0)+\alpha_2*(0,1,0)+\alpha_3*(0,0,1)$. Therefore is $\{(1,0,0),(0,1,0),(0,0,1)\}$ a basis of IR^3.
- Each polynomial of degree p can be written as linear combination of $\{1,x,x^2,...,x^p\}$ and therefore forms a basis for this vector space.

Actually, for both examples we would have to prove condition 2., but it is clear that it holds.

42.3.4 Dimension of a vector space

A *dimension* of a vector space is the number of vectors which are necessary for a basis. A vector space has infinitely many number of basis, but the dimension is uniquely determined. Note that the vector space may have a dimension of infinity, e.g. consider the space of continuous functions.

Examples:

- The dimension of IR^3 is three, the dimension of IR^p is p .
- The dimension of the polynomials of degree p is $p+1$.

42.3.5 Scalar products

A mapping $< .,. >: V \times V \to R$ is called a *scalar product* if the following holds for all $x,x_1,x_2,y,y_1,y_2 \in V$ and $\alpha_1,\alpha_2 inR$:

1. $< \alpha_1 x_1 + \alpha_2 x_2, y >= \alpha_1 < x_1, y > + \alpha_2 < x_2, y >$
2. $< x, \alpha_1 y_1 + \alpha_2 y_2 >= \alpha_1 < x, y_1 > + \alpha_2 < x, y_2 >$
3. $< x,y >= \overline{< y,x >}$ with $\overline{\alpha + \imath\beta} = \alpha - \imath\beta$
4. $< x,x > \geq 0$ with $< x,x >= 0 \Leftrightarrow x = \mathbb{O}$

Examples:

- The typical scalar product in IR^p is $< x,y >= \sum_i x_i y_i$.
- $< f,g >= \int_a^b f(x)*g(x)dx$ is a scalar product on the vector space of polynomials of degree p.

42.3.6 Norm

A *norm* of a vector is a mapping $\|.\| : V \to R$, if holds

1. $\|x\| \geq 0$ for all $x \in V$ and $\|x\| = 0 \Leftrightarrow x = \mathbb{O}$ (positive definiteness)
2. $\|\alpha v\| = \mid \alpha \mid \|x\|$ for all $x \in V$ and all $\alpha \in R$
3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$ (triangle inequality)

Examples:

- The L_q norm of a vector in IR^p is defined as $\|_q = \sqrt[q]{\sum_{i=1}^p x_i^q}$.
- Each scalar product generates a norm by $\| = \sqrt{< x, x >}$, therefore $\| = \sqrt{\int_a^b f^2(x)dx}$ is a norm for the polynomials of degree p.

42.3.7 Orthogonality

Two vectors x and y are *orthogonal* to each other if $< x, y >= 0$. In IR^p it holds that the cosine of the angle between two vectors can expressed as

$\cos(\angle(x,y)) = \frac{<x,y>}{\|\|}$.

If the angle between x and y is ninety degree (orthogonal) then the cosine is zero and it follows that $< x, y >= 0$.

A set of vectors $x_1, ..., x_p$ is called *orthonormal*, if

$$< x_i, x_j >= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}.$$

If we consider a basis $e_1, ..., e_p$ of a vector space then we would like to have a orthonormal basis. Why ?

Since we have a basis, each vector x and y can be expressed by $x = \alpha_1 e_1 + ... + \alpha_p e_p$ and $y = \beta_1 e_1 + ... + \beta_p e_p$. Therefore the scalar product of x and y reduces to

$$\begin{aligned} < x, y > &=< \alpha_1 e_1 + ... + \alpha_p e_p, \beta_1 e_1 + ... + \beta_p e_p > \\ &= \sum_{i=1}^p \sum_{j=1}^p \alpha_i \beta_j < e_i, e_j > \\ &= \sum_{i=1}^p \alpha_i \beta_i < e_i, e_i > \\ &= \alpha_1 \beta_1 + ... + \alpha_p \beta_p. \end{aligned}$$

Consequently, the computation of a scalar product is reduced to simple multiplication and addition if the coefficients are known. Remember that for our polynomials we would have to solve an integral!

42.4 Gram-Schmidt orthogonalization

42.4.1 Algorithm

The aim of the Gram-Schmidt orthogonalization is to find for a set of vectors $x_1, ..., x_p$ an equivalent set of *orthonormal* vectors $o_1, ..., o_p$ such that any vector which can be expressed as linear combination of $x_1, ..., x_p$ can also be expressed as linear combination of $o_1, ..., o_p$:

1. Set $b_1 = x_1$ and $o_1 = b_1/_1\|$

2. For each $i > 1$ set $b_i = x_i - \sum_{j=1}^{i-1} \frac{<x_i, b_j>}{<b_j, b_j>} b_j$ and $o_i = b_i/_i\|$, in each step the vector x_i is projected on b_j and the result is subtracted from x_i.

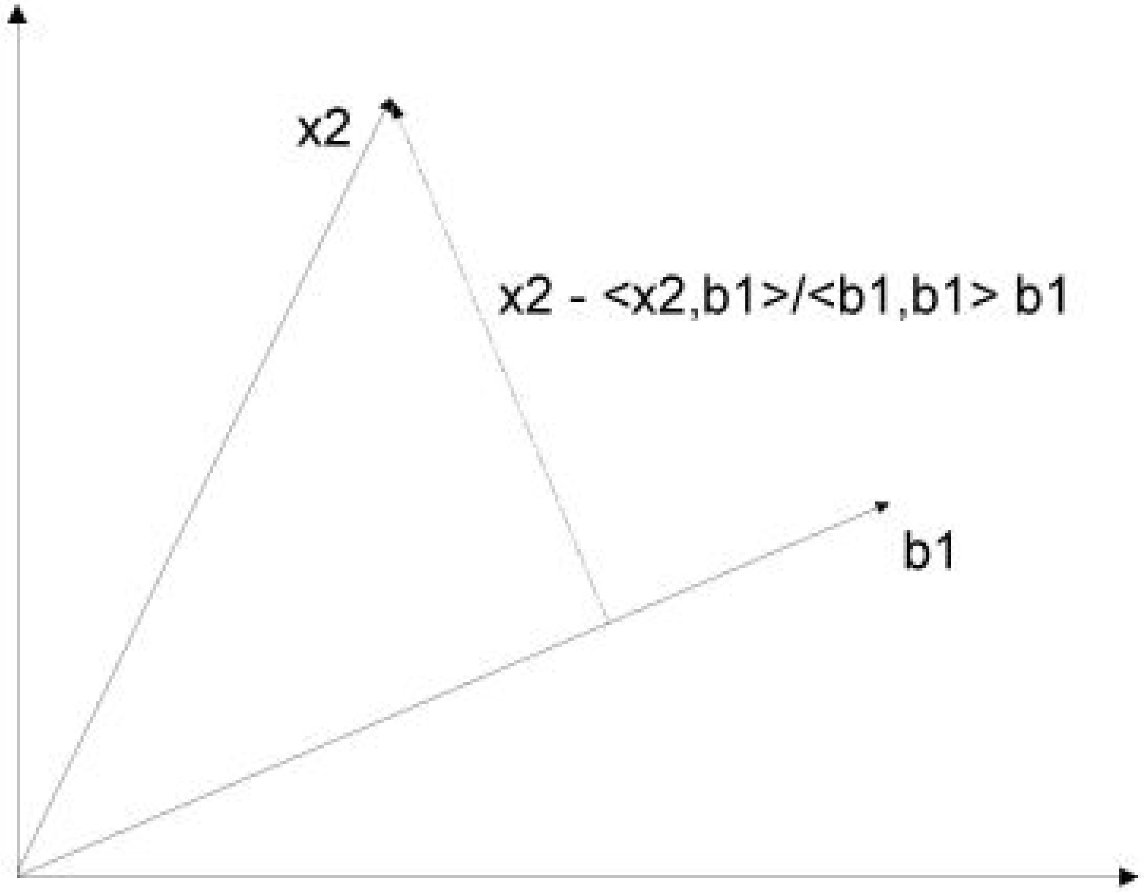


Figure 18

42.4.2 Example

Consider the polynomials of degree two in the interval$[-1, 1]$ with the scalar product $< f, g >= \int_{-1}^{1} f(x)g(x)dx$ and the norm $\| = \sqrt{< f, f >}$. We know that $f_1(x) = 1, f_2(x) = x$ and $f_3(x) = x^2$ are a basis for this vector space. Let us now construct an orthonormal basis:

Step 1a: $b_1(x) = f_1(x) = 1$

Step 1b: $o_1(x) = \frac{b_1(x)}{_1(x)\|} = \frac{1}{\sqrt{<b_1(x),b_1(x)>}} = \frac{1}{\sqrt{\int_{-1}^{1} 1dx}} = \frac{1}{\sqrt{2}}$

Step 2a: $b_2(x) = f_2(x) - \frac{<f_2(x),b_1(x)>}{<b_1(x),b_1(x)>}b_1(x) = x - \frac{\int_{-1}^{1} x\ 1dx}{2}1 = x - \frac{0}{2}1 = x$

Step 2b: $o_2(x) = \frac{b_2(x)}{_2(x)\|} = \frac{x}{\sqrt{<b_2(x),b_2(x)>}} = \frac{x}{\sqrt{\int_{-1}^{1} x^2dx}} = \frac{x}{\sqrt{2/3}} = x\sqrt{3/2}$

Step 3a: $b_3(x) = f_3(x) - \frac{<f_3(x),b_1(x)>}{<b_1(x),b_1(x)>}b_1(x) - \frac{<f_3(x),b_2(x)>}{<b_2(x),b_2(x)>}b_2(x) = x^2 - \frac{\int_{-1}^{1} x^2 1\ dx}{2}1 - \frac{\int_{-1}^{1} x^2 x\ dx}{2/3}x = x^2 - \frac{2/3}{2}1 - \frac{0}{2/3}x = x^2 - 1/3$

Step 3b: $o_3(x) = \frac{b_3(x)}{_3(x)\|} = \frac{x^2-1/3}{\sqrt{<b_3(x),b_3(x)>}} = \frac{x^2-1/3}{\sqrt{\int_{-1}^{1}(x^2-1/3)^2dx}} = \frac{x^2-1/3}{\sqrt{\int_{-1}^{1} x^4-2/3x^2+1/9\ dx}} = \frac{x^2-1/3}{\sqrt{8/45}} = \sqrt{\frac{5}{8}}(3x^2-1)$

It can be proven that $1/\sqrt{2}, x\sqrt{3/2}$ and $\sqrt{\frac{5}{8}}(3x^2-1)$ form a orthonormal basis with the above scalarproduct and norm.

42.4.3 Numerical instability

Consider the vectors $x_1 = (1,\epsilon,0,0), x_2 = (1,0,\epsilon,0)$ and $x_3 = (1,0,0,\epsilon)$. Assume that ϵ is so small that computing $1+\epsilon = 1$ holds on a computer (see HTTP://EN.WIKIPEDIA.ORG/WIKI/MACHINE_EPSILON).[1] Let compute a orthonormal basis for this vectors in IR^4 with the standard scalar product $< x,y >= x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4$ and the norm $\| = \sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2}$.

Step 1a. $b_1 = x_1 = (1,\epsilon,0,0)$

Step 1b. $o_1 = \frac{b_1}{_1\|} = \frac{b_1}{\sqrt{1+\epsilon^2}} = b_1$ with $1+\epsilon^2 = 1$

Step 2a. $b_2 = x_2 - \frac{<x_2,b_1>}{<b_1,b_1>}b_1 = (1,0,\epsilon,0) - \frac{1}{1+\epsilon^2}(1,\epsilon,0,0) = (0,-\epsilon,\epsilon,0)$

Step 2b. $o_2 = \frac{b_2}{_2\|} = \frac{b_2}{\sqrt{2\epsilon^2}} = (0,-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},0)$

Step 3a. $b_3 = x_3 - \frac{<x_3,b_1>}{<b_1,b_1>}b_1 - \frac{<x_3,b_2>}{<b_2,b_2>}b_2 = (1,0,0,\epsilon) - \frac{1}{1+\epsilon^2}(1,\epsilon,0,0) - \frac{0}{2\epsilon^2}(0,-\epsilon,\epsilon,0) = (0,-\epsilon,0,\epsilon)$

Step 3b. $o_3 = \frac{b_3}{_3\|} = \frac{b_3}{\sqrt{2\epsilon^2}} = (0,-\frac{1}{\sqrt{2}},0,\frac{1}{\sqrt{2}})$

It obvious that for the vectors

- $o_1 = (1,\epsilon,0,0)$

- $o_2 = (0,-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},0)$

- $o_3 = (0,-\frac{1}{\sqrt{2}},0,\frac{1}{\sqrt{2}})$

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/MACHINE_EPSILON).

the scalarproduct $< o_2, o_3 >= 1/2 \neq 0$. All other pairs are also not zero, but they are multiplied with ϵ such that we get a result near zero.

42.4.4 Modified Gram-Schmidt

To solve the problem a modified Gram-Schmidt algorithm is used:

1. Set $b_i = x_i$ for all i
2. for each i from 1 to n compute
   a) $o_i = \frac{b_i}{i\|}$
   b) for each j from $i+1$ to n compute $b_j = b_j - < b_j, o_i > o_i$

The difference is that we compute first our new b_i and subtract it from all other b_j. We apply the wrongly computed vector to all vectors instead of computing each b_i separately.

42.4.5 Example (recomputed)

Step 1. $b_1 = (1,\epsilon,0,0)$, $b_2 = (1,0,\epsilon,0)$, $b_3 = (1,0,0,\epsilon)$

Step 2a. $o_1 = \frac{b_1}{1\|} = \frac{b_1}{\sqrt{1+\epsilon^2}} = b_1 = (1,\epsilon,0,0)$ with $1+\epsilon^2 = 1$

Step 2b. $b_2 = b_2 - < b_2, o_1 > o_1 = (1,0,\epsilon,0) - (1,\epsilon,0,0) = (0,-\epsilon,\epsilon,0)$

Step 2c. $b_3 = b_3 - < b_3, o_1 > o_1 = (1,0,0,\epsilon) - (1,\epsilon,0,0) = (0,-\epsilon,0,\epsilon)$

Step 3a. $o_2 = \frac{b_2}{2\|} = \frac{b_2}{\sqrt{2\epsilon^2}} = (0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$

Step 3b. $b_3 = b_3 - < b_3, o_2 > o_2 = (0,-\epsilon,0,\epsilon) - \frac{\epsilon}{\sqrt{2}}(0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0) = (0, -\epsilon/2, -\epsilon/2, \epsilon)$

Step 4a. $o_3 = \frac{b_3}{3\|} = \frac{b_3}{\sqrt{3/2\epsilon^2}} = (0, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}})$

We can easily verify that $< o_2, o_3 >= 0$.

42.5 Application

42.5.1 Exploratory Project Pursuit

In the analysis of high-dimensional data we usually analyze projections of the data. The approach results from the Theorem of Cramer-Wold that states that the multidimensional distribution is fixed if we know all one-dimensional projections. Another theorem states that most (one-dimensional) projections of multivariate data are looking normal, even if the multivariate distribution of the data is highly non-normal.

Therefore in Exploratory Projection Pursuit we jugde the interestingness of a projection by comparison with a (standard) normal distribution. If we assume that the one-dimensional data x are standard normal distributed then after the transformation $z = 2\Phi^{-1}(x) - 1$ with $\Phi(x)$ the cumulative distribution function of the standard normal distribution then z is uniformly distributed in the interval $[-1;1]$.

Thus the interesting can measured by $\int_{-1}^{1}(f(z)-1/2)^2 dx$ with $f(z)$ a density estimated from the data. If the density $f(z)$ is equal to $1/2 < math > intheinterval < math > [-1;1]$ then the integral becomes zero and we have found that our projected data are normally distributed. An value larger than zero indicates a deviation from the normal distribution of the projected data and hopefully an interesting distribution.

42.5.2 Expansion with orthonormal polynomials

Let $L_i(z)$ a set of orthonormal polynomials with the scalar product $< f,g >= \int_{-1}^{1} f(z)g(z)dz$ and the norm $\| = \sqrt{< f,f >}$. What can we derive about a densities $f(z)$ in the interval $[-1;1]$?

If $f(z) = \sum_{i=0}^{I} a_i L_i(z)$ for some maximal degree I then it holds

$\int_{-1}^{1} f(z)L_j(z)dz = \int_{-1}^{1} \sum_{i=0}^{I} a_i L_i(z)L_j(z)dz = a_j \int_{-1}^{1} L_j(z)L_j(z)dz = a_j$

We can also write $\int_{-1}^{1} f(z)L_j(z)dz = E(L_j(z))$ or empirically we get an estimator $\hat{a}_j = \frac{1}{n}\sum_{k=1}^{n} L_j(z_k)$.

We describe the term $1/2 = \sum_{i=1}^{I} b_i L_i(z)$ and get for our integral

$\int_{-1}^{1}(f(z) - 1/2)^2 dz = \int_{-1}^{1}\left(\sum_{i=0}^{I}(a_i - b_i)L_i(z)\right)^2 dz = \sum_{i,j=0}^{I}\int_{-1}^{1}(a_i - b_i)(a_j - b_j)L_i(z)L_j(z)dz = \sum_{i=0}^{I}(a_i - b_i)^2$.

So using a orthonormal function set allows us to reduce the integral to a summation of coefficient which can be estimated from the data by plugging $\hat{a}_j$ in the formula above. The coefficients b_i can be precomputed in advance.

42.5.3 Normalized Legendre polynomials

The only problem left is to find the set of orthonormal polynomials $L_i(z)$ upto degree I. We know that $1, x, x^2, ..., x^I$ form a basis for this space. We have to apply the Gram-Schmidt orthogonalization to find the orthonormal polynomials. This has been started in the FIRST EXAMPLE[2].

The resulting polynomials are called normalized Legendre polynomials. Up to a sacling factor the normalized Legendre polynomials are identical to LEGENDRE POLYNOMIALS[3]. The Legendre polynomials have a recursive expression of the form

$L_i(z) = \frac{(2i-1)L_{i-1}(z)-(i-1)L_{i-2}(z)}{i}$

So computing our integral reduces to computing $L_0(z_k)$ and $L_1(z_k)$ and using the recursive relationship to compute the $\hat{a}_j$'s. Please note that the recursion can be numerically unstable!

[2] HTTP://EN.WIKIBOOKS.ORG/WIKI/STATISTICS:NUMERICAL_METHODS/BASIC_LINEAR_ALGEBRA_AND_GRAM-SCHMIDT_ORTHOGONALIZATION#EXAMPLE

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/LEGENDRE_POLYNOMIALS

42.6 References

- Halmos, P.R. (1974). Finite-Dimensional Vector Spaces, Springer: New York
- Persson, P.O. (2005). Introduction to Numerical Methods, Lecture 5 Gram-Schmidt[4]

[4] http://www-math.mit.edu/~{}persson/18.335/lec5handout6pp.pdf

43 Unconstrained Optimization

43.1 Introduction

In the following we will provide some notes on numerical optimization algorithms. As there are NUMEROUS METHODS[1] out there, we will restrict ourselves to the so-called *Gradient Methods*. There are basically two arguments why we consider this class as a natural starting point when thinking about numerical optimization algorithms. On the one hand, these methods are really workhorses in the field, so their frequent use in practice justifies their coverage here. On the other hand, this approach is highly intuitive in the sense that it somewhat follow naturally from the well-known PROPERTIES OF OPTIMA[2]. In particular we will concentrate on three examples of this class: the *Newtonian Method*, the *Method of Steepest Descent* and the class of *Variable Metric Methods*, nesting amongst others the *Quasi Newtonian Method*.

Before we start we will nevertheless stress that there does not seem to be a "one and only" algorithm but the performance of specific algorithms is always contingent on the specific problem to be solved. Therefore both experience and "trial-and-error" are very important in applied work. To clarify this point we will provide a couple of applications where the performance of different algorithms can be compared graphically. Furthermore a specific example on MAXIMUM LIKELIHOOD ESTIMATION[3] can be found at the end. Especially for statisticians and ECONOMETRICIANS[4] the Maximum Likelihood Estimator is probably the most important example of having to rely on numerical optimization algorithms in practice.

43.2 Theoretical Motivation

Any numerical optimization algorithm has solve the problem of finding "observable" properties of the function such that the computer program knows that a solution is reached. As we are dealing with problems of optimization two well-known results seem to be sensible starting points for such properties.

If f is differentiable and $x^\star$ is a (local) minimum, then

$(1a) \quad Df(x^\star) = 0$

i.e. the Jacobian $Df(x)$ is equal to zero

and

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/OPTIMIZATION_%28MATHEMATICS%29

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/STATIONARY_POINT

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/MAXIMUM_LIKELIHOOD

[4] HTTP://EN.WIKIPEDIA.ORG/WIKI/ECONOMETRICS

If f is twice differentiable and $x^\star$ is a (local) minimum, then

(1*b*) $x^T D^2 f(x^\star) x \geq 0$

i.e. the Hessian $D^2 f(x)$ is POS. SEMIDEFINITE[5].

In the following we will always denote the minimum by $x^\star$. Although these two conditions seem to represent statements that help in finding the optimum $x^\star$, there is the little catch that they give the implications of $x^\star$ being an optimum for the function f. But for our purposes we would need the opposite implication, i.e. finally we want to arrive at a statement of the form: "If some condition $g(f(x^\star))$ is true, then $x^\star$ is a minimum". But the two conditions above are clearly not sufficient in achieving this (consider for example the case of $f(x) = x^3$, with $Df(0) = D^2 f(0) = 0$ but $x^\star \neq 0$). Hence we have to look at an entire neighborhood of $x^\star$ as laid out in the following sufficient condition for detecting optima:

If $Df(x^\star) = 0$ and $x^T D^2 f(z) x \geq 0, \forall x \in \mathbb{R}^n$ and $z \in \mathcal{B}(x^\star, \delta)$, then: $x^\star$ is a local minimum.

Proof: For $x \in \mathcal{B}(x^\star, \delta)$ let $z = x^\star + t(x - x^\star) \in \mathcal{B}$. The TAYLOR APPROXIMATION[6] *yields: $f(x) - f(x^\star) = 0 + \frac{1}{2}(x - x^\star)^T D^2 f(z)(x - x^\star) \geq 0$, where $\mathcal{B}(x^\star, \delta)$ denotes an open ball around $x^\star$, i.e. $\mathcal{B}(x^\star, \delta) = \{x : ||x - x^\star|| \leq \delta\}$ for $\delta > 0$.*

In contrast to the two conditions above, this condition is sufficient for detecting optima - consider the two trivial examples

$f(x) = x^3$ with $Df(x^\star = 0) = 0$ but $x^T D^2 f(z) x = 6zx^2 \ngeq 0 \quad (e.g.\ z = -\frac{\delta}{2})$

and

$f(x) = x^4$ with $Df(x^\star = 0) = 0$ and $x^T D^2 f(z) x = 12z^2 x^2 \geq 0 \quad \forall z.$

Keeping this little caveat in mind we can now turn to the numerical optimization procedures.

43.3 Numerical Solutions

All the following algorithms will rely on the following assumption:

(A1) The set $N(f, f(x^{(0)}) = \{x \in \mathbb{R}^n | f(x) \leq f(x^{(0)})\}$ is COMPACT[7]

where $x^{(0)}$ is some given starting value for the algorithm. The significance of this assumption has to be seen in the *Weierstrass Theorem* which states that every compact set contains its SUPREMUM[8] and its INFIMUM[9]. So *(A1)* ensures that there is some solution in $N(f, f(x^{(0)})$. And at this global minimum $x^\star$ it of course holds true that $D(f(x^\star)) = 0$. So - keeping the discussion above in mind - the optimization problem basically boils down to the question of solving set of equations $D(f(x^\star)) = 0$.

[5] HTTP://EN.WIKIPEDIA.ORG/WIKI/POSITIVE-DEFINITE_MATRIX

[6] HTTP://EN.WIKIPEDIA.ORG/WIKI/TAYLOR%27S_THEOREM

[7] HTTP://EN.WIKIPEDIA.ORG/WIKI/COMPACT_SPACE

[8] HTTP://EN.WIKIPEDIA.ORG/WIKI/SUPREMUM

[9] HTTP://EN.WIKIPEDIA.ORG/WIKI/INFIMUM

43.3.1 The Direction of Descent

The problems with this approach are of course rather generically as $D(f(x^{\star})) = 0$ does hold true for MAXIMA AND SADDLE POINTS[10] as well. Hence, good algorithms should ensure that both maxima and saddle points are ruled out as potential solutions. Maxima can be ruled out very easily by requiring $f(x^{(k+1)}) < f(x^{(k)})$ i.e. we restrict ourselves to a SEQUENCE[11] $\{x^{(k)}\}_k$ such that the function value decreases in every step. The question is of course if this is always possible. Fortunately it is. The basic insight why this is the case is the following. When constructing the mapping $x^{(k+1)} = \varphi(x^{(k)})$ (i.e. the rule how we get from $x^{(k)}$ to $x^{(k+1)}$) we have two degrees of freedoms, namely

- the direction $d^{(k)}$ and
- the step length $\sigma^{(k)}$.

Hence we can choose in which direction we want to move to arrive at $x^{(k+1)}$ *and* how far this movement has to be. So if we choose $d^{(k)}$ and $\sigma^{(k)}$ in the "right way" we can effectively ensure that the function value decreases. The formal representation of this reasoning is provided in the following

Lemma: If $d^{(k)} \in \mathbb{R}^n$ *and* $Df(x)^T d^{(k)} < 0$ *then:* $\exists \bar{\sigma} > 0$ such that

$f(x + \sigma^{(k)} d^{(k)}) < f(x) \quad \forall \sigma \in (0, \bar{\sigma})$

Proof: As $Df(x)^T d^{(k)} < 0$ *and* $Df(x)^T d^{(k)} = \lim_{\sigma \to 0} \frac{f(x+\sigma^{(k)} d^{(k)}) - f(x)}{\sigma^{(k)}}$, *it follows that* $f(x + \sigma^{(k)} d^{(k)}) < f(x)$ *for* $\sigma^{(k)}$ *small enough.*

43.3.2 The General Procedure of Descending Methods

A direction vector $d^{(k)}$ that satisfies this condition is is called a *Direction of Descent*. In practice this Lemma allows us to use the following procedure to numerically solve optimization problems.

1. Define the SEQUENCE[12] $\{x^{(k)}\}_k$ recursively via $x^{(k+1)} = x^{(k)} + \sigma^{(k)} d^{(k)}$

2. Choose the direction $d^{(k)}$ from local information at the point $x^{(k)}$

3. Choose a step size $\sigma^{(k)}$ that ensures CONVERGENCE[13] of the algorithm.

4. Stop the iteration if $|f(x^{(k+1)}) - f(x^{(k)})| < \epsilon$ where $\epsilon > 0$ is some chosen tolerance value for the minimum

This procedure already hints that the choice of $d^{(k)}$ and $\sigma^{(k)}$ are not separable, but rather dependent. Especially note that even if the method is a descending method (i.e. both $d^{(k)}$ and $\sigma^{(k)}$ are chosen according to *Lemma 1*) the convergence to the minimum is not guaranteed. At a first glance this may seem a bit puzzling. If we found a sequence $\{x^{(k)}\}_k$ such that the function value decreases at every step, one might think that at some stage,

[10] HTTP://EN.WIKIPEDIA.ORG/WIKI/STATIONARY_POINT
[11] HTTP://EN.WIKIPEDIA.ORG/WIKI/SEQUENCE
[12] HTTP://EN.WIKIPEDIA.ORG/WIKI/SEQUENCE
[13] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONVERGENT_SERIES

i.e. in the limit of k tending to infinity we should reach the solution. Why this is not the case can be seen from the following example borrowed from W. Alt (2002, p. 76).

Example 1

- Consider the following example which does not converge although it is clearly descending. Let the criterion function be given by

$f(x) = x^2$, let the starting value be $x^{(0)} = 1$, consider a (constant) direction vector $d^{(k)} = -1$ and choose a step width of $\sigma^{(k)} = (\frac{1}{2})^{k+2}$. Hence the recursive definition of the SEQUENCE[14] $\{x^{(k)}\}_k$ follows as

(2) $x^{(k+1)} = x^{(k)} + (\frac{1}{2})^{k+2}(-1) = x^{(k-1)} - (\frac{1}{2})^{k+1} - (\frac{1}{2})^{k+2} = x^{(0)} - \sum_{j=0}^{k}(\frac{1}{2})^{j+2}$.

Note that $x^{(k)} > 0 \, \forall \, k$ and hence $f(x^{(k+1)}) < f(x^{(k)}) \, \forall \, k$, so that it is clearly a descending method. Nevertheless we find that

(3) $lim_{k\to\infty} x^{(k)} = lim_{k\to\infty} x^{(0)} - \sum_{j=0}^{k-1}(\frac{1}{2})^{j+2} = lim_{k\to\infty} 1 - \frac{1}{4}(\frac{1-(\frac{1}{2})^k}{\frac{1}{2}}) = lim_{k\to\infty} \frac{1}{2} + (\frac{1}{2})^{k+1} = \frac{1}{2} \neq 0 = x^\star$.

The reason for this non-convergence has to be seen in the stepsize $\sigma^{(k)}$ decreasing too fast. For large k the steps $x^{(k+1)} - x^{(k)}$ get so small that convergence is precluded. Hence we have to link the stepsize to the direction of descend $d^{(k)}$.

43.3.3 Efficient Stepsizes

The obvious idea of such a linkage is to require that the actual descent is proportional to a first order approximation, i.e. to choose $\sigma^{(k)}$ such that there is a constant $c_1 > 0$ such that

(4) $f(x^{(k)} + \sigma^{(k)} d^{(k)}) - f(x^{(k)}) \leq c_1 \sigma^{(k)} D(f(x^{(k)})) d^{(k)} < 0$.

Note that we still look only at descending directions, so that $Df(x^{(k)})^T d^{(k)} < 0$ as required in Lemma 1 above. Hence, the compactness of $N(f, f(x^{(k)}))$ implies the CONVERGENCE[15] of the LHS and by (4)

(5) $\lim_{k\to\infty} \sigma^{(k)} D(f(x^{(k)})) d^{(k)} = 0$.

Finally we want to choose a sequence $\{x^{(k)}\}_k$ such that $\lim_{k\to\infty} D(f(x^{(k)})) = 0$ because that is exactly the necessary first order condition we want to solve. Under which conditions does (5) in fact imply $\lim_{k\to\infty} D(f(x^{(k)})) = 0$? First of all the stepsize $\sigma^{(k)}$ must not go to zero too quickly. That is exactly the case we had in the example above. Hence it seems sensible to bound the stepsize from below by requiring that

(6) $\sigma^{(k)} \geq -c_2 \frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||^2} > 0$

for some constant $c_2 > 0$. Substituting (6) into (5) finally yields

(7) $f(x^{(k)} + \sigma^{(k)} d^{(k)}) - f(x^{(k)}) \leq -c(\frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||})^2, \quad c = c_1 c_2$

[14] HTTP://EN.WIKIPEDIA.ORG/WIKI/SEQUENCE

[15] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONVERGENT_SERIES

where again the COMPACTNESS[16] of $N(f, f(x^{(k)}))$ ensures the CONVERGENCE[17] of the LHS and hence

(8) $lim_{k\to\infty} - c(\frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||})^2 = lim_{k\to\infty} \frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||} = 0$

Stepsizes that satisfy (4) and (6) are called *efficient stepsizes* and will be denoted by $\sigma_E^{(k)}$. The importance of condition (6) is illustated in the following continuation of Example 1.

Example 1 (continued)

- Note that it is exactly the failure of (6) that induced Exmaple 1 not to converge. Substituting the stepsize of the example into (6) yields

(6.1) $\sigma^{(k)} = (\frac{1}{2})^{(k+2)} \geq -c_2 \frac{2x^{(k)}(-1)}{1} = c_2 \cdot 2(\frac{1}{2} + (\frac{1}{2})^{k+1}) \Leftrightarrow \frac{1}{4(1+2^{(k)})} \geq c_2 > 0$

so there is *no* constant $c_2 > 0$ satisfying this inequality for all k as required in (6). Hence the stepsize is not bounded from below and decreases too fast. To really acknowledge the importance of (6), let us change the example a bit and assume that $\sigma^{(k)} = (\frac{1}{2})^{k+1}$. Then we find that

(6.2) $lim_{k\to\infty} x^{(k+1)} = lim_{k\to\infty} x^{(0)} - \frac{1}{2}\sum_i (\frac{1}{2})^i = lim_{k\to\infty} (\frac{1}{2})^{k+1} = 0 = x^\star,$

i.e. CONVERGENCE[18] actually does take place. Furthermore recognize that this example actually does satisfy condition (6) as

(6.3) $\sigma^{(k)} = (\frac{1}{2})^{(k+1)} \geq -c_2 \frac{2x^{(k)}(-1)}{1} = c_2 \cdot 2(\frac{1}{2})^k \Leftrightarrow \frac{1}{4} \geq c_2 > 0.$

43.3.4 Choosing the Direction d

We have already argued that the choice of $\sigma^{(k)}$ and $d^{(k)}$ is intertwined. Hence the choice of the "right" $d^{(k)}$ is always contingent on the respective stepsize $\sigma^{(k)}$. So what does "right" mean in this context? Above we showed in equation (8) that choosing an *efficient stepsize* implied

(8′) $lim_{k\to\infty} - c(\frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||})^2 = lim_{k\to\infty} \frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||} = 0.$

The "right" direction vector will therefore guarantee that (8’) implies that

(9) $lim_{k\to\infty} Df(x^{(k)}) = 0$

as (9) is the condition for the chosen sequence $\{x^{(k)}\}_k$ to converge. So let us explore what directions could be chosen to yield (9). Assume that the stepsize σ_k is *efficient* and define

(10) $\beta^{(k)} = \frac{Df(x^{(k)})^T d^{(k)}}{||Df(x^{(k)})|| ||d^{(k)}||} \quad \Leftrightarrow \quad \beta^{(k)} ||Df(x^{(k)})|| = \frac{Df(x^{(k)})^T d^{(k)}}{||d^{(k)}||}$

By (8’) and (10) we know that

(11) $lim_{k\to\infty} \beta^{(k)} ||Df(x^{(k)})|| = 0.$

[16] HTTP://EN.WIKIPEDIA.ORG/WIKI/COMPACT_SPACE

[17] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONVERGENT_SERIES

[18] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONVERGENT_SERIES

So if we bound $\beta^{(k)}$ from below (i.e. $\beta^{(k)} \leq -\delta < 0$), (11) implies that

(12) $lim_{k\to\infty}\beta^{(k)}||Df(x^{(k)})|| = lim_{k\to\infty}||Df(x^{(k)})|| = lim_{k\to\infty}Df(x^{(k)}) = 0,$

where (12) gives just the condition of the sequence $\{x^{(k)}\}_k$ converging to the solution $x^\star$. As (10) defines the direction vector $d^{(k)}$ implicitly by $\beta^{(k)}$, the requirements on $\beta^{(k)}$ translate directly into requirements on $d^{(k)}$.

43.3.5 Why Gradient Methods?

When considering the conditions on $\beta^{(k)}$ it is clear where the term *Gradient Methods* originates from. With $\beta^{(k)}$ given by

$\beta_k = \frac{D(f(x))d^{(k)}}{||Df(x^{(k)})||||d^{(k)}||} = cos(Df(x^{(k)}), d^{(k)})$

we have the following result

Given that $\sigma^{(k)}$ was chosen efficiently and $d^{(k)}$ satisfies

(13) $cos(Df(x^{(k)}), d^{(k)}) = \beta_k \leq -\delta < 0$

we have

(14) $lim_{k\to\infty}Df(x^{(k)}) \to 0$

Hence: Convergence takes place if the angle between the negative gradient at $x^{(k)}$ and the direction $d^{(k)}$ is consistently smaller than the right angle. Methods relying on $d^{(k)}$ satisfying (13) are called Gradient Methods.

In other words: As long as one is not moving ORTHOGONAL[19] to the gradient and if the stepsize is chosen efficiently, *Gradient Methods* guarantee convergence to the solution $x^\star$.

43.3.6 Some Specific Algorithms in the Class of Gradient Methods

Let us now explore three specific algorithms of this class that differ in their respective choice of $d^{(k)}$.

The Newtonian Method

The *Newtonian Method*[20] is by far the most popular method in the field. It is a well known method to solve for the ROOTS[21] of all types of equations and hence can be easily applied to optimization problems as well. The main idea of the Newtonian method is to linearize the system of equations to arrive at

(15) $g(x) = g(\hat{x}) + Dg(\hat{x})^T(x - \hat{x}) = 0.$

[19] HTTP://EN.WIKIPEDIA.ORG/WIKI/ORTHOGONAL

[20] HTTP://EN.WIKIPEDIA.ORG/WIKI/NEWTON_METHOD

[21] HTTP://EN.WIKIPEDIA.ORG/WIKI/ROOT_%28MATHEMATICS%29

(15) can easily be solved for x as the solution is just given by (assuming $Dg(\hat{x})^T$ to be NON-SINGULAR[22])

(16) $x = \hat{x} - [Dg(\hat{x})^T]^{-1} g(\hat{x})$.

For our purposes we just choose $g(x)$ to be the gradient $Df(x)$ and arrive at

(17) $d_N^{(k)} = x^{(k+1)} - x^{(k)} = -[D^2 f(x^{(k)})]^{-1} Df(x^{(k)})$

where $d_N^{(k)}$ is the so-called *Newtonian Direction*.

Properties of the Newtonian Method

Analyzing (17) elicits the main properties of the Newtonian method:

- If $D^2 f(x^{(k)})$ is POSITIVE DEFINITE[23], d_N^k is a direction of descent in the sense of *Lemma 1*.
- The *Newtonian Method* uses local information of the first *and* second derivative to calculate d_N^k.
- As

(18) $x^{(k+1)} = x^{(k)} + d_N^{(k)}$

the *Newtonian Method* uses a fixed stepsize of $\sigma^{(k)} = 1$. Hence the Newtonian method is *not* necessarily a descending method in the sense of *Lemma 1*. The reason is that the fixed stepsize $\sigma^{(k)} = 1$ might be larger than the critical stepsize $\bar{\sigma_k}$ given in *Lemma 1*. Below we provide the Rosenbrock function as an example where the *Newtonian Method* is not descending.

- The Method can be time-consuming as calculating $[D^2 f(x^{(k)})]^{-1}$ for every step k can be cumbersome. In applied work one could think about approximations. One could for example update the Hessian only every sth step or one could rely on *local approximations*. This is known as the *Quasi-Newtonian-Method* and will be discussed in the section about *Variable Metric Methods*.
- To ensure the method to be decreasing one could use an efficient stepsize $\sigma_E^{(k)}$ and set

(19) $x^{(k+1)} = x^{(k)} - \sigma_E^{(k)} d_N^{(k)} = x^{(k)} - \sigma_E^{(k)} [D^2 f(x^k)]^{-1} Df(x^{(k)})$

Method of Steepest Descent

Another frequently used method is the *Method of Steepest Descent*[24]. The idea of this method is to choose the direction $d^{(k)}$ so that the decrease in the function value f is maximal. Although this procedure seems at a first glance very sensible, it suffers from the fact that it uses effectively less information than the *Newtonian Method* by ignoring the Hessian's

[22] HTTP://EN.WIKIPEDIA.ORG/WIKI/SINGULAR_MATRIX
[23] HTTP://EN.WIKIPEDIA.ORG/WIKI/POSITIVE-DEFINITE_MATRIX
[24] HTTP://EN.WIKIPEDIA.ORG/WIKI/STEEPEST_DESCENT

information about the curvature of the function. Especially in the applications below we will see a couple of examples of this problem.

The direction vector of the *Method of Steepest Descent* is given by

(20) $d_{SD}^{(k)} = argmax_{d:||d||=r}\{-Df(x^{(k)})^T d\} = argmin_{d:||d||=r}\{Df(x^{(k)})^T d\} = -r\frac{Df(x)}{||Df(x)||}$

Proof: By the Cauchy-Schwartz Inequality[25] *it follows that*

(21) $\frac{Df(x)^T d}{||Df(x)||||d||} \geq -1 \quad \Leftrightarrow \quad Df(x)^T d \geq -r||Df(x)||.$

Obviously (21) holds with equality for $d^{(k)} = d_{SD}^{(k)}$ given in (20).

Note especially that for $r = ||Df(x)||$ we have $d_{SD}^{(k)} = -Df(x^{(k)})$, i.e. we just "walk" in the direction of the negative gradient. In contrast to the *Newtonian Method* the *Method of Steepest Descent* does not use a fixed stepsize but chooses an efficient stepsize $\sigma_E^{(k)}$. Hence the *Method of Steepest Descent* defines the sequence $\{x^{(k)}\}_k$ by

(22) $x^{(k+1)} = x^{(k)} + \sigma_E^{(k)} d_{SD}^{(k)},$

where $\sigma_E^{(k)}$ is an efficient stepsize and $d_{SD}^{(k)}$ the Direction of Steepest Descent given in (20).

Properties of the Method of Steepest Descent

- With $d_{SD}^{(k)} = -r\frac{Df(x)}{||Df(x)||}$ the Method of Steepest Descent defines a direction of descent in the sense of *Lemma 1*, as

$Df(x)^T d_{SD}^{(k)} = Df(x)^T(-r\frac{Df(x)}{||Df(x)||}) = -\frac{r}{||Df(x)||}Df(x)^T Df(x) < 0.$

- The *Method of Steepest Descent* is only locally sensible as it ignores second order information.
- Especially when the criterion function is flat (i.e. the solution $x^\star$ lies in a "valley") the sequence defined by the Method of Steepest Descent fluctuates wildly (see the applications below, especially the example of the Rosenbrock function).
- As it does not need the Hessian, calculation and implementation of the *Method of Steepest Descent* is easy and fast.

Variable Metric Methods

A more general approach than both the *Newtonian Method* and the *Method of Steepest Descent* is the class of *Variable Metric Methods.* Methods in this class rely on the updating formula

(23) $x^{k+1} = x^k - \sigma_E^{(k)}[A^k]^{-1}Df(x^k).$

[25] http://en.wikipedia.org/wiki/Cauchy-Schwartz_inequality

If A^k is a SYMMETRIC[26] and POSITIVE DEFINITE[27] matrix, (23) defines a descending method as $[A^k]^{-1}$ is positive definite if and only if A^k is positive definite as well. To see this: just consider the SPECTRAL DECOMPOSITION[28]

$$(24) \quad A^k = \Gamma\Lambda\Gamma^T$$

where Γ and Λ are the matrices with EIGENVECTORS[29] and EIGENVALUES[30] respectively. If A^k is positive definite, all eigenvalues λ_i are strictly positive. Hence their inverse λ_i^{-1} are positive as well, so that $[A^k]^{-1} = \Gamma\Lambda^{-1}\Gamma^T$ is clearly positive definite. But then, substitution of $d^{(k)} = [A^k]^{-1}Df(x^k)$ yields

$$(25) \quad Df(x^k)^T d^{(k)} = -Df(x^k)^T[A^k]^{-1}Df(x^k) \equiv -v^T[A^k]^{-1}v \leq 0,$$

i.e. the method is indeed descending. Up to now we have not specified the matrix A^k, but is easily seen that for two specific choices, the *Variable Metric Method* just coincides with the *Method of Steepest Descent* and the *Newtonian Method* respectively.

- For $A^k = \mathcal{I}$ (with $\mathcal{I}$ being the IDENTITY MATRIX[31]) it follows that

$$(22') \quad x^{k+1} = x^k - \sigma_E^{(k)} Df(x^k)$$

which is just the *Method of Steepest Descent.*

- For $A^k = D^2f(x^k)$ it follows that

$$(19') \quad x^{k+1} = x^k - \sigma_E^{(k)}[D^2f(x^k)]^{-1}Df(x^k)$$

which is just the *Newtonian Method* using a stepsize $\sigma_E^{(k)}$.

The Quasi Newtonian Method

A further natural candidate for a *Variable Metric Method* is the *Quasi Newtonian Method.* In contrast to the standard *Newtonian Method* it uses an efficient stepsize so that it is a descending method and in contrast to the *Method of Steepest Descent* it does not fully ignore the local information about the curvature of the function. Hence the *Quasi Newtonian Method* is defined by the two requirements on the matrix A^k:

- A^k should approximate the Hessian $D^2f(x^k)$ to make use of the information about the curvature and
- the update $A^k \to A^{k+1}$ should be easy so that the algorithm is still relatively fast (even in high dimensions).

To ensure the first requirement, A^{k+1} should satisfy the so-called *Quasi-Newtonian-Equation*

$$(26) \quad A^{k+1}(x^{(k+1)} - x^{(k)}) = Df(x^{(k+1)}) - Df(x^{(k)})$$

as all A^k satisfying (26) reflect information about the Hessian. To see this, consider the function $g(x)$ defined as

[26] HTTP://EN.WIKIPEDIA.ORG/WIKI/SYMMETRIC_MATRIX
[27] HTTP://EN.WIKIPEDIA.ORG/WIKI/POSITIVE-DEFINITE_MATRIX
[28] HTTP://EN.WIKIPEDIA.ORG/WIKI/SPECTRAL_DECOMPOSITION
[29] HTTP://EN.WIKIPEDIA.ORG/WIKI/EIGENVECTORS
[30] HTTP://EN.WIKIPEDIA.ORG/WIKI/EIGENVECTORS
[31] HTTP://EN.WIKIPEDIA.ORG/WIKI/IDENTITY_MATRIX

(27) $g(x) = f(x^{k+1}) + Df(x^{k+1})^T(x - x^{k+1}) + \frac{1}{2}(x - x^{k+1})^T A^{k+1}(x - x^{k+1})$.

Then it is obvious that $g(x^{k+1}) = f(x^{k+1})$ and $Dg(x^{k+1}) = Df(x^{k+1})$. So $g(x)$ and $f(x)$ are reasonably similar in the neighborhood of $x^{(k+1)}$. In order to ensure that $g(x)$ is also a good approximation at $x^{(k)}$, we want to choose A^{k+1} such that the gradients at $x^{(k)}$ are identical. With

(28) $Dg(x^k) = Df(x^{k+1}) - A^{k+1}(x^{k+1} - x^k)$

it is clear that $Dg(x^k) = Df(x^k)$ if A^{k+1} satisfies the *Quasi Newtonian Equation* given in (26). But then it follows that

(29) $A^{k+1}(x^{k+1} - x^k) = Df(x^{k+1}) - Dg(x^k) = Df(x^{k+1}) - Df(x^k) = D^2f(\lambda x^{(k)} + (1 - \lambda)x^{(k+1)})(x^{k+1} - x^k)$.

Hence as long as $x^{(k+1)}$ and $x^{(k)}$ are not too far apart, A^{k+1} satisfying (26) is a good approximation of $D^2f(x^{(k)})$.

Let us now come to the second requirement that the update of the A^k should be easy. One specific algorithm to do so is the so-called *BFGS-Algorithm*[32]. The main merit of this algorithm is the fact that it uses only the already calculated elements $\{x^{(k)}\}_k$ and $\{Df(x^{(k)})\}_k$ to construct the update $A^{(k+1)}$. Hence no new entities have to be calculated but one has only to keep track of the x-sequence and sequence of gradients. As a starting point for the BFGS-Algorithm one can provide any positive definite matrix (e.g. the identity matrix or the Hessian at $x^{(0)}$). The *BFGS-Updating-Formula* is then given by

(30) $A^k = A^{k-1} - \frac{(A^{k-1})^T \gamma_{k-1}^T \gamma_{k-1} A^{k-1}}{\gamma_{k-1}^T A^{k-1} \gamma_{k-1}} + \frac{\Delta_{k-1}\Delta_{k-1}^T}{\Delta_{k-1}^T \gamma_{k-1}}$

where $\Delta_{k-1} = Df(x^{(k)}) - Df(x^{(k-1)})$ and $\gamma_{k-1} = x^{(k)} - x^{(k-1)}$. Furthermore (30) ensures that all A^k are positive definite as required by *Variable Metric Methods* to be descending.

Properties of the Quasi Newtonian Method

- It uses second order information about the curvature of $f(x)$ as the matrices A^k are related to the Hessian $D^2f(x)$.
- Nevertheless it ensures easy and fast updating (e.g. by the BFGS-Algorithm) so that it is faster than the standard Newtonian Method.
- It is a descending method as A^k are positive definite.
- It is relatively easy to implement as the BFGS-Algorithm is available in most numerical or statistical software packages.

43.4 Applications

To compare the methods and to illustrate the differences between the algorithms we will now evaluate the performance of the *Steepest Descent Method*, the standard *Newtonian*

[32] HTTP://EN.WIKIPEDIA.ORG/WIKI/BFGS_METHOD

Method and the *Quasi Newtonian Method* with an efficient stepsize. We use two classical functions in this field, namely the Himmelblau and the Rosenbrock function.

43.4.1 Application I: The Himmelblau Function

The Himmelblau function is given by

(31) $f(x,y) = (x^2+y-11)^2+(x+y^2-7)^2$

This fourth order polynomial has four minima, four saddle points and one maximum so there are enough possibilities for the algorithms to fail. In the following pictures we display the CONTOUR PLOT[33] and the 3D plot of the function for different starting values.

In Figure 1 we display the function and the paths of all three methods at a starting value of $(2,-4)$. Obviously the three methods do not find the same minimum. The reason is of course the different direction vector of the *Method of Steepest Descent* - by ignoring the information about the curvature it chooses a totally different direction than the two *Newtonian Methods* (see especially the right panel of Figure 1).

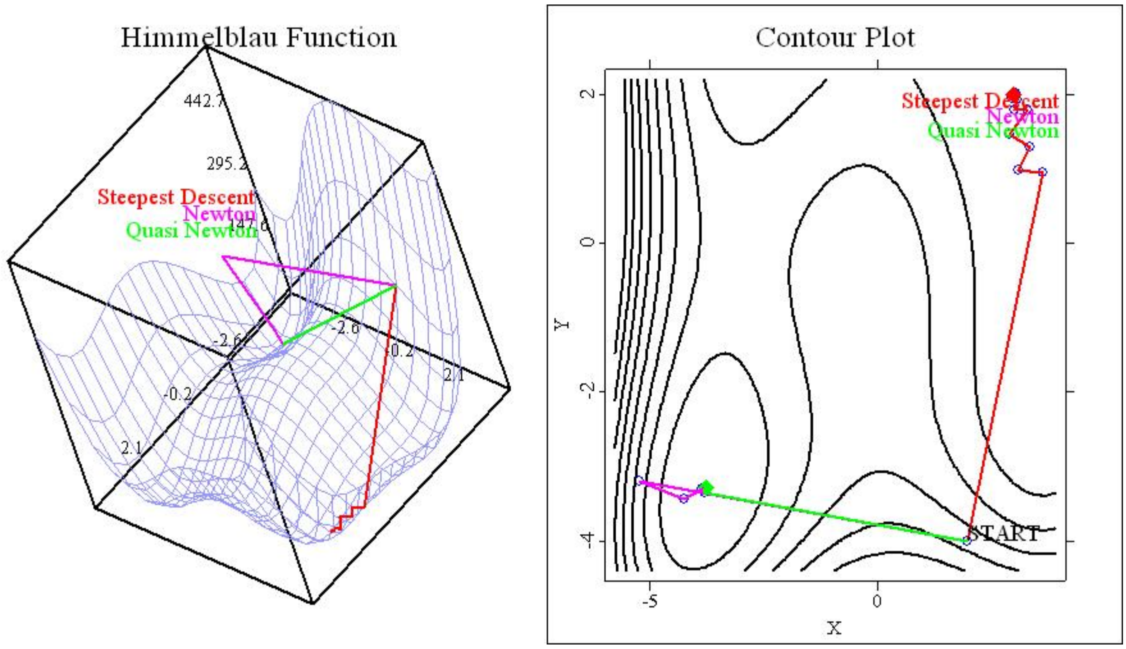


Figure 19: Figure 1: The two Newton Methods converge to the same, the Method of Steepest Descent to a different minimum.

Consider now the starting value $(4.5,-0.5)$, displayed in Figure 2. The most important thing is of course that now *all* methods find different solutions. That the *Method of Steepest Descent* finds a different solution than the two *Newtonian Methods* is again not that suprising. But that the two *Newtonian Methods* converge to different solution shows the significance of the stepsize σ. With the *Quasi-Newtonian Method* choosing an efficient stepsize in the first iteration, both methods have different stepsizes *and* direction vectors for

[33] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONTOUR_LINE

all iterations after the first one. And as seen in the picture: the consequence may be quite significant.

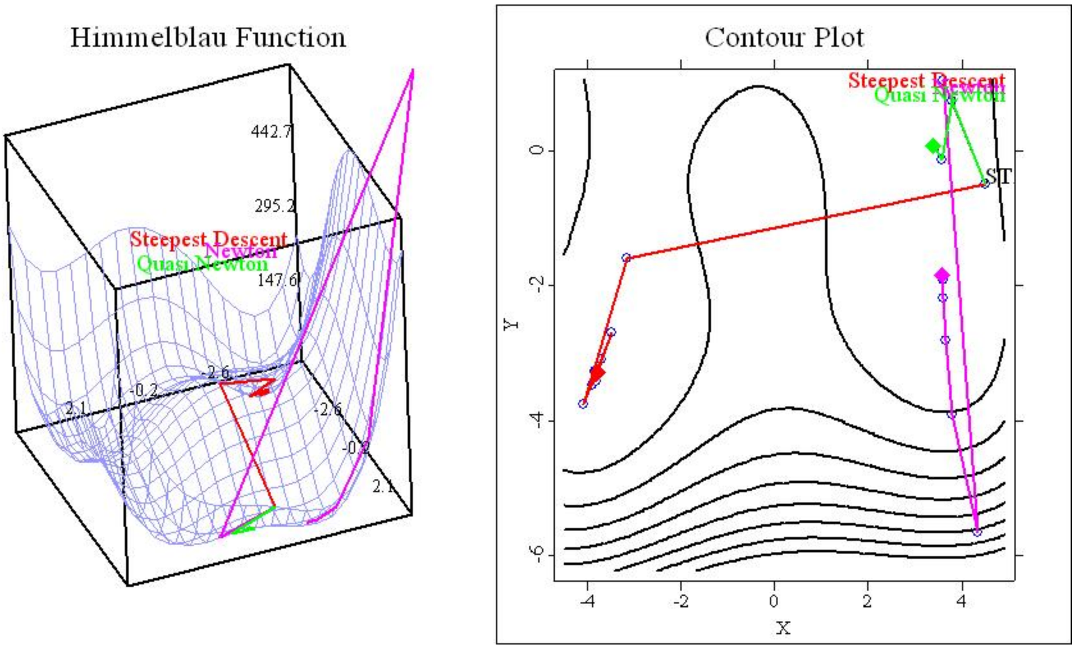


Figure 20: Figure 2: Even all methods find different solutions.

43.4.2 Application II: The Rosenbrock Function

The Rosenbrock function is given by

(32) $f(x,y) = 100(y-x^2)^2 + (1-x)^2$

Although this function has only one minimum it is an interesting function for optimization problems. The reason is the very flat valley of this U-shaped function (see the right panels of Figures 3 and 4). Especially for econometricians[34] this function may be interesting because in the case of Maximum Likelihood estimation flat criterion functions occur quite frequently. Hence the results displayed in Figures 3 and 4 below seem to be rather generic for functions sharing this problem.

My experience when working with this function and the algorithms I employed is that Figure 3 (given a starting value of $(2,-5)$) seems to be quite characteristic. In contrast to the Himmelblau function above, all algorithms found the same solution and given that there is only one minimum this could be expected. More important is the path the different methods choose as is reflects the different properties of the respective methods. It is seen that the *Method of Steepest Descent* fluctuates rather wildly. This is due to the fact that it does not use information about the curvature but rather jumps back and forth between the "hills" adjoining the valley. The two Newtonian Methods choose a more direct path as they use the second order information. The main difference between the two Newtonian

[34] http://en.wikipedia.org/wiki/Econometrics

Methods is of course the stepsize. Figure 3 shows that the *Quasi Newtonian Method* uses very small stepsizes when working itself through the valley. In contrast, the stepsize of the *Newtonian Method* is fixed so that it jumps directly in the direction of the solution. Although one might conclude that this is a disadvantage of the *Quasi Newtonian Method*, note of course that in general these smaller stepsizes come with benefit of a higher stability, i.e. the algorithm is less likely to jump to a different solution. This can be seen in Figure 4.

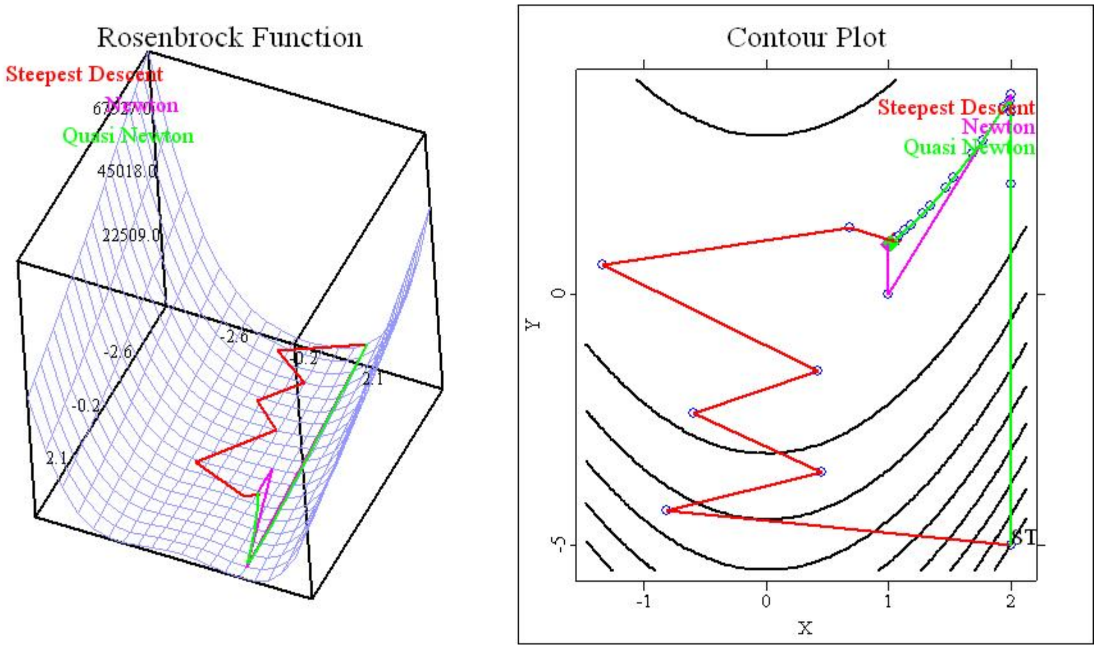


Figure 21: Figure 3: All methods find the same solution, but the Method of Steepest Descent fluctuates heavily.

Figure 4, which considers a starting value of $(-2, -2)$, shows the main problem of the *Newtonian Method* using a fixed stepsize - the method might "overshoot" in that it is not descending. In the first step, the *Newtonian Method* (displayed as the purple line in the figure) jumps out of the valley to only bounce back in the next iteration. In this case convergence to the minimum still occurs as the gradient at each side points towards the single valley in the center, but one can easily imagine functions where this is not the case. The reason of this jump are the second derivatives which are very small so that the step $[Df(x^{(k)})]^{-1}Df(x^{(k)}))$ gets very large due to the inverse of the Hessian. In my experience I would therefore recommend to use efficient stepsizes to have more control over the paths the respective Method chooses.

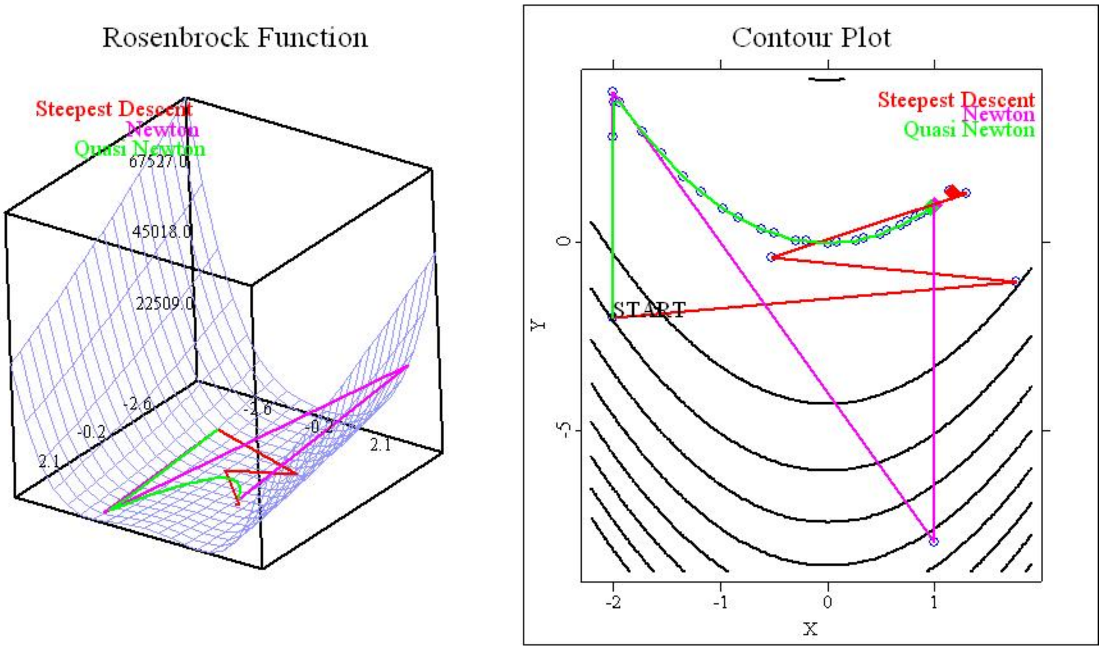


Figure 22: Figure 2: Overshooting of the Newtonian Method due to the fixed stepsize.

43.4.3 Application III: Maximum Likelihood Estimation

For econometricians and statisticians the MAXIMUM LIKELIHOOD ESTIMATOR[35] is probably the most important application of numerical optimization algorithms. Therefore we will briefly show how the estimation procedure fits in the framework developed above.

As usual let

(33) $f(Y|X;\theta)$

be the CONDITIONAL DENSITY[36] of Y given X with parameter θ and

(34) $l(\theta;Y|X)$

the CONDITIONAL LIKELIHOOD FUNCTION[37] for the parameter θ

If we assume the data to be INDEPENDENTLY, IDENTICALLY DISTRIBUTED (IID)[38] then the sample log-likelihood follows as

(35) $\mathcal{L}(\theta;Y_1,...,Y_N) = \sum_i^N \mathcal{L}(\theta;Y_i) = \sum_i^N log(l(\theta;Y_i)).$

Maximum Likelihood estimation therefore boils down to maximize (35) with respect to the parameter θ. If we for simplicity just decide to use the *Newtonian Method* to solve that problem, the sequence $\{\theta^{(k)}\}_k$ is recursively defined by

[35] HTTP://EN.WIKIPEDIA.ORG/WIKI/MAXIMUM_LIKELIHOOD

[36] HTTP://EN.WIKIPEDIA.ORG/WIKI/CONDITIONAL_DISTRIBUTION

[37] HTTP://EN.WIKIPEDIA.ORG/WIKI/LIKELIHOOD_FUNCTION

[38] HTTP://EN.WIKIPEDIA.ORG/WIKI/IID

(36) $D_\theta\mathcal{L}(\theta^{(k+1)}) = D_\theta\mathcal{L}(\theta^{(k)}) + D_{\theta\theta}\mathcal{L}(\theta^{(k)})(\theta^{(k+1)} - \theta^{(k)}) = 0 \Leftrightarrow \theta^{(k+1)} = \theta^{(k)} - [D_{\theta\theta}\mathcal{L}(\theta^{(k)})]^{-1}D_\theta\mathcal{L}(\theta^{(k)})$

where $D_\theta\mathcal{L}$ and $D_{\theta\theta}\mathcal{L}$ denotes the first and second derivative with respect to the parameter vector θ and $[D_{\theta\theta}\mathcal{L}(\theta^{(k)})]^{-1}D_\theta\mathcal{L}(\theta^{(k)})$ defines the *Newtonian Direction* given in (17). As Maximum Likelihood estimation always assumes that the conditional density (i.e. the distribution of the error term) is known up to the parameter θ, the methods described above can readily be applied.

A Concrete Example of Maximum Likelihood Estimation

Assume a simple linear model

(37*a*) $Y_i = \beta_1 + \beta_x X_i + U_i$

with $\theta = (\beta_1, \beta_2)'$. The conditional distribution Y is then determined by the one of U, i.e.

(37*b*) $p(Y_i - \beta_1 - \beta_x X_i) \equiv p_{|X_i}(Y_i) = p(U_i)$,

where p denotes the DENSITY FUNCTION[39]. Generally, there is no closed form solution of maximizing (35) (at least if U does not happen to be NORMALLY DISTRIBUTED[40]), so that numerical methods have to be employed. Hence assume that U follows STUDENT'S T-DISTRIBUTION[41] with m DEGREES OF FREEDOM[42] so that (35) is given by

(38) $\mathcal{L}(\theta; Y_{|X}) = \sum log(\frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi m}\Gamma(\frac{m}{2})}(1 + \frac{(y_i - x_i^T\beta)^2}{m})^{-\frac{m+1}{2}})$

where we just used the definition of the density function of the t-distribution. (38) can be simplified to

(39) $\mathcal{L}(\theta; Y_{|X}) = N[log(\Gamma(\frac{m+1}{2})) - log(\sqrt{\pi m}\Gamma(\frac{m}{2}))] - \frac{m+1}{2}\sum log(1 + \frac{(y_i - x_i^T\beta)^2}{m})$

so that (if we assume that the degrees of freedom m are known)

(40) $argmax\{\mathcal{L}(\theta; Y_{|X})\} = argmax\{-\frac{m+1}{2}\sum log(1 + \frac{(y_i - x_i^T\beta)^2}{m})\} = argmin\{\sum log(1 + \frac{(y_i - x_i^T\beta)^2}{m})\}$.

With the criterion function

(41) $f(\beta_1, \beta_2) = \sum log(1 + \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{m})$

the methods above can readily applied to calculate the Maximum Likelihood Estimator $(\hat{\beta}_{1,ML}, \hat{\beta}_{2,ML})$ maximizing (41).

43.5 References

- Alt, W. (2002): "Nichtlineare Optimierung", Vieweg: Braunschweig/Wiesbaden

[39] HTTP://EN.WIKIPEDIA.ORG/WIKI/DENSITY_FUNCTION
[40] HTTP://EN.WIKIPEDIA.ORG/WIKI/NORMAL_DISTRIBUTION
[41] HTTP://EN.WIKIPEDIA.ORG/WIKI/STUDENT%27S_T-DISTRIBUTION
[42] HTTP://EN.WIKIPEDIA.ORG/WIKI/DEGREES_OF_FREEDOM_%28STATISTICS%29

- Härdle, W. and Simar, L. (2003): "Applied Multivariate Statistical Analysis", Springer: Berlin Heidelberg
- Königsberger, K. (2004): "Analysis I", Springer: Berlin Heidelberg
- Ruud, P. (2000): "Classical Econometric Theory", Oxford University Press: New York

44 Quantile Regression

Quantile Regression as introduced by Koenker and Bassett (1978) seeks to complement classical linear regression analysis. Central hereby is the extension of "ordinary quantiles from a location model to a more general class of linear models in which the conditional quantiles have a linear form" (Buchinsky (1998), p. 89). In Ordinary Least Squares (OLS[1]) the primary goal is to determine the conditional mean of random variable Y, given some explanatory variable x_i, reaching the expected value $E[Y|x_i]$. Quantile Regression goes beyond this and enables one to pose such a question at any quantile of the conditional distribution function. The following seeks to introduce the reader to the ideas behind Quantile Regression. First, the issue of QUANTILES[2] is addressed, followed by a brief outline of least squares estimators focusing on Ordinary Least Squares. Finally, Quantile Regression is presented, along with an example utilizing the Boston Housing data set.

44.1 Preparing the Grounds for Quantile Regression

44.1.1 What are Quantiles

Gilchrist (2001, p.1) describes a quantile as "simply the value that corresponds to a specified proportion of an (ordered) sample of a population". For instance a very commonly used quantile is the MEDIAN[3] M, which is equal to a proportion of 0.5 of the ordered data. This corresponds to a quantile with a probability of 0.5 of occurrence. Quantiles hereby mark the boundaries of equally sized, consecutive subsets. (Gilchrist, 2001)

More formally stated, let Y be a continuous random variable with a distribution function $F_Y(y)$ such that

$(1) F_Y(y) = P(Y \leq y) = \tau$

which states that for the distribution function $F_Y(y)$ one can determine for a given value y the probability τ of occurrence. Now if one is dealing with quantiles, one wants to do the opposite, that is one wants to determine for a given probability τ of the sample data set the corresponding value y. A $\tau^{th}-$quantile refers in a sample data to the probability τ for a value y.

$(2) F_Y(y_\tau) = \tau$

Another form of expressing the $\tau^{th}-$quantile mathematically is following:

$(3) y_\tau = F_Y^{-1}(\tau)$

[1] HTTP://EN.WIKIPEDIA.ORG/WIKI/OLS

[2] HTTP://EN.WIKIPEDIA.ORG/WIKI/QUANTILES

[3] HTTP://EN.WIKIPEDIA.ORG/WIKI/MEDIAN

y_τ is such that it constitutes the inverse of the function $F_Y(\tau)$ for a probability τ.

Note that there are two possible scenarios. On the one hand, if the distribution function $F_Y(y)$ is monotonically increasing, quantiles are well defined for every $\tau \in (0;1)$. However, if a distribution function $F_Y(y)$ is not strictly monotonically increasing , there are some τs for which a unique quantile can not be defined. In this case one uses the smallest value that y can take on for a given probability τ.

Both cases, with and without a strictly monotonically increasing function, can be described as follows:

$(4) y_\tau = F_Y^{-1}(\tau) = inf\{y|F_Y(y) \geq \tau\}$

That is y_τ is equal to the inverse of the function $F_Y(\tau)$ which in turn is equal to the infimum of y such that the distribution function $F_Y(y)$ is greater or equal to a given probability τ, i.e. the $\tau^{th}-$quantile. (Handl (2000))

However, a problem that frequently occurs is that an empirical distribution function is a step function. Handl (2000) describes a solution to this problem. As a first step, one reformulates equation 4 in such a way that one replaces the continuous random variable Y with n, the observations, in the distribution function $F_Y(y)$, resulting in the empirical distribution function $F_n(y)$. This gives the following equation:

$(5) \hat{y}_\tau = inf\{y|F_n(y) \geq \tau\}$

The empirical distribution function can be separated into equally sized, consecutive subsets via the the number of observations n. Which then leads one to the following step:

$(6) \hat{y}_\tau = y_{(i)}$

with $i = 1,...,n$ and $y_{(1)},...,y_{(n)}$ as the sorted observations. Hereby, of course, the range of values that y_τ can take on is limited simply by the observations $y_{(i)}$ and their nature. However, what if one wants to implement a different subset, i.e. different quantiles but those that can be derived from the number of observations n?

Therefore a further step necessary to solving the problem of a step function is to smooth the empirical distribution function through replacing it a with continuous linear function $\tilde{F}(y)$. In order to do this there are several algorithms available which are well described in Handl (2000) and more in detail with an evaluation of the different algorithms and their efficiency in computer packages in Hyndman and Fan (1996). Only then one can apply any division into quantiles of the data set as suitable for the purpose of the analysis. (Handl (2000))

44.1.2 Ordinary Least Squares

In regression analysis the researcher is interested in analyzing the behavior of a dependent variable y_i given the information contained in a set of explanatory variables x_i. Ordinary Least Squares is a standard approach to specify a linear regression model and estimate its unknown parameters by minimizing the sum of squared errors. This leads to an approximation of the mean function of the conditional distribution of the dependent variable. OLS achieves the property of BLUE, it is the best, linear, and unbiased estimator, if following four assumptions hold:

1. The explanatory variable x_i is non-stochastic
2. The expectations of the error term ϵ_i are zero, i.e. $E[\epsilon_i] = 0$
3. Homoscedasticity - the variance of the error terms ϵ_i is constant, i.e. $var(\epsilon_i) = \sigma^2$
4. No autocorrelation, i.e. $cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$

However, frequently one or more of these assumptions are violated, resulting in that OLS is not anymore the best, linear, unbiased estimator. Hereby Quantile Regression can tackle following issues: (i), frequently the error terms are not necessarily constant across a distribution thereby violating the axiom of homoscedasticity. (ii) by focusing on the mean as a measure of location, information about the tails of a distribution are lost. (iii) OLS is sensitive to extreme outliers that can distort the results significantly. (Montenegro (2001))

44.2 Quantile Regression

44.2.1 The Method

Quantile Regression essentially transforms a conditional distribution function into a conditional quantile function by slicing it into segments. These segments describe the cumulative distribution of a conditional dependent variable Y given the explanatory variable x_i with the use of quantiles as defined in equation 4.

For a dependent variable Y given the explanatory variable $X = x$ and fixed τ, $0 < \tau < 1$, the conditional quantile function is defined as the $\tau - th$ quantile $Q_{Y|X}(\tau|x)$ of the conditional distribution function $F_{Y|X}(y|x)$. For the estimation of the location of the conditional distribution function, the conditional median $Q_{Y|X}(0,5|x)$ can be used as an alternative to the conditional mean. (Lee (2005))

One can nicely illustrate Quantile Regression when comparing it with OLS. In OLS, modeling a conditional distribution function of a random sample $(y_1, ..., y_n)$ with a parametric function $\mu(x_i, \beta)$ where x_i represents the independent variables, β the corresponding estimates and μ the conditional mean, one gets following minimization problem:

$(7) min_{\beta \in \Re} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2$

One thereby obtains the conditional expectation function $E[Y|x_i]$. Now, in a similar fashion one can proceed in Quantile Regression. Central feature thereby becomes ρ_τ, which serves as a check function.

$$(8) \rho_\tau(x) = \begin{cases} \tau * x & \text{if } x \geq 0 \\ (\tau - 1) * x & \text{if } x < 0 \end{cases}$$

This check-function ensures that

1. all ρ_τ are positive
2. the scale is according to the probability τ

Such a function with two supports is a must if dealing with L1 distances, which can become negative.

In Quantile Regression one minimizes now following function:

$(9) min_{\beta \in \Re} \sum_{i=1}^{n} \rho_\tau (y_i - \xi(x_i, \beta))$

Here, as opposed to OLS, the minimization is done for each subsection defined by ρ_τ, where the estimate of the τ^{th}-quantile function is achieved with the parametric function $\xi(x_i, \beta)$. (Koenker and Hallock (2001))

Features that characterize Quantile Regression and differentiate it from other regression methods are following:

1. The entire conditional distribution of the dependent variable Y can be characterized through different values of τ

2. Heteroscedasticity can be detected

3. If the data is heteroscedastic, median regression estimators can be more efficient than mean regression estimators

4. The minimization problem as illustrated in equation 9 can be solved efficiently by linear programming methods, making estimation easy

5. Quantile functions are also equivariant to monotone transformations. That is $Q_{h(Y|X)}(x_\tau) = h(Q_{(Y|X)}(x_\tau))$, for any function

6. Quantiles are robust in regards to outliers (Lee (2005))

44.2.2 A graphical illustration of Quantile Regression

Before proceeding to a numerical example, the following subsection seeks to graphically illustrate the concept of Quantile Regression. First, as a starting point for this illustration, consider figure 1. For a given explanatory value of x_i the density for a conditional dependent variable Y is indicated by the size of the balloon. The bigger the balloon, the higher is the density, with the MODE[4], i.e. where the density is the highest, for a given x_i being the biggest balloon. Quantile Regression essentially connects the equally sized balloons, i.e. probabilities, across the different values of x_i, thereby allowing one to focus on the interrelationship between the explanatory variable x_i and the dependent variable Y for the different quantiles, as can be seen in figure 2. These subsets, marked by the quantile lines, reflect the probability density of the dependent variable Y given x_i.

[4] HTTP://EN.WIKIPEDIA.ORG/WIKI/MODE

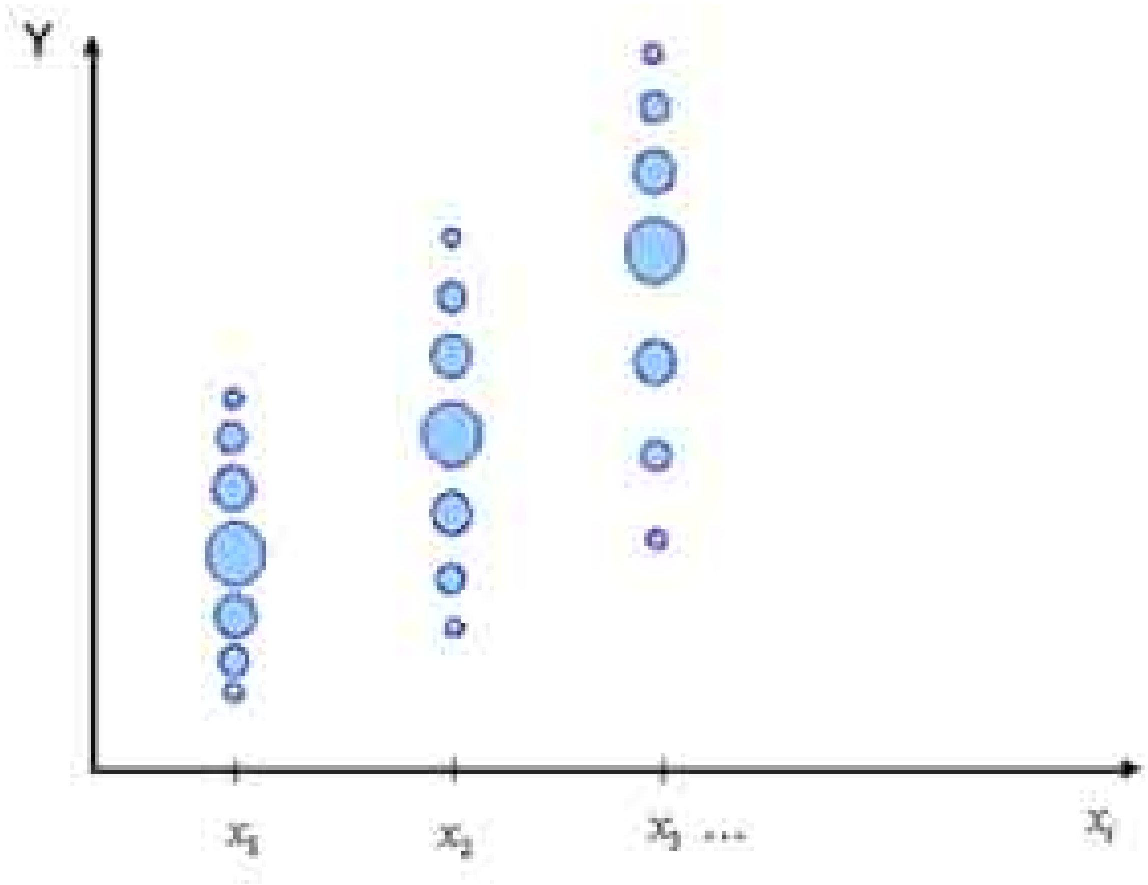


Figure 23: Figure 1: Probabilities of occurrence for individual explanatory variables

The example used in figure 2 is originally from Koenker and Hallock (2000), and illustrates a classical empirical application, Ernst Engel's (1857) investigation into the relationship of household food expenditure, being the dependent variable, and household income as the explanatory variable. In Quantile Regression the conditional function of $Q_{Y|X}(\tau|x)$ is segmented by the τ^{th}-quantile. In the analysis, the τ^{th}-quantiles $\tau \in \{0,05;0,1;0,25;0,5;0,75;0,9;0,95\}$, indicated by the thin blue lines that separate the different color sections, are superimposed on the data points. The conditional median $(\tau = 0,5)$ is indicated by a thick dark blue line, the conditional mean by a light yellow line. The color sections thereby represent the subsections of the data as generated by the quantiles.

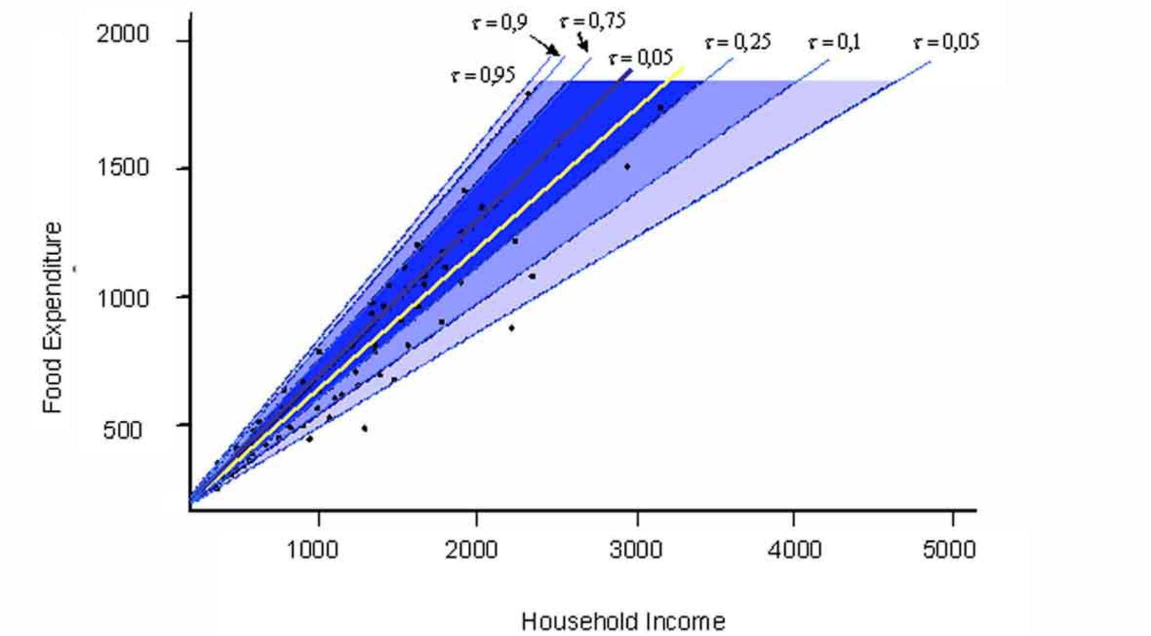


Figure 24: Figure 2: Engels Curve, with the median highlighted in dark blue and the mean in yellow

Figure 2 can be understood as a contour plot representing a 3-D graph, with food expenditure and income on the respective y and x axis. The third dimension arises from the probability density of the respective values. The density of a value is thereby indicated by the darkness of the shade of blue, the darker the color, the higher is the probability of occurrence. For instance, on the outer bounds, where the blue is very light, the probability density for the given data set is relatively low, as they are marked by the quantiles 0,05 to 0,1 and 0,9 to 0,95. It is important to notice that figure 2 represents for each subsections the individual probability of occurrence, however, quantiles utilize the cumulative probability of a conditional function. For example, τ of 0,05 means that 5% of observations are expected to fall below this line, a τ of 0,25 for instance means that 25% of the observations are expected to fall below this and the 0,1 line.

The graph in figure 2, suggests that the error variance is not constant across the distribution. The dispersion of food expenditure increases as household income goes up. Also the data is skewed to the left, indicated by the spacing of the quantile lines that decreases above the median and also by the relative position of the median which lies above the mean. This suggests that the axiom of homoscedasticity is violated, which OLS relies on. The statistician is therefore well advised to engage in an alternative method of analysis such as Quantile Regression, which is actually able to deal with heteroscedasticity.

44.2.3 A Quantile Regression Analysis

In order to give a numerical example of the analytical power of Quantile Regression and to compare it within the boundaries of a statistical application with OLS the following section will be analyzing some selected variables of the Boston Housing dataset which is available at the md-base website. The data was first analyzed by Belsley, Kuh, and Welsch (1980).

The original data comprised 506 observations for 14 variables stemming from the census of the Boston metropolitan area.

This analysis utilizes as the dependent variable the median value of owner occupied homes (a metric variable, abbreviated with H) and investigates the effects of 4 independent variables as shown in table 1. These variables were selected as they best illustrate the difference between OLS and Quantile Regression. For the sake of simplicity of the analysis, it was neglected for now to deal with potential difficulties related to finding the correct specification of a parametric model. A simple linear regression model therefore was assumed. For the estimation of asymptotic standard errors see for example Buchinsky (1998), which illustrates the design-matrix bootstrap estimator or alternatively Powell (1986) for kernel based estimation of asymptotic standard errors.

Table1: The explanatory variablesName	**Short**	**What it is**	**type**
NonrTail	T	Proportion of non-retail business acres	metric
NoorOoms	O	Average number of rooms per dwelling	metric
Age	A	Proportion of owner-built dwellings prior to 1940	metric
PupilTeacher	P	Pupil-teacher ratio	metric

In the following firstly an OLS model was estimated. Three digits after the comma were indicated in the tables as some of the estimates turned out to be very small.

(10)$E[H_i|T_i,O_i,A_i,P_i] = \alpha + \beta T_i + \delta O_i + \gamma A_i + \lambda P_i$

Computing this via XploRe one obtains the results as shown in the table below.

Table2: OLS estimates$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\gamma}$	$\hat{\lambda}$
36,459	0,021	38,010	0,001	-0,953

Analyzing this data set via Quantile Regression, utilizing the τ^{th} quantiles $\tau \in (0,1;0,3;0,5;0,7;0,9)$ the model is characterized as follows:

(11)$Q_H[\tau|T_i,O_i,A_i,P_i] = \alpha_\tau + \beta_\tau T_i + \delta_\tau O_i + \gamma_\tau A_i + \lambda_\tau P_i$

Just for illustrative purposes and to further foster the understanding of the reader for Quantile Regression, the equation for the $0,1^{th}$ quantile is briefly illustrated, all others follow analogous:

$$(12) min\,[\rho_{0,1}(y_1 - x_1\beta) + \rho_{0,1}(y_2 - x_2\beta) + ... + \rho_{0,1}(y_n - x_n\beta)]$$

equation 12 with $\rho_{0,1}(y_i - x_i\beta) = \begin{cases} 0,1(y_i - x_i\beta) & \text{if } (y_i - x_i\beta) > 0 \\ -0,9(y_i - x_i\beta) & \text{if } (y_i - x_i\beta) < 0 \end{cases}$

Table3: Quantile Regression estimatesτ	$\hat{\alpha}_\tau$	$\hat{\beta}_\tau$	$\hat{\delta}_\tau$	$\hat{\gamma}_\tau$	$\hat{\lambda}_\tau$
0,1	23,442	0,087	29,606	-0,022	-0,443
0,3	15,7130	-0,001	45,281	-0,037	-0,617
0,5	14,8500	0,022	53,252	-0,031	-0,737
0,7	20,7910	-0,021	50,999	-0,003	-0,925
0,9	34,0310	-0,067	51,353	0,004	-1,257

Now if one compares the results for the estimates of OLS from table 2 and Quantile Regression, table 3, one finds that the latter method can make much more subtle inferences of the effect of the explanatory variables on the dependent variable. Of particular interest are thereby quantile estimates that are relatively different as compared to other quantiles for the same estimate.

Probably the most interesting result and most illustrative in regards to an understanding of the functioning of Quantile Regression and pointing to the differences with OLS are the results for the independent variable of the proportion of non-retail business acres (T_i). OLS indicates that this variable has a positive influence on the dependent variable, the value of homes, with an estimate of $\hat{\beta} = 0,021$, i.e. the value of houses increases as the proportion of non-retail business acres (T_i) increases in regards to the Boston Housing data.

Looking at the output that Quantile Regression provides us with, one finds a more differentiated picture. For the 0,1 quantile, we find an estimate of $\hat{\beta}_{0,1} = 0,087$ which would suggest that for this low quantile the effect seems to be even stronger than is suggested by OLS. Here house prices go up when the proportion of non-retail businesses (T_i) goes up, too. However, considering the other quantiles, this effect is not quite as strong anymore, for the 0,7th and 0,9th quantile this effect seems to be even reversed indicated by the parameter $\hat{\beta}_{0,7} = -0,021$ and $\hat{\beta}_{0,9} = -0,062$. These values indicate that in these quantiles the house price is negatively influenced by an increase of non-retail business acres (T_i). The influence of non-retail business acres (T_i) seems to be obviously very ambiguous on the dependent variable of housing price, depending on which quantile one is looking at. The general recommendation from OLS that if the proportion of non-retail business acres (T_i) increases, the house prices would increase can obviously not be generalized. A policy recommendation on the OLS estimate could therefore be grossly misleading.

One would intuitively find the statement that the average number of rooms of a property (O_i) positively influences the value of a house, to be true. This is also suggested by OLS with an estimate of $\hat{\delta} = 38,099$. Now Quantile Regression also confirms this statement, however, it also allows for much subtler conclusions. There seems to be a significant difference between the 0,1 quantile as opposed to the rest of the quantiles, in particular the 0,9th quantile. For the lowest quantile the estimate is $\hat{\delta}_{0,1} = 29,606$, whereas for the 0,9th quantile it is $\hat{\delta}_{0,9} = 51,353$. Looking at the other quantiles one can find similar values for the Boston housing data set as for the 0,9th, with estimates of $\hat{\delta}_{0,3} = 45,281$, $\hat{\delta}_{0,5} = 53,252$, and $\hat{\delta}_{0,7} = 50,999$ respectively. So for the lowest quantile the influence of additional number of rooms (O_i) on the house price seems to be considerably smaller then for all the other quantiles.

Another illustrative example is provided analyzing the proportion of owner-occupied units built prior to 1940 (A_i) and its effect on the value of homes. Whereas OLS would indicate this variable has hardly any influence with an estimate of $\hat{\gamma} = 0,001$, looking at Quantile Regression one gets a different impression. For the 0,1th quantile, the age has got a negative influence on the value of the home with $\hat{\gamma}_{0,1} = -0,022$. Comparing this with the highest quantile where the estimate is $\hat{\gamma}_{0,9} = 0,004$, one finds that the value of the house is suddenly now positively influenced by its age. Thus, the negative influence is confirmed by all other quantiles besides the highest, the 0,9th quantile.

Last but not least, looking at the pupil-teacher ratio (P_i) and its influence on the value of houses, one finds that the tendency that OLS indicates with a value of $\hat{\lambda} = -0,953$ to be also reflected in the Quantile Regression analysis. However, in Quantile Regression one can see that the influence on the housing price of the pupils-teacher ratio (P_i) gradually increases over the different quantiles, from the 0,1th quantile with an estimate of $\hat{\lambda}_{0,1} = -0,443$ to the 0,9th quantile with a value of $\hat{\lambda}_{0,9} = -1,257$.

This analysis makes clear, that Quantile Regression allows one to make much more differentiated statements when using Quantile Regression as opposed to OLS. Sometimes OLS estimates can even be misleading what the true relationship between an explanatory and a dependent variable is as the effects can be very different for different subsection of the sample.

44.3 Conclusion

For a distribution function $F_Y(y)$ one can determine for a given value of y the probability τ of occurrence. Now quantiles do exactly the opposite. That is, one wants to determine for a given probability τ of the sample data set the corresponding value y. In OLS, one has the primary goal of determining the conditional mean of random variable Y, given some explanatory variable x_i , $E[Y|x_i]$. Quantile Regression goes beyond this and enables us to pose such a question at any quantile of the conditional distribution function. It focuses on the interrelationship between a dependent variable and its explanatory variables for a given quantile. Quantile Regression overcomes thereby various problems that OLS is confronted with. Frequently, error terms are not constant across a distribution, thereby violating the axiom of homoscedasticity. Also, by focusing on the mean as a measure of location, information about the tails of a distribution are lost. And last but not least, OLS is sensitive to extreme outliers, which can distort the results significantly. As has been indicated in the small example of the Boston Housing data, sometimes a policy based upon an OLS analysis might not yield the desired result as a certain subsection of the population does not react as strongly to this policy or even worse, responds in a negative way, which was not indicated by OLS.

44.4 References

Abrevaya, J. (2001): “The effects of demographics and maternal behavior on the distribution of birth outcomes,” in Economic Application of Quantile Regression, ed. by B. Fitzenberger, R. Koenker, and J. A. Machade, pp. 247–257. Physica-Verlag Heidelberg, New York.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980): Applied Multivariate Statistical Analysis. Regression Diagnostics, Wiley.

Buchinsky, M. (1998): “Recent Advances in Quantile Regression Models: A Practical Guidline for Empirical Research,” Journal of Human Resources, 33(1), 88–126.

Cade, B.S. and B.R. Noon (2003): A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1(8): 412-420. http://www.fort.usgs.gov/products/publications/21137/21137.pdf

Cizek, P. (2003): “Quantile Regression,” in XploRe Application Guide, ed. by W. Härdle, Z. Hlavka, and S. Klinke, chap. 1, pp. 19–48. Springer, Berlin.

Curry, J., and J. Gruber (1996): “Saving Babies: The Efficacy and Costs of Recent Changes in the Medicaid Eligibility of Pregnant Women,” Journal of Political Economy, 104, 457–470.

Handl, A. (2000): “Quantile,” available at http://www.wiwi.uni-bielefeld.de/˜frohn/Lehre/Datenanalyse/Skript/daquantile.pdf

Härdle, W. (2003): Applied Multivariate Statistical Analysis. Springer Verlag, Heidelberg.
Hyndman, R. J., and Y. Fan (1996): “Sample Quantiles in Statistical Packages,” The American Statistician, 50(4), 361 – 365.

Jeffreys, H., and B. S. Jeffreys (1988): Upper and Lower Bounds. Cambridge University Press.

Koenker, R., and G. W. Bassett (1978): “Regression Quantiles,” Econometrica, 46, 33–50.

Koenker, R., and G. W. Bassett (1982): “Robust tests for heteroscedasticity based on Regression Quantiles,” Econometrica, 61, 43–61.

Koenker, R., and K. F. Hallock (2000): “Quantile Regression an Introduction,” available at http://www.econ.uiuc.edu/˜roger/research/intro/intro.html

Koenker, R., and K. F. Hallock (2001): “Quantile Regression,” Journal of Economic Perspectives, 15(4), 143–156.

Lee, S. (2005): “Lecture Notes for MECT1 Quantile Regression,” available at http://www.homepages.ucl.ac.uk/˜uctplso/Teaching/MECT/lecture8.pdf

Lewit, E. M., L. S. Baker, H. Corman, and P. Shiono (1995): “The Direct Costs of Low Birth Weight,” The Future of Children, 5, 35–51.

mdbase (2005): “Statistical Methodology and Interactive Datanalysis,” available at http://www.quantlet.org/mdbase/

Montenegro, C. E. (2001): “Wage Distribution in Chile: Does Gender Matter? A Quantile Regression Approach,” Working Paper Series 20, The World Bank, Development Research Group.

Powell, J. (1986): “Censored Regression Quantiles,” Journal of Econometrics, 32, 143– 155.

Scharf, F. S., F. Juanes, and M. Sutherland (1998): “Inferring Ecologiocal Relationships from the Edges of Scatter Diagrams: Comparison of Regression Techniques,” Ecology, 79(2), 448–460.

XploRe (2006): “XploRe,” available at http://www.xplore-stat.de/index_js.html

45 Numerical Comparison of Statistical Software

45.1 Introduction

Statistical computations require an extra accuracy and are open to some errors such as truncation or cancellation error etc. These errors occur as a result of binary representation and finite precision and may cause inaccurate results. In this work we are going to discuss the accuracy of the statistical software, different tests and methods available for measuring the accuracy and the comparison of different packages.

45.1.1 Accuracy of Software

Accuracy can be defined as the correctness of the results. When a statistical software package is used, it is assumed that the results are correct in order to comment on these results. On the other hand it must be accepted that computers have some limitations. The main problem is that the available precision provided by computer systems is limited. It is clear that statistical software can not deliver such accurate results, which exceed these limitations. However statistical software should recognize its limits and give clear indication that these limits are reached. We have two types of precision generally used today:

- Single precision
- Double precision

Binary Representation and Finite Precision

As we discussed above under the problem of software accuracy lay the binary representation and finite precision. In computer we don't have real numbers. But we represent them with a finite approximation.

Example: Assume that we want to represent 0.1 in single precision. The result will be as follows:

0.1 = .00011001100110011001100110 = 0.99999964 (McCullough,1998)

It is clear that we can only approximate to 0.1 in binary form. This problem grows, if we try to subtract two large numbers which differs only in the decimals. For instance 100000.1-100000 = .09375

With single precision we can only represent 24 significant binary digits, with other word 6-7 decimal digits. In double precision it is possible to represent 53 significant binary digits and

15-17 significant decimal digits. Limitations of binary representation create five distinct numerical ranges, which cause the loss of accuracy:

- negative overflow
- negative underflow
- zero
- positive underflow
- positive overflow

Overflow means that values have grown too large for the representation. Underflow means that values are so small and so close to zero that causes to set to zero. Single and double precision representations have different ranges.

Results of Binary Representation

This limitations cause different errors in different situations:

- Cancellation error results from subtracting two nearly equal numbers.
- Accumulation errors are successive rounding errors in a series of calculations summed up to a total error. In this type of errors it is possible that only the rightmost digits of the result is affected or the result has no single accurate digits.
- Another result of binary representation and finite precision is that two formulas which are algebraically equivalent may not be equivalent numerically. For instance:

$$\sum_{n=1}^{10000} n^{-2}$$

$$\sum_{n=1}^{10000} (10001-n)^{-2}$$

First formula adds the numbers in ascending order, whereas the second in descending order. In the first formula the smallest numbers reached at the very end of the computation, so that these numbers are all lost to rounding error. The error is 650 times greater than the second.(McCullough,1998)

- Truncation error can be defined as approximation error which results from the limitations of binary representation.

Example:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

Difference between the true value of sin(x) and the result achieved by summing up finite number of terms is truncation error. (McCullough,1998)

- Algorithmic errors are another reason of inaccuracies. There can be different ways of calculating a quantity and these different methods may be unequally accurate. For example according to Sawitzki (1994) in a single precision environment using the following formula in order to calculate variance :

$$S^2 = (1/(1-n)(\sum x_i^2 - n\bar{x}^2))$$

45.1.2 Measuring Accuracy

Due to limits of the computers some problems occur in calculating statistical values. We need a measure which shows us the degree of accuracy of a computed value. This measurement base on the difference between the computed value (q) and the real value (c).An oft-used measure is LRE (number of the correct significant digits)(McCullough,1998)

$$LRE = -\log_{10}[|q-c|/|c|]$$

Rules:

- q should be close to c (less than 2). If they are not, set LRE to zero
- If LRE is greater than number of the digits in c, set LRE to number of the digits in c.
- If LRE is less than unity, set it to zero.

45.2 Testing Statistical Software

In this part we are going to discuss two different tests which aim for measuring the accuracy of the software: Wilkinson Test (Wilkinson, 1985) and NIST StRD Benchmarks.

45.2.1 Wilkinson's Statistic Quiz

Wilkinson dataset “NASTY” which is employed in Wilkinson's Statistic Quiz is a dataset created by Leland Wilkinson (1985). This dataset consist of different variables such as “Zero” which contains only zeros, “Miss” with all missing values, etc. NASTY is a reasonable dataset in the sense of values it contains. For instance the values of “Big” in “NASTY” are less than U.S. Population or “Tiny” is comparable to many values in engineering. On the other hand the exercises of the “Statistic Quiz” are not meant to be reasonable. These tests are designed to check some specific problems in statistical computing. Wilkinson's Statistics Quiz is an entry level test.

45.2.2 NIST StRD Benchmarks

These benchmarks consist of different datasets designed by National Institute of Standards and Technology in different levels of difficulty. The purpose is to test the accuracy of statistical software regarding to different topics in statistics and different level of difficulty. In the webpage of “Statistical Reference Datasets” Project there are five groups of datasets:

- Analysis of Variance
- Linear Regression
- Markov Chain Monte Carlo
- Nonlinear Regression
- Univariate Summary Statistics

In all groups of benchmarks there are three different types of datasets: Lower level difficulty datasets, average level difficulty datasets and higher level difficulty datasets. By using these datasets we are going to explore whether the statistical software deliver accurate results to 15 digits for some statistical computations.

There are 11 datasets provided by NIST among which there are six datasets with lower level difficulty, two datasets with average level difficulty and one with higher level difficulty. Certified values to 15 digits for each dataset are provided for the mean (μ), the standard deviation (σ), the first-order autocorrelation coefficient (ρ).

In group of ANOVA-datasets there are 11 datasets with levels of difficulty, four lower, four average and three higher. For each dataset certified values to 15 digits are provided for between treatment degrees of freedom, within treatment. degrees of freedom, sums of squares, mean squares, the F-statistic , the R^2, the residual standard deviation. Since most of the certified values are used in calculating the F-statistic, only its LRE λ_F will be compared to the result of regarding statistical software.

For testing the linear regression results of statistical software NIST provides 11 datasets with levels of difficulty two lower, two average and seven higher. For each dataset we have the certified values to 15 digits for coefficient estimates, standard errors of coefficients, the residual standard deviation, R^2, the analysis of variance for linear regression table, which includes the residual sum of squares. LREs for the least accurate coefficients λ_β, standard errors λ_σ and Residual sum of squares λ_r will be compared. In nonliner regression dataset group there are 27 datasets designed by NIST with difficulty eight lower ,eleven average and eight higher. For each dataset we have certified values to 11 digits provided by NIST for coefficient estimates, standard errors of coefficients, the residual sum of squares, the residual standard deviation, the degrees of freedom.

In the case of calculation of nonlinear regression we apply curve fitting method. In this method we need starting values in order to initialize each variable in the equation. Then we generate the curve and calculate the convergence criterion (ex. sum of squares). Then we adjust the variables to make the curve closer to the data points. There are several algorithms for adjusting the variables:

- The method of Marquardt and Levenberg
- The method of linear descent
- The method of Gauss-Newton

One of these methods is applied repeatedly, until the difference in the convergence criterion is smaller than the convergence tolerance.

NIST provides also two sets of starting values: Start I (values far from solution), Start II (values close to solution). Having Start II as initial values makes it easier to reach an accurate solution. Therefore Start I solutions will be preffered.

Other important settings are as follows:

- the convergence tolerance (ex. 1E-6)
- the method of solution (ex. Gauss Newton or Levenberg Marquardt)
- the convergence criterion (ex. residual sum of squares (RSS) or square of the maximum of the parameter differences)

We can also choose between numerical and analytic derivatives.

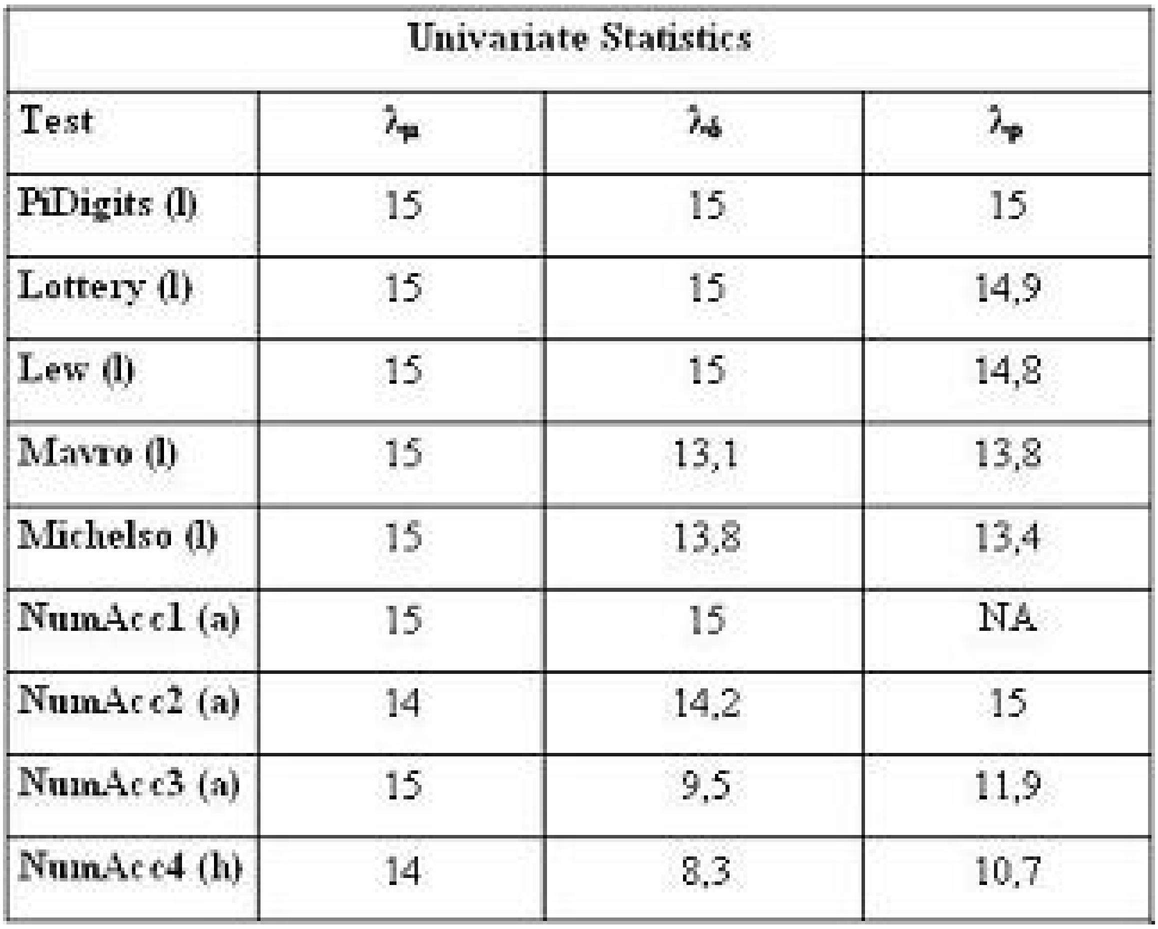
45.3 Testing Examples

45.3.1 Testing Software Package: SAS, SPSS and S-Plus

In this part we are going to discuss the test results of three statistical software packages applied by M.D. McCullough. In McCullough's work SAS 6.12, SPSS 7.5 and S-Plus 4.0 are tested and compared in respect to certified LRE values provided by NIST. Comparison will be handled according to the following parts:

- Univariate Statistics
- ANOVA
- Linear Regression
- Nonlinear Regression

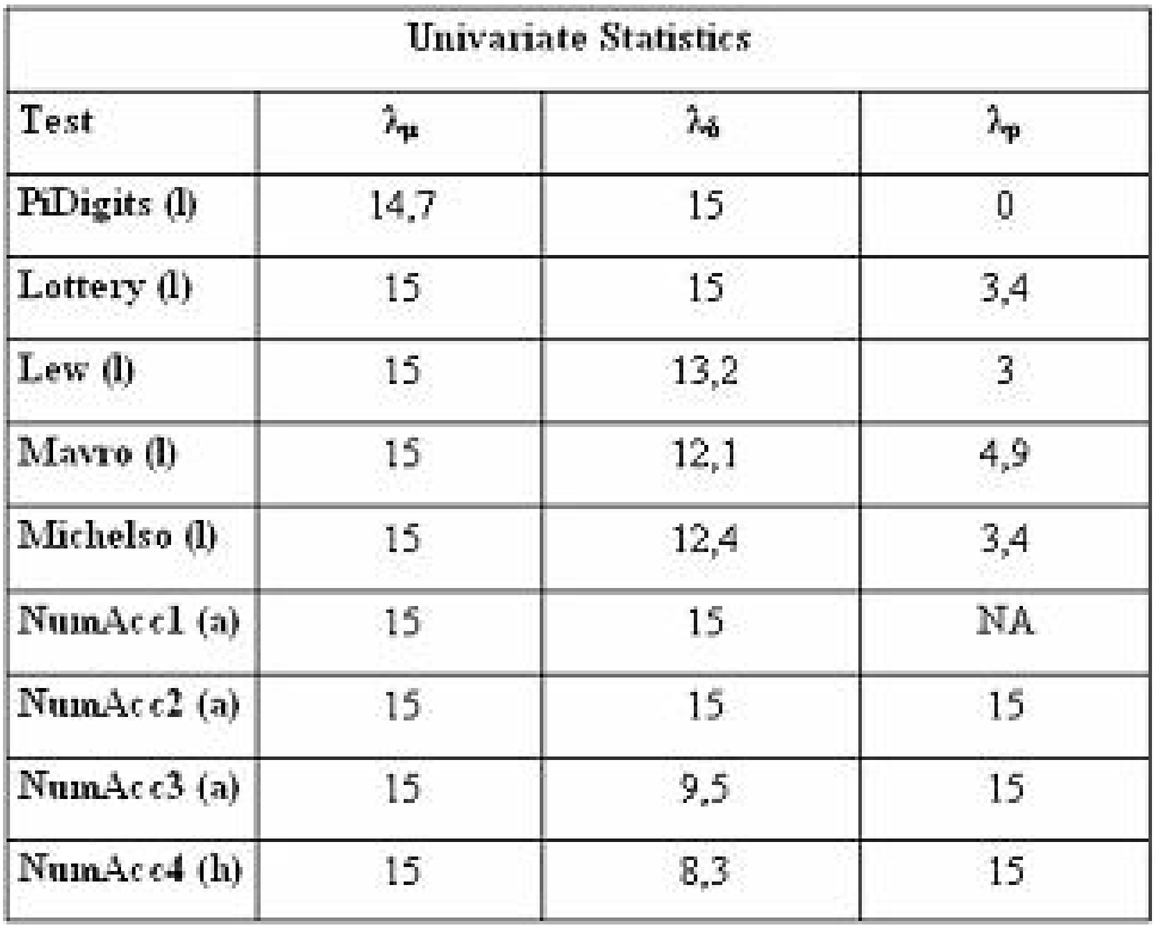
Univariate Statistics



Univariate Statistics			
Test	λ_{μ}	λ_{σ}	λ_{ρ}
PiDigits (l)	15	15	15
Lottery (l)	15	15	14,9
Lew (l)	15	15	14,8
Mavro (l)	15	13,1	13,8
Michelso (l)	15	13,8	13,4
NumAcc1 (a)	15	15	NA
NumAcc2 (a)	14	14,2	15
NumAcc3 (a)	15	9,5	11,9
NumAcc4 (h)	14	8,3	10,7

Figure 25: Table 1: Results from SAS for Univariate Statistics (McCullough,1998)

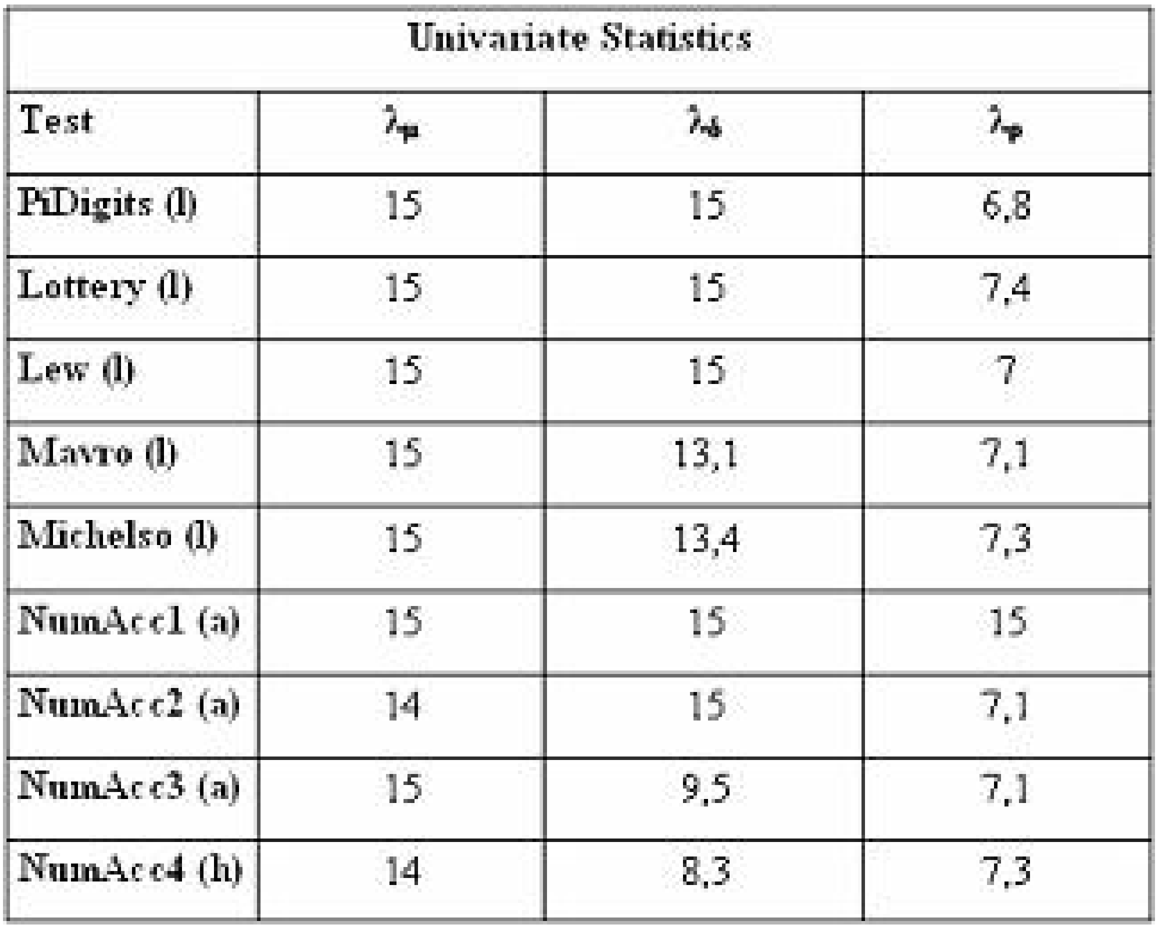
All values calculated in SAS seem to be more or less accurate. For the dataset NumAcc1 p-value can not be calculated because of the insufficient number of observations. Calculating standard deviation for datasets NumAcc3 (average difficulty) and NumAcc 4 (high difficulty) seem to stress SAS.



Univariate Statistics			
Test	λ_μ	λ_σ	λ_p
PiDigits (l)	14,7	15	0
Lottery (l)	15	15	3,4
Lew (l)	15	13,2	3
Mavro (l)	15	12,1	4,9
Michelso (l)	15	12,4	3,4
NumAcc1 (a)	15	15	NA
NumAcc2 (a)	15	15	15
NumAcc3 (a)	15	9,5	15
NumAcc4 (h)	15	8,3	15

Figure 26: Table 2: Results from SPSS for Univariate Statistics (McCullough,1998)

All values calculated for mean and standard deviation seem to be more or less accurate. For the dataset NumAcc1 p-value can not be calculated because of the insufficient number of observations.Calculating standard deviation for datasets NumAcc3 and -4 seem to stress SPSS,as well. For p-values SPSS represent results with only 3 decimal digits which causes an understate of first and an overstate of last p-values regarding to accuracy.

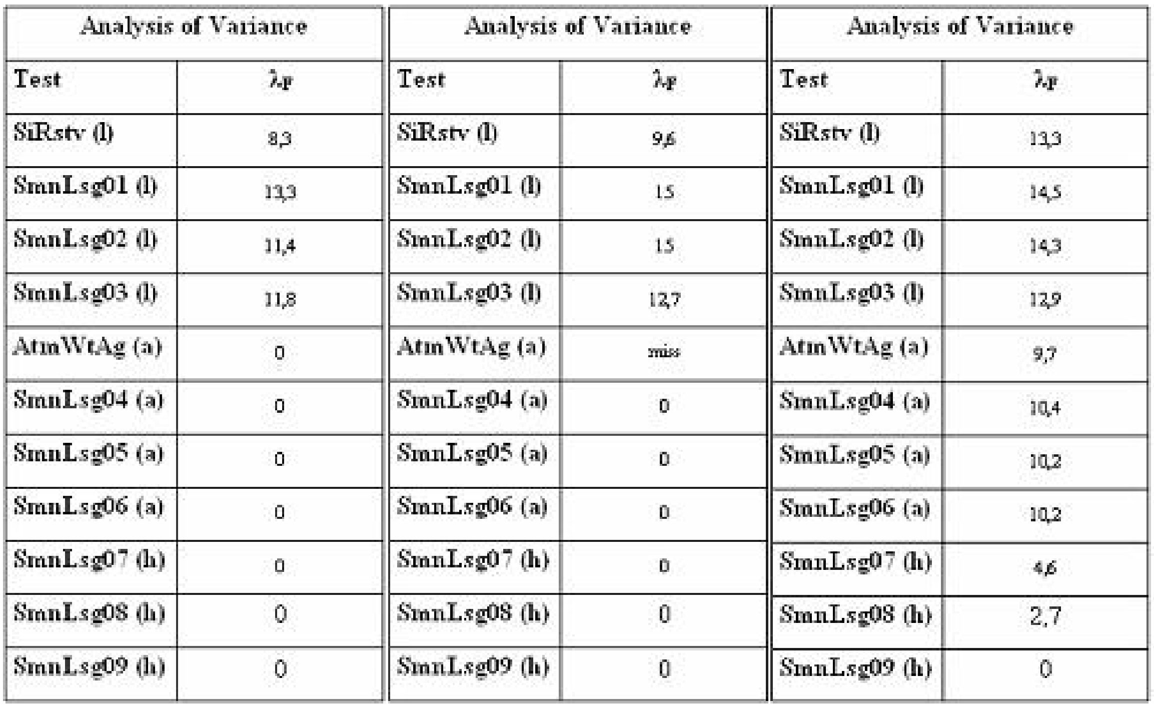


Univariate Statistics			
Test	λ_μ	λ_σ	λ_ρ
PiDigits (l)	15	15	6,8
Lottery (l)	15	15	7,4
Lew (l)	15	15	7
Mavro (l)	15	13,1	7,1
Michelso (l)	15	13,4	7,3
NumAcc1 (a)	15	15	15
NumAcc2 (a)	14	15	7,1
NumAcc3 (a)	15	9,5	7,1
NumAcc4 (h)	14	8,3	7,3

Figure 27: Table 3: Results from S-Plus for Univariate Statistics (McCullough,1998)

All values calculated for mean and standard deviation seem to be more or less accurate. S-Plus have also problems in calculating standard deviation for datasets NumAcc3 and -4. S-Plus does not show a good performance in calculating the p-values.

Analysis of Variance



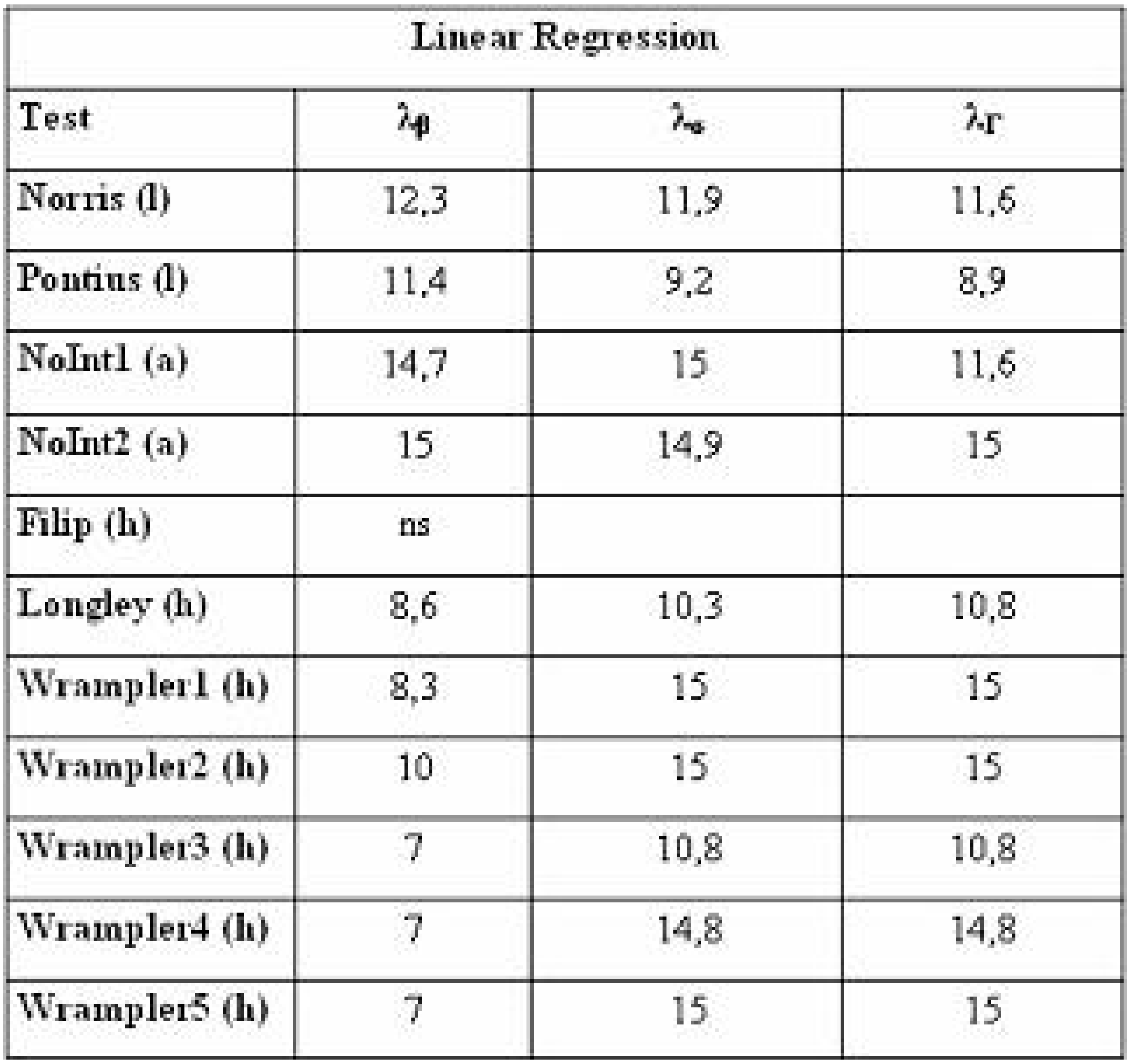
Analysis of Variance		Analysis of Variance		Analysis of Variance	
Test	λ_F	Test	λ_F	Test	λ_F
SiRstv (l)	8,3	SiRstv (l)	9,6	SiRstv (l)	13,3
SmnLsg01 (l)	13,3	SmnLsg01 (l)	1,5	SmnLsg01 (l)	14,5
SmnLsg02 (l)	11,4	SmnLsg02 (l)	1,5	SmnLsg02 (l)	14,3
SmnLsg03 (l)	11,8	SmnLsg03 (l)	12,7	SmnLsg03 (l)	12,9
AtmWtAg (a)	0	AtmWtAg (a)	miss	AtmWtAg (a)	9,7
SmnLsg04 (a)	0	SmnLsg04 (a)	0	SmnLsg04 (a)	10,4
SmnLsg05 (a)	0	SmnLsg05 (a)	0	SmnLsg05 (a)	10,2
SmnLsg06 (a)	0	SmnLsg06 (a)	0	SmnLsg06 (a)	10,2
SmnLsg07 (h)	0	SmnLsg07 (h)	0	SmnLsg07 (h)	4,6
SmnLsg08 (h)	0	SmnLsg08 (h)	0	SmnLsg08 (h)	2,7
SmnLsg09 (h)	0	SmnLsg09 (h)	0	SmnLsg09 (h)	0

Figure 28: Table 4: Results from SAS for Analysis of Variance(McCullough,1998)

Results:

- SAS can solve only the ANOVA problems of lower level difficulty.
- F-Statistics for datasets of average or higher difficulty can be calculated with very poor performance and zero digit accuracy.
- SPSS can display accurate results for datasets with lower level difficulty, like SAS.
- Performance of SPSS in calculating ANOVA is poor.
- For dataset “AtmWtAg” SPSS displays no F-Statistic which seems more logical instead of displaying zero accurate results.
- S-Plus handels ANOVA problem better than other softwares.
- Even for higher difficulty datasets this package can display more accurate results than other. But still results for datasets with high difficulty are not enough accurate.
- S-Plus can solve the average difficulty problems with a sufficient accuracy.

Linear Regression

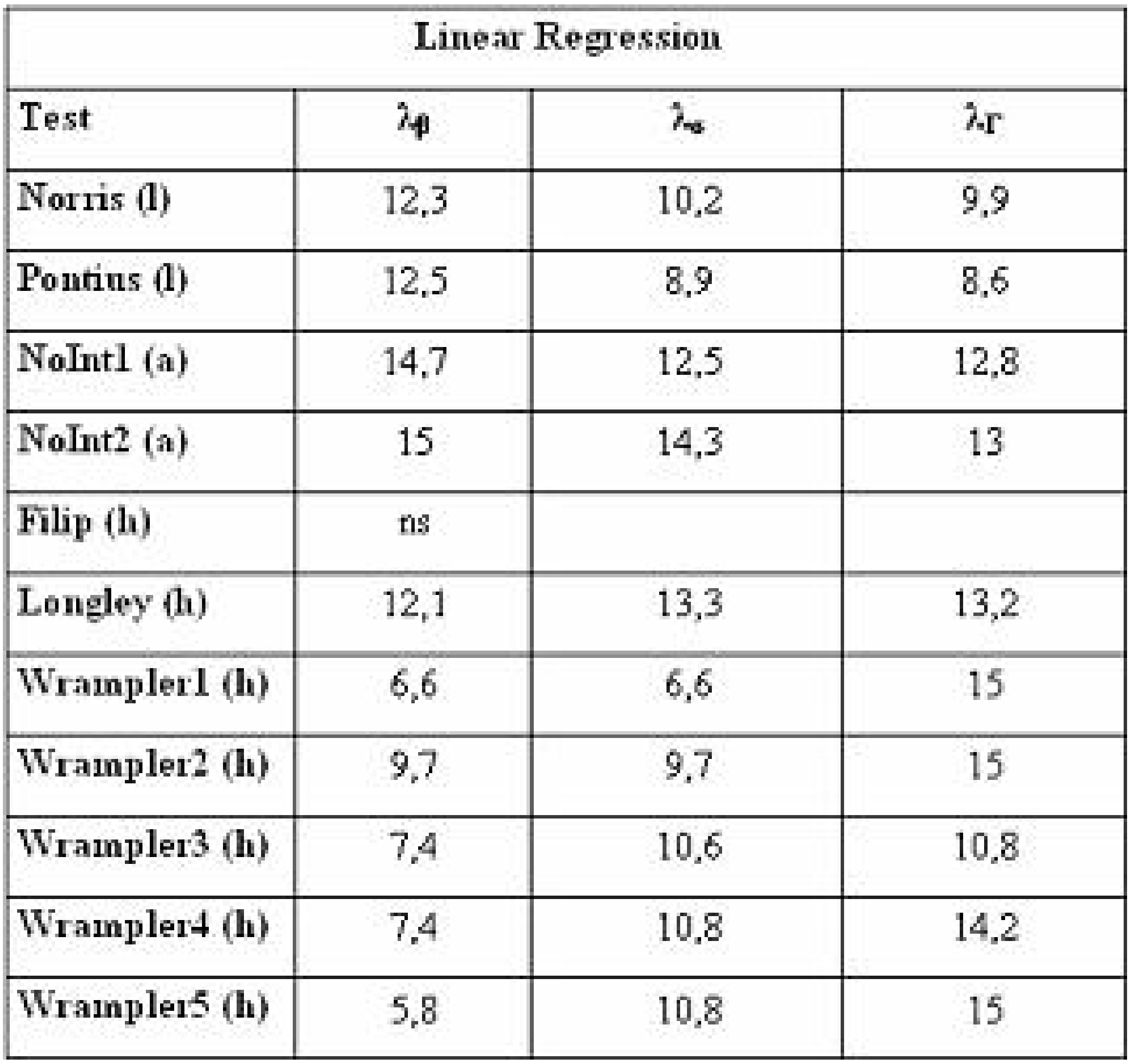


Linear Regression

Test	λ_β	λ_σ	λ_r
Norris (l)	12,3	11,9	11,6
Pontius (l)	11,4	9,2	8,9
NoInt1 (a)	14,7	15	11,6
NoInt2 (a)	15	14,9	15
Filip (h)	ns		
Longley (h)	8,6	10,3	10,8
Wrampler1 (h)	8,3	15	15
Wrampler2 (h)	10	15	15
Wrampler3 (h)	7	10,8	10,8
Wrampler4 (h)	7	14,8	14,8
Wrampler5 (h)	7	15	15

Figure 29: Table 5: Results from SAS for Linear Regression(McCullough,1998)

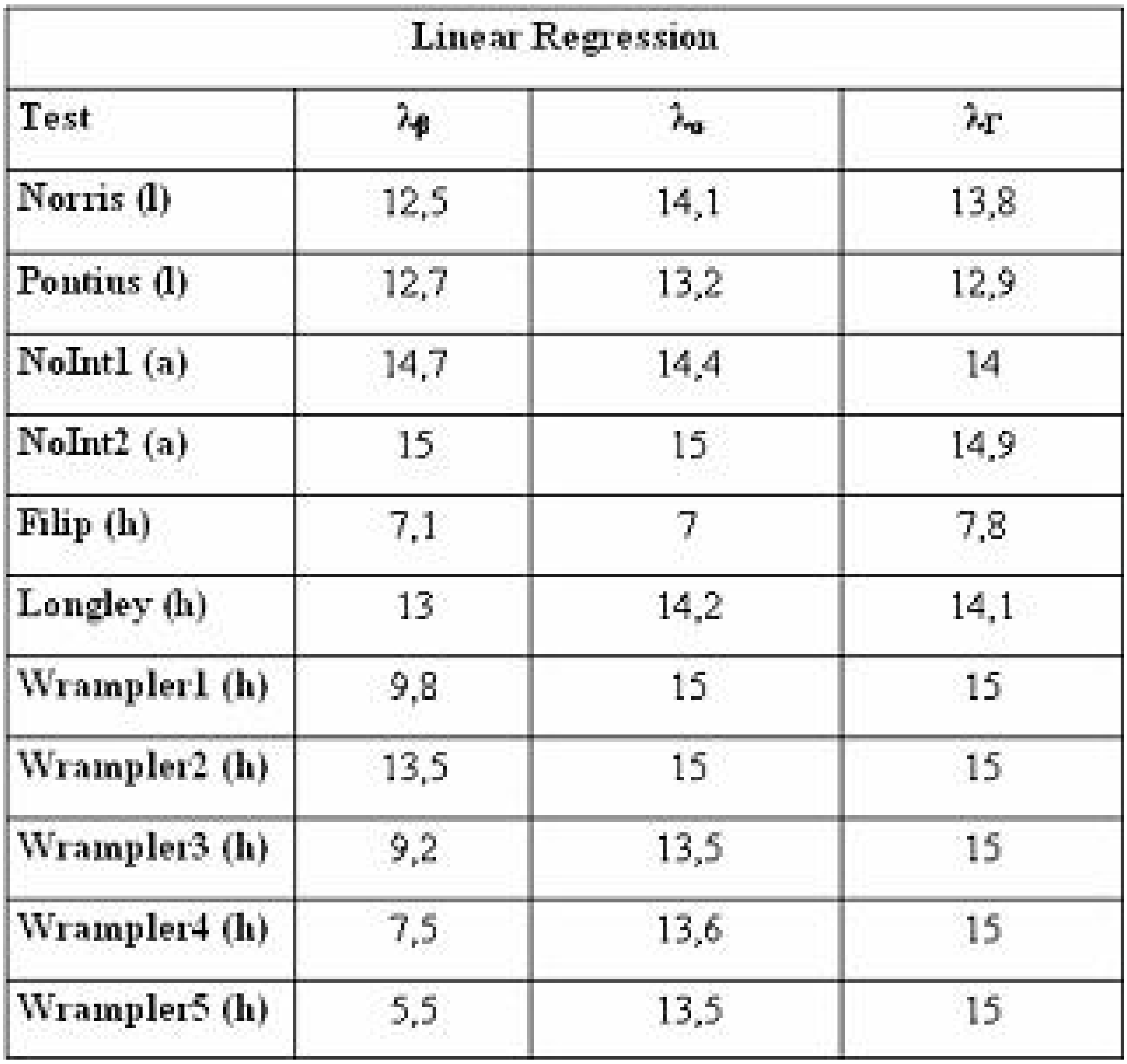
SAS delivers no solution for dataset Filip which is ten degree polynomial. Except Filip SAS can display more or less accurate results. But the performance seems to decrease for higher difficulty datasets, especially in calculating coefficients



Linear Regression			
Test	λ_β	λ_σ	λ_r
Norris (l)	12,3	10,2	9,9
Pontius (l)	12,5	8,9	8,6
NoInt1 (a)	14,7	12,5	12,8
NoInt2 (a)	15	14,3	13
Filip (h)	ns		
Longley (h)	12,1	13,3	13,2
Wrampler1 (h)	6,6	6,6	15
Wrampler2 (h)	9,7	9,7	15
Wrampler3 (h)	7,4	10,6	10,8
Wrampler4 (h)	7,4	10,8	14,2
Wrampler5 (h)	5,8	10,8	15

Figure 30: Table 6: Results from SPSS for Linear Regression(McCullough,1998)

SPSS has also Problems with “Filip” which is a 10 degree polynomial. Many packages fail to compute values for it. Like SAS, SPSS delivers lower accuracy for high level datasets

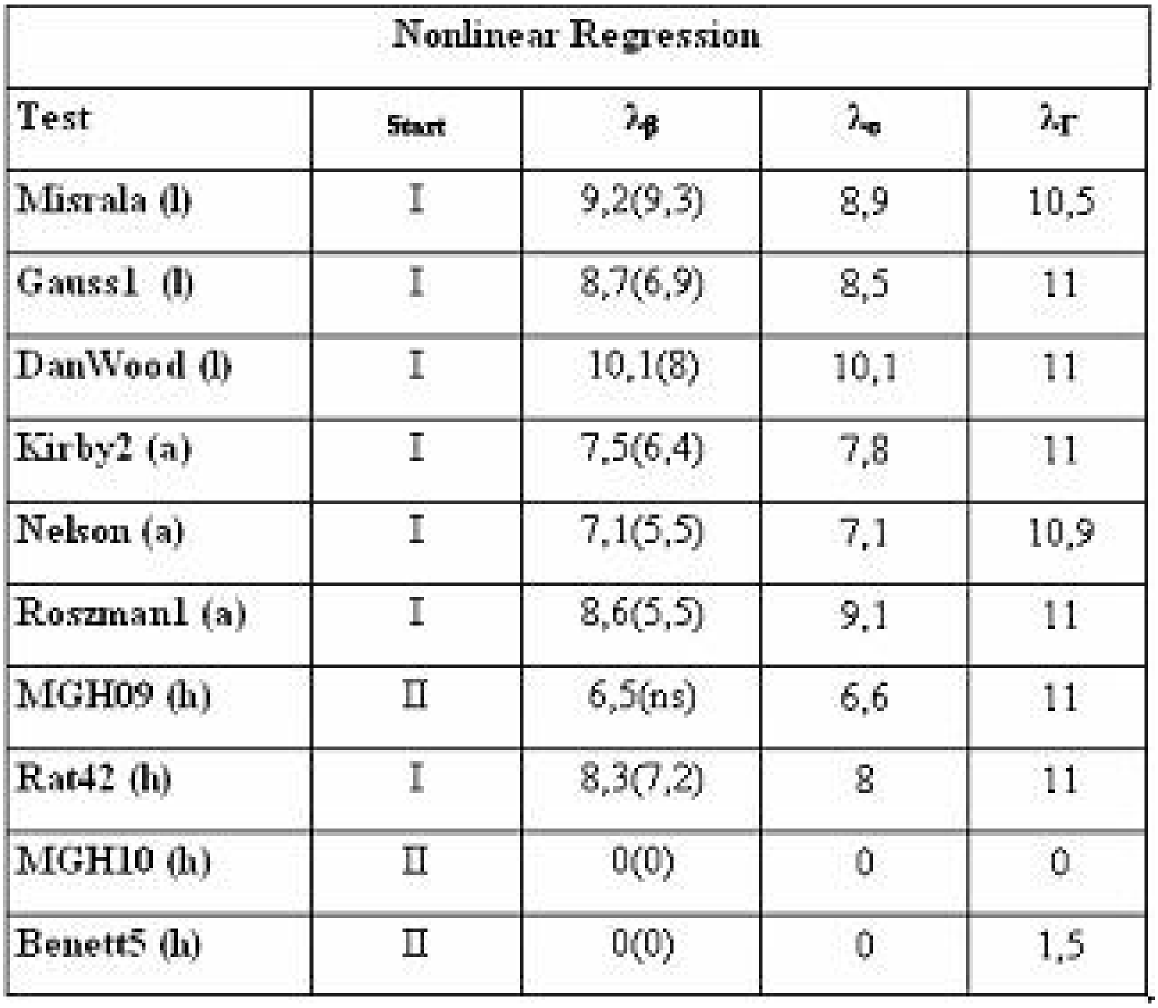


Linear Regression			
Test	λ_β	λ_σ	λ_r
Norris (l)	12,5	14,1	13,8
Pontius (l)	12,7	13,2	12,9
NoInt1 (a)	14,7	14,4	14
NoInt2 (a)	15	15	14,9
Filip (h)	7,1	7	7,8
Longley (h)	13	14,2	14,1
Wrampler1 (h)	9,8	15	15
Wrampler2 (h)	13,5	15	15
Wrampler3 (h)	9,2	13,5	15
Wrampler4 (h)	7,5	13,6	15
Wrampler5 (h)	5,5	13,5	15

Figure 31: Table 7: Results from S-Plus for Linear Regression(McCullough,1998)

S-Plus is the only package which delivers a result for dataset “Filip”. The accuracy of Result for Filip seem not to be poor but average. Even for higher difficulty datasets S-Plus can calculate more accurate results than other software packages. Only coefficients for datasets “Wrampler4” and “-5” is under the average accuracy.

Nonlinear Regression



Nonlinear Regression

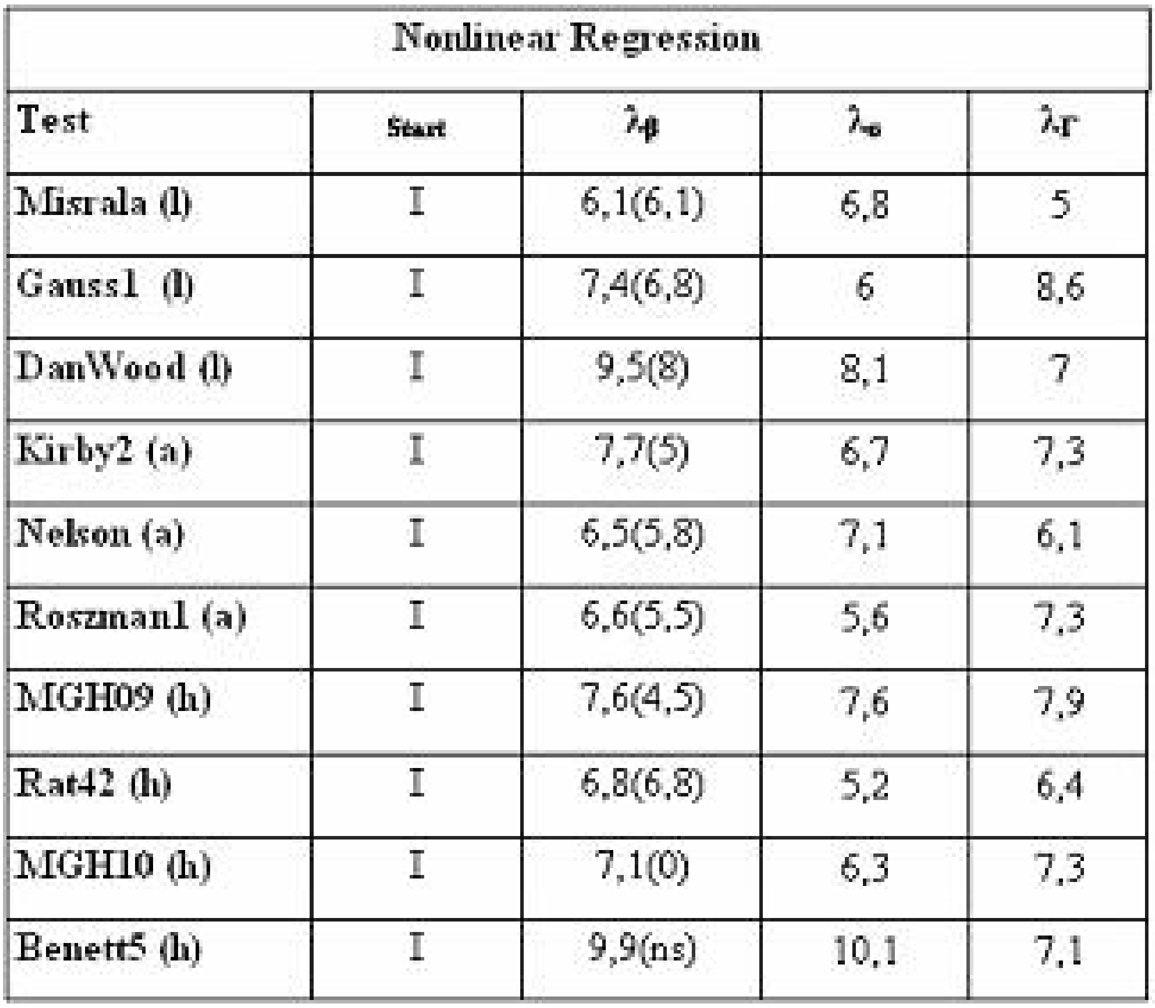
Test	Start	λ_β	λ_σ	λ_r
Misrala (l)	I	9,2(9,3)	8,9	10,5
Gauss1 (l)	I	8,7(6,9)	8,5	11
DanWood (l)	I	10,1(8)	10,1	11
Kirby2 (a)	I	7,5(6,4)	7,8	11
Nelson (a)	I	7,1(5,5)	7,1	10,9
Roszman1 (a)	I	8,6(5,5)	9,1	11
MGH09 (h)	II	6,5(ns)	6,6	11
Rat42 (h)	I	8,3(7,2)	8	11
MGH10 (h)	II	0(0)	0	0
Benett5 (h)	II	0(0)	0	1,5

Figure 32: Table 8: Results from SAS for Nonlinear Regression(McCullough,1998)

For the nonlinear Regression two setting combinations are tested for each software, because different settings make a difference in the results.As we can see in the table in SAS preffered combination produce better results than default combination. In this table results produced using default combination are in paranthesis. Because 11 digits are provided for certified values by NIST, we are looking for LRE values of 11.

Preffered combination :

- Method:Gauss-Newton
- Criterion: PARAM
- Tolerance: 1E-6



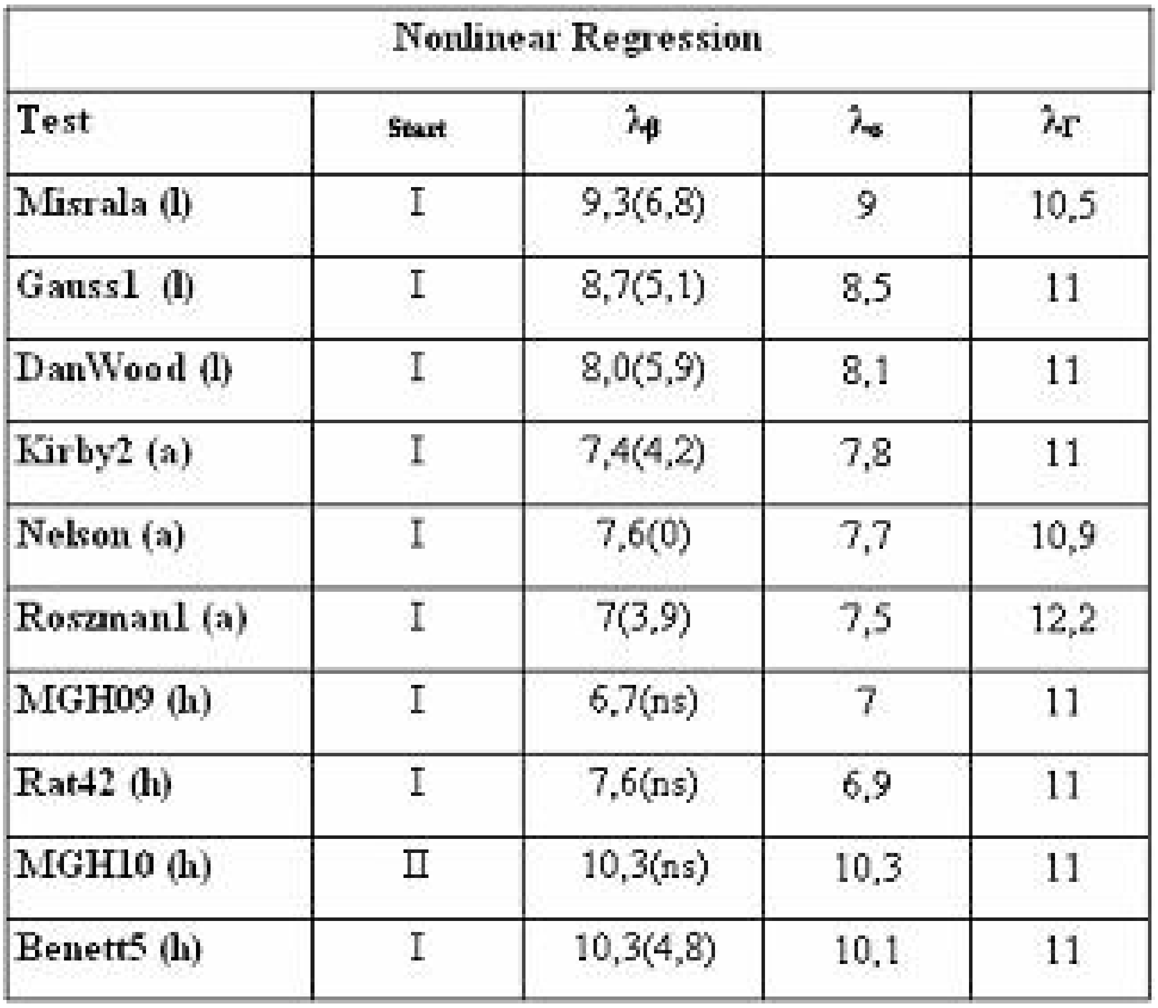
Nonlinear Regression				
Test	Start	λ_β	λ_σ	λ_r
Misrala (l)	I	6,1(6,1)	6,8	5
Gauss1 (l)	I	7,4(6,8)	6	8,6
DanWood (l)	I	9,5(8)	8,1	7
Kirby2 (a)	I	7,7(5)	6,7	7,3
Nelson (a)	I	6,5(5,8)	7,1	6,1
Roszman1 (a)	I	6,6(5,5)	5,6	7,3
MGH09 (h)	I	7,6(4,5)	7,6	7,9
Rat42 (h)	I	6,8(6,8)	5,2	6,4
MGH10 (h)	I	7,1(0)	6,3	7,3
Benett5 (h)	I	9,9(ns)	10,1	7,1

Figure 33: Table 9: Results from SPSS for Nonlinear Regression(McCullough,1998)

Also in SPSS preffered combination shows a better performance than default options. All problems are solved with initial values “start I” whereas in SAS higher level datasets are solved with Start II values.

Preffered Combination:

- Method:Levenberg-Marquardt
- Criterion:PARAM
- Tolerance: 1E-12



Nonlinear Regression				
Test	Start	λ_β	λ_σ	λ_r
Misrala (l)	I	9,3(6,8)	9	10,5
Gauss1 (l)	I	8,7(5,1)	8,5	11
DanWood (l)	I	8,0(5,9)	8,1	11
Kirby2 (a)	I	7,4(4,2)	7,8	11
Nelson (a)	I	7,6(0)	7,7	10,9
Roszman1 (a)	I	7(3,9)	7,5	12,2
MGH09 (h)	I	6,7(ns)	7	11
Rat42 (h)	I	7,6(ns)	6,9	11
MGH10 (h)	II	10,3(ns)	10,3	11
Benett5 (h)	I	10,3(4,8)	10,1	11

Figure 34: Table 10: Results from S-Plus for Nonlinear Regression(McCullough,1998)

As we can see in the table preffered combination is also in S-Plus better than default combination. All problems except “MGH10” are solved with initial values “start I”. We may say that S-Plus showed a better performance than other software in calculating nonlinear regression.

Preffered Combination:

- Method:Gauss-Newton
- Criterion:RSS
- Tolerance: 1E-6

Results of the Comparison

All packages delivered accurate results for mean and standard deviation in univariate statistics.There are no big differences between the tested statistical software packages. In ANOVA calculations SAS and SPSS can not pass the average difficulty problems, whereas S-Plus delivered more accurate results than others. But for high difficulty datasets it also produced poor results. Regarding linear regression problems all packages seem to be reliable. If we examine the results for all software packages, we can say that the success in calculating the results for nonlinear regression greatly depends on the chosen options.

Other important results are as follows:

- S-Plus solved from Start II one time.
- SPSS never used Start II as initial values, but produce one time zero accurate digits.
- SAS used Start II three times and produced three times zero accurate digits.

45.3.2 Comparison of different versions of SPSS

In this part we are going to compare an old version with a new version of SPSS in order to see whether the problems in older version are solved in the new one. In this part we compared SPSS version 7.5 with SPSS version 12.0. LRE values for version 7.5 are taken from an article by B.D. McCullough (see references). We also applied these tests to version 12.0 and calculated regarding LRE values. We chose one dataset from each difficulty groups and applied univariate statistics, ANOVA and linear regression in version 12.0. Source for the datasets is NIST Statistical Reference Datasets Archive. Then we computed LRE values for each dataset by using the certified values provided by NIST in order to compare two versions of SPSS.

Univariate Statistics

Difficulty: Low

Our first dataset is PiDigits with lower level difficulty which is designed by NIST in order to detect the deficiencies in calculating univariate statistical values.

Certified Values for PiDigits are as follows:

- Sample Mean : 4.53480000000000
- Sample Standard Deviation : 2.86733906028871

As we can see in the table 13 the results from SPSS 12.0 match the certified values provided by NIST. Therefore our LREs for mean and standard deviation are λ_μ: 15, λ_δ: 15. In version 7.5 LRE values were λ_μ: 14.7, λ_δ: 15. (McCullough,1998)

Difficulty: Average

Second dataset is NumAcc3 with average difficulty from NIST datasets for univariate statistics. Certified Values for NumAcc3 are as follows:

- Sample Mean : 1000000.2
- Sample Standard Deviation : 0.1

In the table 14 we can see that calculated mean value is the same with the certified value by NIST. Therefore our LREs for mean is λ_μ: 15. However the standard deviation value differs from the certified value. So the calculation of LRE for standard deviation is as follows:

λ_δ : -log10 |0,10000000003464-0,1|/|0,1| = 9.5

LREs for SPSS v 7.5 were λ_μ: 15, λ_δ: 9.5. (McCullough,1998)

Difficulty: High

Last dataset in univariate statistics is NumAcc4 with high level of difficulty. Certified Values for NumAcc4 are as follows:

- Sample Mean : 10000000.2
- Sample Standard Deviation : 0.1

Also for this dataset we do not have any problems with computed mean value. Therefore LRE is λ_μ: 15. However the standard deviation value does not match to the certified one. So we should calculate the LRE for standard deviation as follows:

λ_δ : -log10 |0,10000000056078-0,1|/|0,1| = 8.3

LREs for SPSS v 7.5 were λ_μ: 15, λ_δ : 8.3 (McCullough,1998)

For this part of our test we can say that there is no difference between two versions of SPSS. For average and high difficulty datasets delivered standard deviation results have still an average accuracy.

Analysis of Variance

Difficulty: Low

The dataset which we used for testing SPSS 12.0 regarding lower difficulty level problems is SiRstv. Certified F Statistic for SiRstv is 1.18046237440255E+00

- LRE : λ_F: -log10 | 1,18046237440224- 1,18046237440255|/ |1,18046237440255| = 12,58
- LRE for SPSS v 7.5 : λ_F : 9,6 (McCullough, 1998)

Difficulty: Average

Our dataset for average difficulty problems is AtmWtAg . Certified F statistic value for AtmWtAg is 1.59467335677930E+01.

- LREs : λ_F : -log10 | 15,9467336134506- 15,9467335677930|/| 15,9467335677930| = 8,5
- LREs for SPSS v 7.5 : λ_F: miss

Difficulty: High

We used the dataset SmnLsg07 in order to test high level difficulty problems. Certified F value for SmnLsg07 is 2.10000000000000E+01

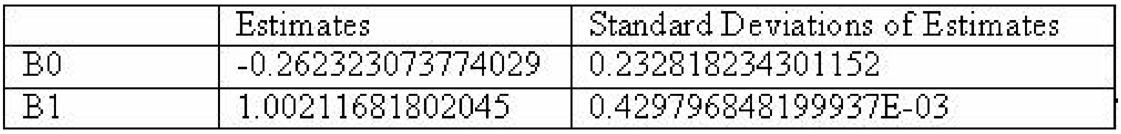
- LREs : λ_F : -log10 | 21,0381922055595 - 21|/| 21| = 2,7
- LREs for SPSS v 7.5 : λ_F: 0

ANOVA results computed in version 12.0 are better than those calculated in version 7.5. However the accuracy degrees are still too low.

Linear Regression

Difficulty: Low

Our lower level difficulty dataset is Norris for linear regression. Certified values for Norris are as follows:

- Sample Residual Sum of Squares : 26.6173985294224
- 

	Estimates	Standard Deviations of Estimates
B0	-0.262323073774029	0.232818234301152
B1	1.00211681802045	0.429796848199937E-03

Figure 35: Table 17: Coefficient estimates for Norris(www.itl.nist.gov)

- LREs : λ_r : 9,9 λ_β : 12,3 λ_σ : 10,2
- LREs for SPSS v 7.5 : λ_r: 9,9 , λ_β : 12,3 , λ_σ : 10,2 (McCullough, 1998)

Difficulty: Average

We used the dataset NoInt1 in order to test the performance in average difficulty dataset. Regression model is as follows:

y = B1*x + e

Certified Values for NoInt1 :

- Sample Residual Sum of Squares : 127,272727272727
- Coefficient estimate : 2.07438016528926, standard deviation : 0.16528925619834E-0(www.itl.nist.gov)
- LREs: λ_r:12,8 λ_β: 15 λ_σ: 12,9
- LREs for SPSS v. 7.5 : λ_r: 12,8 , λ_β: 14,7 , λ_σ: 12,5 (McCullough, 1998)

Difficulty: High

Our high level difficulty dataset is Longley designed by NIST.

- Model: y =B0+B1*x1 + B2*x2 + B3*x3 + B4*x4 + B5*x5 + B6*x6 +e
- LREs :
  - λ_r: -log10 |836424,055505842-836424,055505915|/ |836424,055505915| = 13,1
  - λ_β : 15
  - λ_σ : -log10 | 0,16528925619836E-01 – 0,16528925619834E-01|/ |0,16528925619834E-01| = 12,9
- LREs for SPSS v. 7.5 : λ_r: 12,8 , λ_β : 14,7 , λ_σ : 12,5 (McCullough, 1998)

As we conclude from the computed LREs, there is no big difference between the results of two versions for linear regression.

45.4 Conclusion

By applying these test we try to find out whether the software are reliable and deliver accurate results or not. However based on the results we can say that different software packages deliver different results for same the problem which can lead us to wrong interpretations for statistical research questions.

In specific we can that SAS, SPSS and S-Plus can solve the linear regression problems better in comparision to ANOVA Problems. All three of them deliver poor results for F statistic calculation.

From the results of comparison two different versions of SPSS we can conclude that the difference between the accuracy of the results delivered by SPSS v.12 and v.7.5 is not great considering the difference between the version numbers. On the other hand SPSS v.12 can handle the ANOVA Problems much better than old version. However it has still problems in higher difficulty problems.

45.5 References

- McCullough, B.D. 1998, 'Assessing The Reliability of Ststistical Software: Part I',*The American Statistician*, Vol.52, No.4, pp.358-366.
- McCullough, B.D. 1999, 'Assessing The Reliability of Ststistical Software: Part II', *The American Statistician*, Vol.53, No.2, pp.149-159
- Sawitzki, G. 1994, 'Testing Numerical Reliability of Data Analysis Systems', *Computational Statistics & Data Analysis*, Vol.18, No.2, pp.269-286
- Wilkinson, L. 1993, 'Practical Guidelines for Testing Statistical Software' in *25th Conference on Statistical Computing at Schloss Reisenburg*, ed. P. Dirschedl& R. Ostermnann, Physica Verlag
- National Institute of Standards and Technology. (1 September 2000). The Statistical Reference Datasets: Archives, [Online], Available from: <HTTP://WWW.ITL.NIST.GOV/DIV898/STRD/GENERAL/DATAARCHIVE.HTML[1]> [10 November 2005].

[1] HTTP://WWW.ITL.NIST.GOV/DIV898/STRD/GENERAL/DATAARCHIVE.HTML

46 Numerics in Excel

The purpose of this paper is to evaluate the accuracy of MS Excel in terms of statistical procedures and to conclude whether the MS Excel should be used for (statistical) scientific purposes or not. The evaulation is made for Excel versions 97, 2000, XP and 2003.

According to the literature, there are three main problematic areas for Excel if it is used for statistical calculations. These are

- probability distributions,
- univariate statistics, ANOVA and Estimations (both linear and non-linear)
- random number generation.

If the results of statistical packages are assessed, one should take into account that the acceptable accuracy of the results should be achieved in double precision (which means that a result is accepted as accurate if it possesses 15 accurate digits) given that the reliable algorithms are capable of delivering correct results in double precision, as well. If the reliable algorithms can not retrieve results in double precision, it is not fair to anticipate that the package (evaluated) should achieve double precision. Thus we can say that the correct way for evaluating the statistical packages is assessing the quality of underlying algorithm of statistical calculations rather than only counting the accurate digits of results. Besides, test problems must be reasonable which means they must be amenable to solution by known reliable algorithms. (McCullough & Wilson, 1999, S. 28)

In further sections, our judgement about the accuracy of MS Excel will base on certified values and tests. As basis we have Knüsel's ELV software for probability distributions, StRD (Statistical Reference Datasets) for Univariate Statistics, ANOVA and Estimations and finally Marsaglia's DIEHARD for Random Number Generation. Each of the tests and certified values will be explained in the corresponding sections.

46.1 Assessing Excel Results for Statistical Distributions

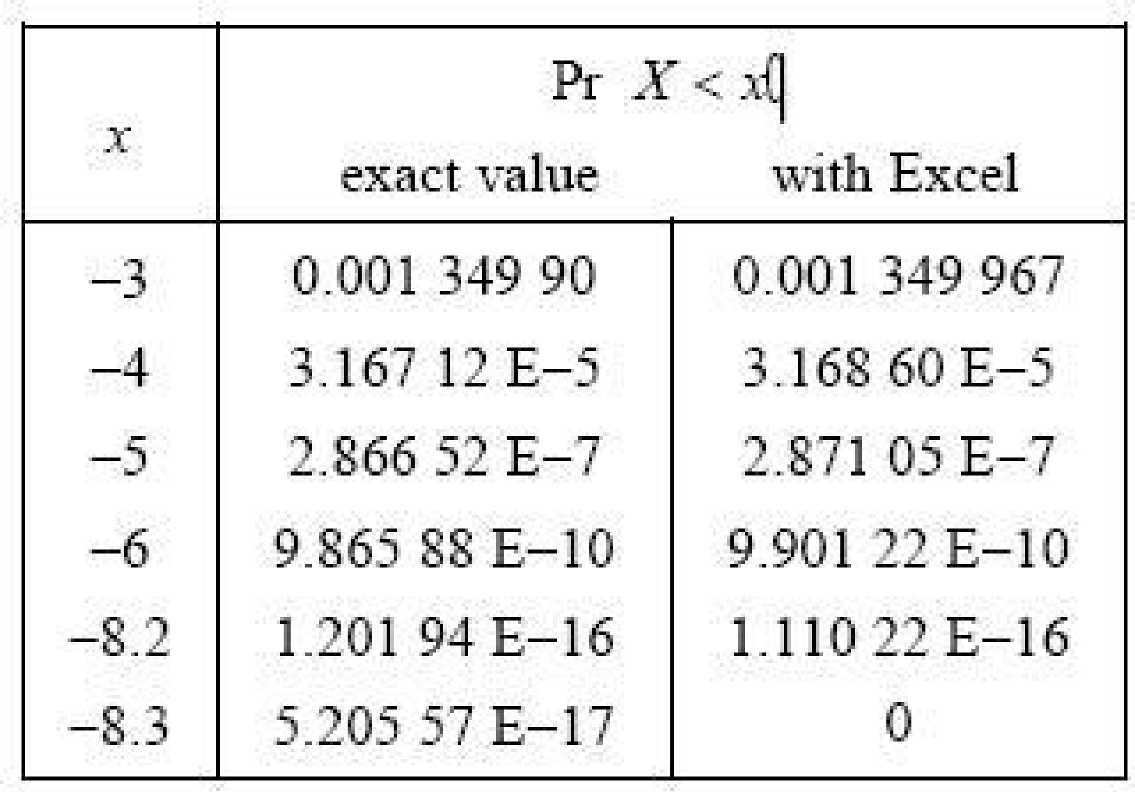
As we mentioned above our judgement about Excel's calculations for probability distributions will base on Knüsel's ELV Program which can compute probabilities and quantiles of some elementary statistical distributions. Using ELV, the upper and lower tail probabilities of all distributions are computed with six significant digits for probabilities as small as 10–100 and upper and lower quantiles are computed for all distributions for tail probabilities P with 10–12 ≤ P ≤ $\frac{1}{2}$. (Knüsel, 2003, S.1)

In our benchmark Excel should display no inaccurate digits. If six digits are displayed, then all six digits should be correct. If the algorithm is only accurate to two digits, then only two digits should be displayed so as not to mislead the user (McCullough & Wilson, 2005, S. 1245)

In the following sub-sections the exact values in the tables are retrieved from Knüsel's ELV software and the acceptable accuracy is in single presicion, because even the best algorithms can not achieve 15 correct digits in most cases, if the probability distributions are issued.

46.1.1 Normal Distribution

- *Excel Function:*NORMDIST
- *Parameters:* mean = 0, variance = 1, x (critical value)
- *Computes:* the tail probability Pr X ≤ x, whereas X denotes a random variable with a standard normal distribution (with mean 0 and variance 1)



x	Pr $X < x$ exact value	with Excel
−3	0.001 349 90	0.001 349 967
−4	3.167 12 E−5	3.168 60 E−5
−5	2.866 52 E−7	2.871 05 E−7
−6	9.865 88 E−10	9.901 22 E−10
−8.2	1.201 94 E−16	1.110 22 E−16
−8.3	5.205 57 E−17	0

Figure 36: Table 1: (Knüsel, 1998, S.376)

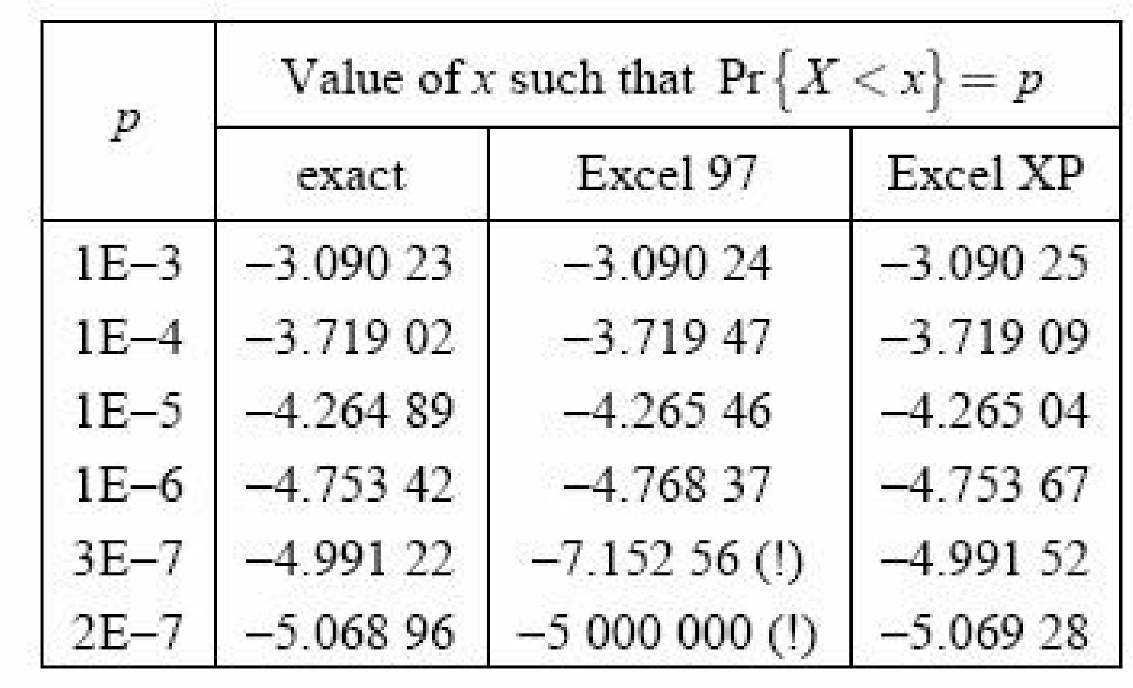
As we can see in table 1, Excel 97, 2000 and XP encounter problems and computes small probabilities in tail incorrectly (i.e for x = -8,3 or x = -8.2) However, this problem is fixed in Excel 2003 (Knüsel, 2005, S.446).

46.1.2 Inverse Normal Distribution

- *Excel Function:* NORMINV
- *Parameters:* mean = 0, variance = 1, p (probability for X < x)
- *Computes:* the x value (quantile)

X denotes a random variable with a standard normal distribution. In contrast to “NORMDIST” function issued in the last section, p is given and quantile is computed.

If used, Excel 97 prints out quantiles with 10 digits although none of these 10 digits may be correct if p is small. In Excel 2000 and XP, Microsoft tried to fix errors, although results are not sufficient (See table 2). However in Excel 2003 the problem is fixed entirely. (Knüsel, 2005, S.446)



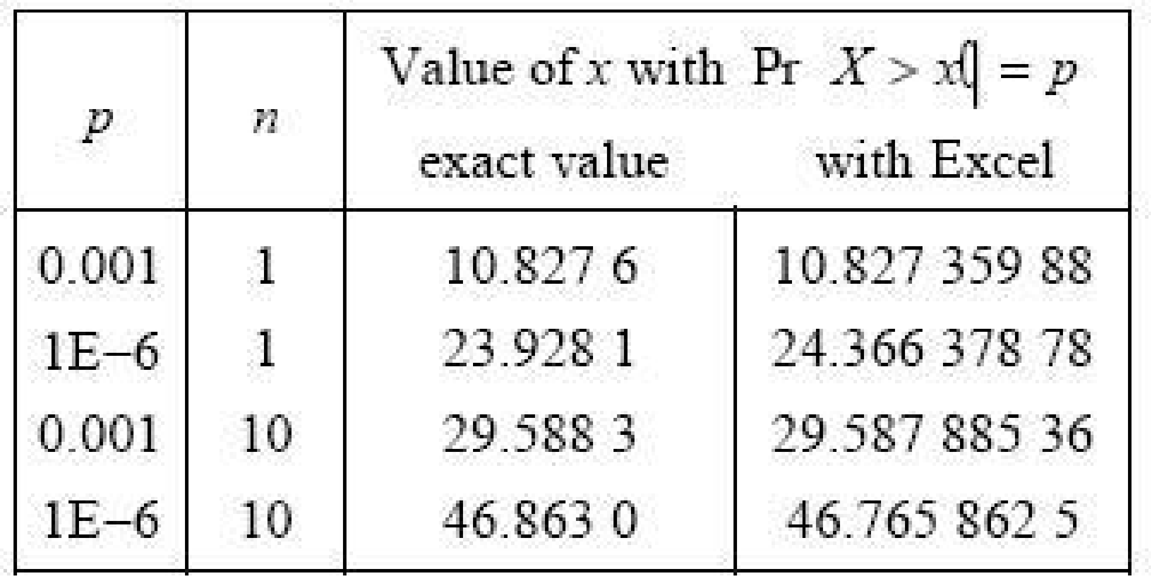
p	Value of x such that $\Pr\{X < x\} = p$		
	exact	Excel 97	Excel XP
1E–3	–3.090 23	–3.090 24	–3.090 25
1E–4	–3.719 02	–3.719 47	–3.719 09
1E–5	–4.264 89	–4.265 46	–4.265 04
1E–6	–4.753 42	–4.768 37	–4.753 67
3E–7	–4.991 22	–7.152 56 (!)	–4.991 52
2E–7	–5.068 96	–5 000 000 (!)	–5.069 28

Figure 37: Table 2: (Knüsel, 2002, S.110)

46.1.3 Inverse Chi-Square Distribution

- *Excel Function:* CHIINV
- *Parameters:* p (probability for X > x), n (degrees of freedom)
- *Computes:* the x value (quantile)

X denotes a random variable with a chi-square distribution with n degrees of freedom.



p	n	Value of x with $\Pr\ X > x] = p$	
		exact value	with Excel
0.001	1	10.827 6	10.827 359 88
1E–6	1	23.928 1	24.366 378 78
0.001	10	29.588 3	29.587 885 36
1E–6	10	46.863 0	46.765 862 5

Figure 38: Table 3: (Knüsel , 1998, S. 376)

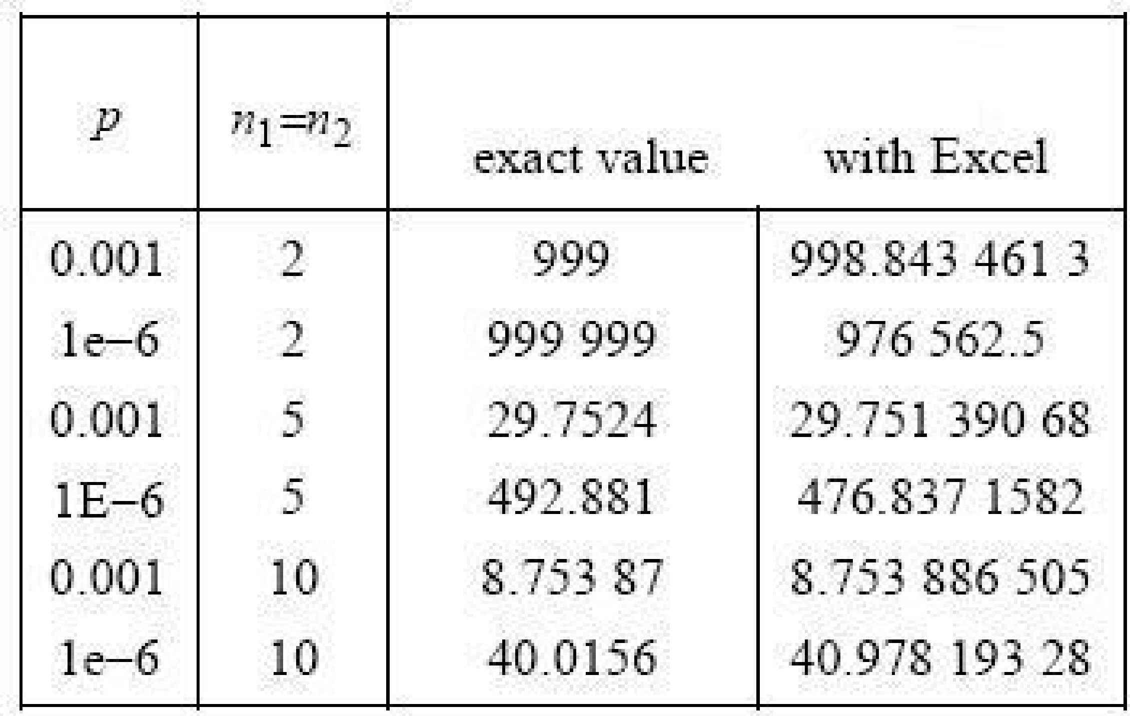
Old Excel Versions: Although the old Excel versions show ten significant digits, only very few of them are accurate if p is small (See table 3). Even if p is not small, the accurate digits are not enough to say that Excel is sufficient for this distribution.

Excel 2003: Problem was fixed. (Knüsel, 2005, S.446)

46.1.4 Inverse F Distribution

- *Excel Function:* FINV
- *Parameters:* p (probability for X > x), n1, n2 (degrees of freedom)
- *Computes:* the x value (quantile)

X denotes a random variable with a F distribution with n1 and n2 degrees of freedom.



p	$n_1 = n_2$	exact value	with Excel
0.001	2	999	998.843 461 3
1e–6	2	999 999	976 562.5
0.001	5	29.7524	29.751 390 68
1E–6	5	492.881	476.837 1582
0.001	10	8.753 87	8.753 886 505
1e–6	10	40.0156	40.978 193 28

Figure 39: Table 4: (Knüsel , 1998, S. 377)

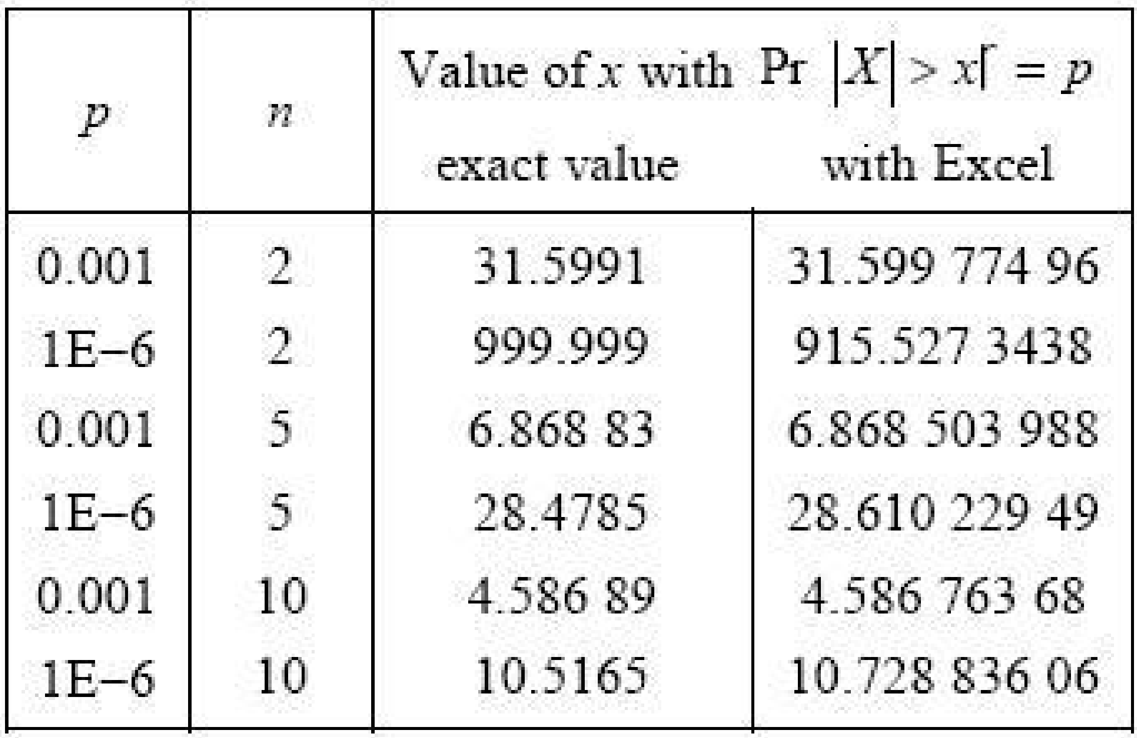
Old Excel Versions: Excel prints out x values with 7 or more significant digits although only one or two of these many digits are correct if p is small (See table 4).

Excel 2003: Problem fixed. (Knüsel, 2005, S.446)

46.1.5 Inverse t Distribution

- *Excel Function:* TINV
- *Parameters:* p (probability for |X| > x), n (degree of freedom)
- *Computes:* the x value (quantile)

X denotes a random variable with a t distribution with n degrees of freedom. Please note that the |X| value causes a 2 tailed computation. (lower tail & high tail)



p	n	Value of x with Pr $\|X\| > x\lceil = p$	
		exact value	with Excel
0.001	2	31.5991	31.599 774 96
1E−6	2	999.999	915.527 3438
0.001	5	6.868 83	6.868 503 988
1E−6	5	28.4785	28.610 229 49
0.001	10	4.586 89	4.586 763 68
1E−6	10	10.5165	10.728 836 06

Figure 40: Table 5: (Knüsel , 1998, S. 377)

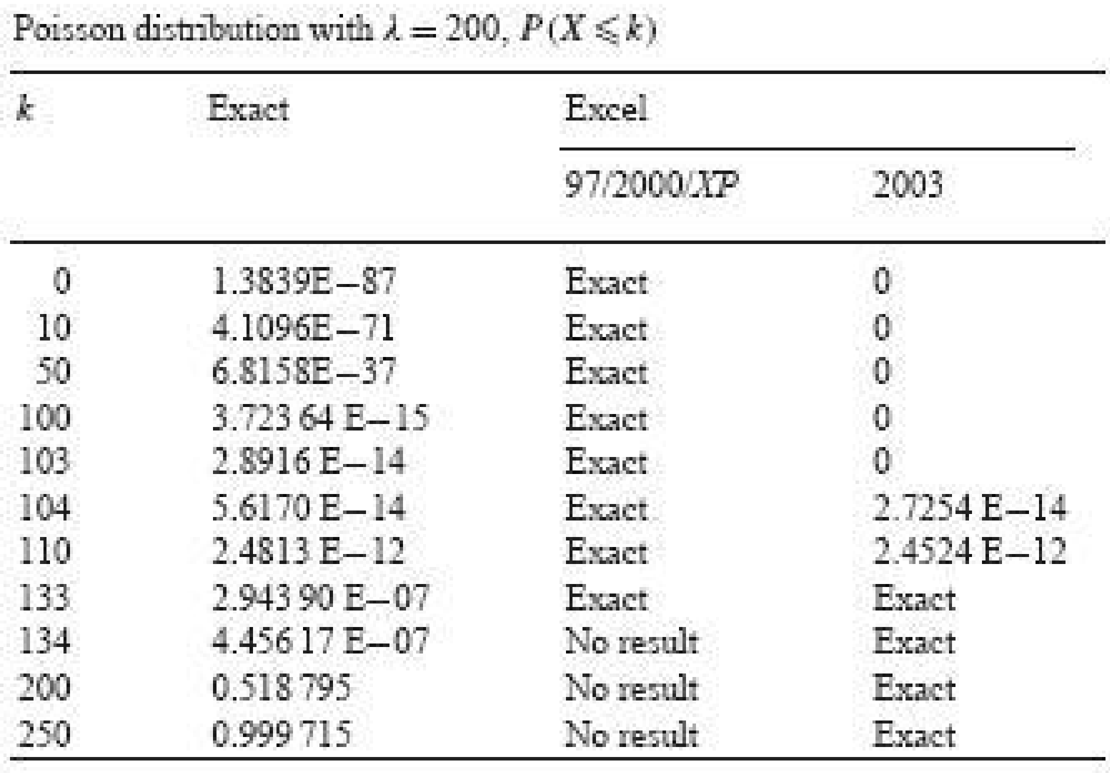
Old Excel Versions: Excel prints out quantiles with 9 or more significant digits although only one or two of these many digits are correct if p is small (See table 5).

Excel 2003: Problem fixed. (Knüsel, 2005, S.446)

46.1.6 Poisson Distribution

- *Excel Function:* Poisson
- *Parameters:* λ (mean), k (number of cases)
- *Computes:* the tail probability Pr X ≤ k

X denotes a random variable with a Poisson distribution with given parameters.



Poisson distribution with $\lambda = 200$, $P(X \leqslant k)$

k	Exact	Excel	
		97/2000/XP	2003
0	1.3839E−87	Exact	0
10	4.1096E−71	Exact	0
50	6.8158E−37	Exact	0
100	3.723 64 E−15	Exact	0
103	2.8916 E−14	Exact	0
104	5.6170 E−14	Exact	2.7254 E−14
110	2.4813 E−12	Exact	2.4524 E−12
133	2.943 90 E−07	Exact	Exact
134	4.456 17 E−07	No result	Exact
200	0.518 795	No result	Exact
250	0.999 715	No result	Exact

Figure 41: Table 6: (McCullough & Wilson, 2005, S.1246)

Old Excel Versions: correctly computes very small probabilities but gives no result for central probabilities near the mean (in the range about 0.5). (See table 6)

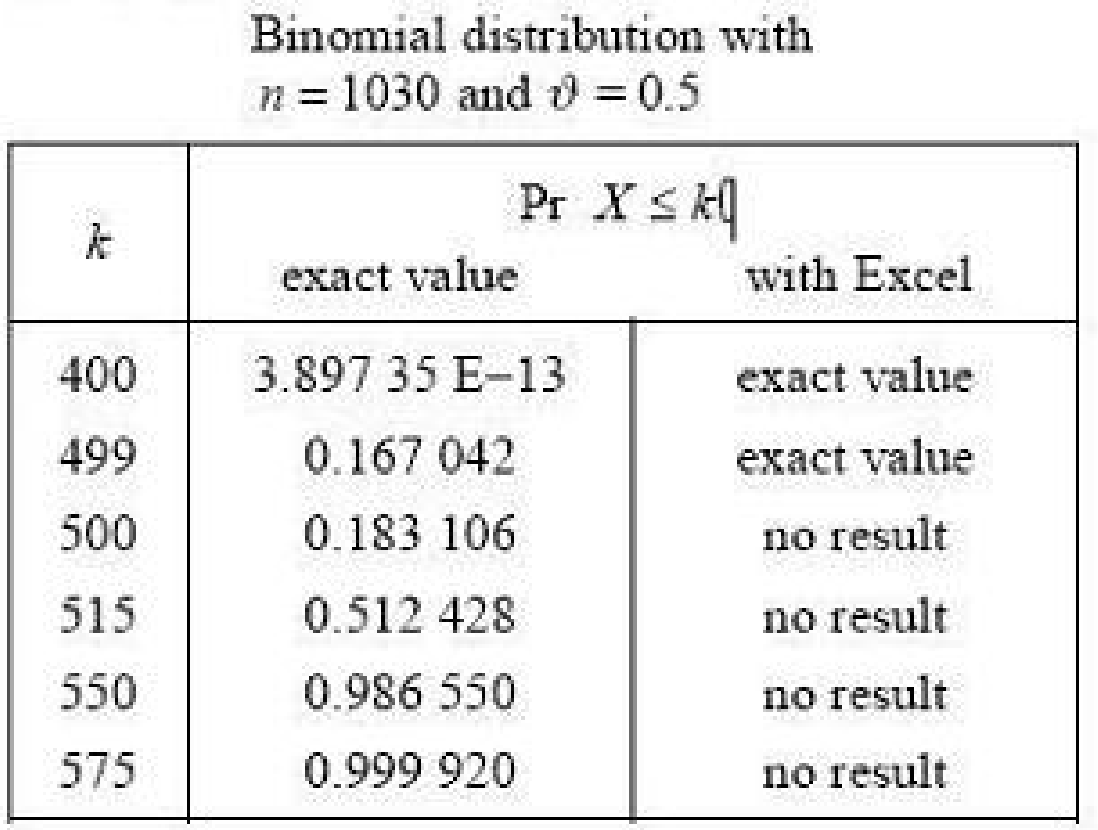
Excel 2003: The central probabilities are fixed. However, inaccurate results in the tail. (See table 6)

The strange behaivour of Excel can be encountered for values λ▯▯▯150. (Knüsel, 1998, S.375) It fails even for probabilities in the central range between 0.01 and 0.99 and even for parameter values that cannot be judged as too extreme.

46.1.7 Binomial Distribution

- *Excel Function:* BINOMDIST
- *Parameters:* n (= number of trials) , υ(= probability for a success) , k(number of successes)
- *Computes:* the tail probability Pr X ≤ k

-X denotes a random variable with a binoamial distribution with given parameters



Binomial distribution with
$n = 1030$ and $\vartheta = 0.5$

k	Pr $X \leq k$	
	exact value	with Excel
400	3.897 35 E−13	exact value
499	0.167 042	exact value
500	0.183 106	no result
515	0.512 428	no result
550	0.986 550	no result
575	0.999 920	no result

Figure 42: Table 7: (Knüsel, 1998, S.375)

Old Excel Versions: As we see in table 7, old versions of Excel correctly computes very small probabilities but gives no result for central probabilities near the mean (same problem with Poisson distribuiton on old Excel versions)

Excel 2003: The central probabilities are fixed. However, inaccurate results in the tail. (Knüsel, 2005, S.446). (same problem with Poisson distribuiton on Excel 2003).

This strange behaivour of Excel can be encountered for values n > 1000. (Knüsel, 1998, S.375) It fails even for probabilities in the central range between 0.01 and 0.99 and even for parameter values that cannot be judged as too extreme.

46.1.8 Other problems

- Excel 97, 2000 and XP includes flaws by computing the hypergeometric distribution (HYPERGEOM). For some values (N > 1030) no result is retrieved. This is prevented on Excel 2003, but there is still no option to compute tail probabilities. So computation of Pr {X = k} is possible, but computation of Pr {X ≤ k} is not. (Knüsel, 2005, S.447)
- Function GAMMADIST for gamma distribution retreives incorrect values on Excel 2003. (Knüsel, 2005, S.447-448)
- Also the function BETAINV for inverse beta distribution computes incorrect values on Excel 2003 (Knüsel, 2005, S. 448)

46.2 Assessing Excel Results for Univariate Statistics, ANOVA and Estimation (Linear & Non-Linear)

Our judgement about Excel's calculations for univariate statistics, ANOVA and Estimation will base on StRD which is designed by Statistical Engineering Division of National Institute of Standards and Technology (NIST) to assist researchers in benchmarking statistical software packages explicitly. StRD has reference datasets (real-world and generated datasets) with certified computational results that enable the objective evaluation of statistical Software. It comprises four suites of numerical benchmarks for statistical software: univariate summary statistics, one way analysis of variance, linear regression and nonlinear regression and it includes several problems for each suite of tests. All problems have a difficulty level:low, average or high.

By assessing Excel results in this section we are going to use LRE (log relative error) which can be used as a score for accuracy of results of statistical packages. The number of correct digits in results can be calculated via log relative error. Please note that for double precision the computed LRE is in the range 0 - 15, because we can have max. 15 correct digits in double precision.

Formula LRE:

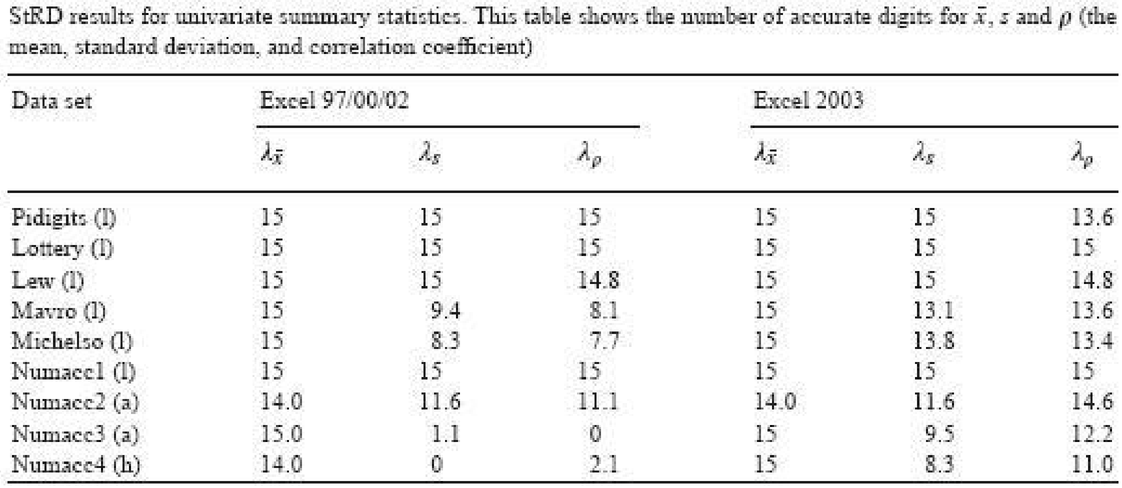
$\lambda = LRE(x) = -log_{10}\left(\frac{|x-c|}{|x|}\right)$

c: the correct answer (certified computational result) for a particular test problem

x: answer of Excel for the same problem

46.2.1 Univariate Statistics

- *Excel Functions:* - AVERAGE, STDEV, PEARSON (also CORREL)
- *Computes (respectively):* mean, standard deviation, correlation coefficient



StRD results for univariate summary statistics. This table shows the number of accurate digits for $\bar{x}$, s and ρ (the mean, standard deviation, and correlation coefficient)

Data set	Excel 97/00/02			Excel 2003		
	$\lambda_{\bar{x}}$	λ_s	λ_ρ	$\lambda_{\bar{x}}$	λ_s	λ_ρ
Pidigits (l)	15	15	15	15	15	13.6
Lottery (l)	15	15	15	15	15	15
Lew (l)	15	15	14.8	15	15	14.8
Mavro (l)	15	9.4	8.1	15	13.1	13.6
Michelso (l)	15	8.3	7.7	15	13.8	13.4
Numacc1 (l)	15	15	15	15	15	15
Numacc2 (a)	14.0	11.6	11.1	14.0	11.6	14.6
Numacc3 (a)	15.0	1.1	0	15	9.5	12.2
Numacc4 (h)	14.0	0	2.1	15	8.3	11.0

Figure 43: Table 8: (McCullough & Wilson, 2005, S.1247)

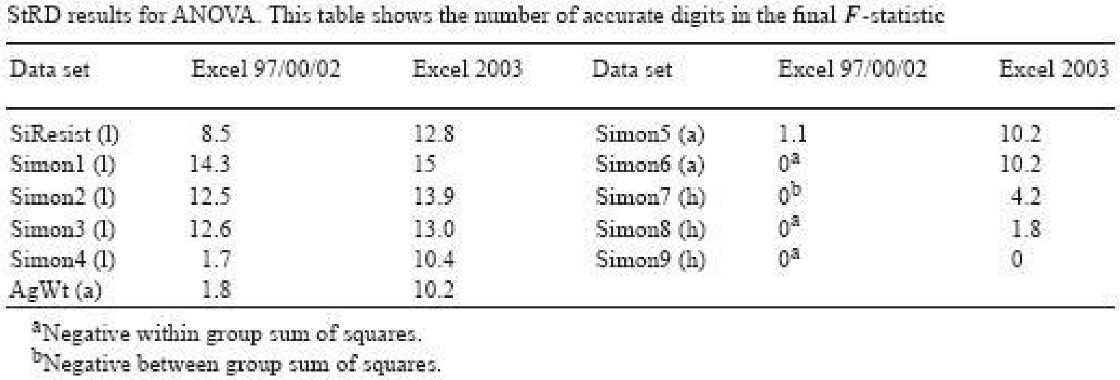
Old Excel Versions: an unstable algorithm for calculation of the sample variance and the correlation coefficient is used. Even for the low difficulty problems (datasets with letter “l” in table 8) the old versions of Excel fail.

Excel 2003: Problem was fixed and the performance is acceptable. The accurate digits less than 15 don’t indicate an unsuccessful implementation because even the reliable algorithms can not retrieve 15 correct digits for these average and high difficulty problems (datasets with letters “a” and “h” in table 8) of StRD.

46.2.2 ONEWAY ANOVA

- *Excel Function:* Tools – Data Analysis – ANOVA: Single Factor (requires Analysis Toolpak)
- *Computes:* df, ss, ms, F-statistic

Since ANOVA produces many numerical results (such as df, ss, ms, F), here only the LRE for the final F-statistic is presented. Before assessing Excel’s performance one should consider that a reliable algorithm for one way Analysis of Variance can deliver 8-10 digits for the average difficulty problems and 4-5 digits for higher difficulty problems.



StRD results for ANOVA. This table shows the number of accurate digits in the final *F*-statistic

Data set	Excel 97/00/02	Excel 2003	Data set	Excel 97/00/02	Excel 2003
SiResist (l)	8.5	12.8	Simon5 (a)	1.1	10.2
Simon1 (l)	14.3	15	Simon6 (a)	0[a]	10.2
Simon2 (l)	12.5	13.9	Simon7 (h)	0[b]	4.2
Simon3 (l)	12.6	13.0	Simon8 (h)	0[a]	1.8
Simon4 (l)	1.7	10.4	Simon9 (h)	0[a]	0
AgWt (a)	1.8	10.2			

[a]Negative within group sum of squares.
[b]Negative between group sum of squares.

Figure 44: Table 9: (McCullough & Wilson, 2005, S.1248)

Old Excel Versions: Considering numerical solutions, delivering only a few digits of accuracy for difficult problems is not an evidence for bad software, but retrieving 0 accurate digits for average difficulty problems indicates bad software when calculating ANOVA. (McCullough & Wilson, 1999, S. 31). For that reason Excel versions prior than Excel 2003 has an acceptable performance only on low-difficulty problems. It retrieves zero accurate digits for difficult problems. Besides, negative results for “within group sum of squares” and “between group sum of squares” are the further indicators of a bad algorithm used for Excel. (See table 9)

Excel 2003: Problem was fixed (See table 9). The zero digits of accuracy for the Simon 9 test is no cause for concern, since this also occurs when reliable algorithms are employed. Therefore the performance is acceptable. (McCullough & Wilson, 2005, S. 1248)

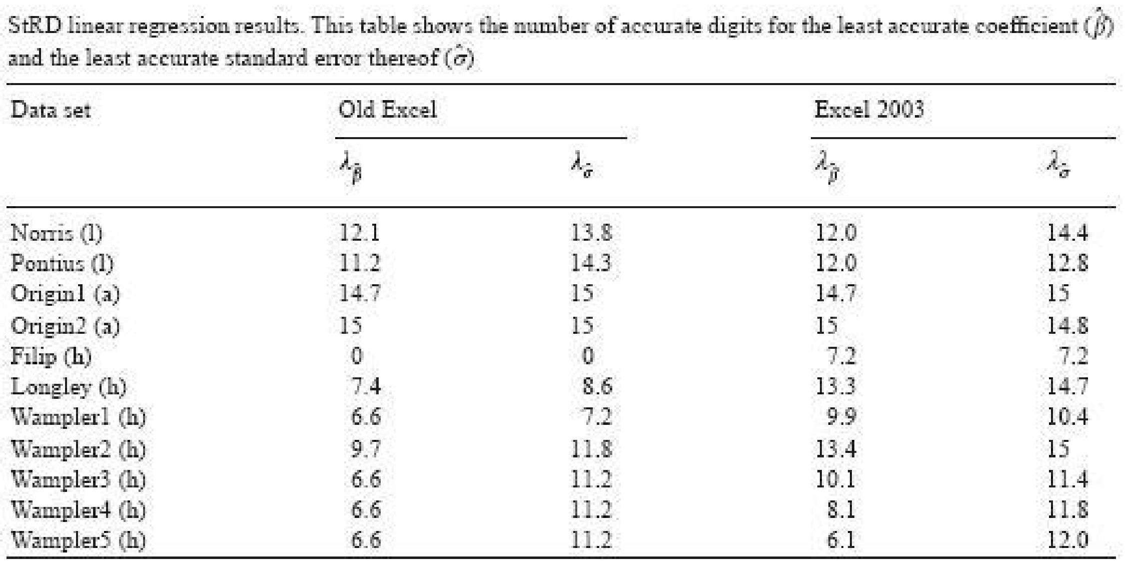
46.2.3 Linear Regression

- *Excel Function:* LINEST
- *Computes:* All numerical results required by Linear Regression

Since LINEST produces many numerical results for linear regression, only the LRE for the coefficients and standard errors of coefficients are taken into account. Table 9 shows the lowest LRE values for each dataset as the weakest link in the chain in order to reflect the worst estimations (smallest λ_β-LRE and λ_σ-LRE) made by Excel for each linear regression function.

Old Excel Versions: either doesn't check for near-singularity of the input matrix or checking it incorrectly, so the results for ill-conditioned Dataset "Filip (h)" include not a single correct digit. Actually, Excel should have refused the solution and commit a warning to user about the near singularity of data matrix. (McCullough & Wilson, 1999, S.32,33) . However, in this case, the user is mislead.

Excel 2003: Problem is fixed and Excel 2003 has an acceptable performance. (see table 10)



StRD linear regression results. This table shows the number of accurate digits for the least accurate coefficient ($\hat{\beta}$) and the least accurate standard error thereof ($\hat{\sigma}$)

Data set	Old Excel		Excel 2003	
	$\lambda_{\hat{\beta}}$	$\lambda_{\hat{\sigma}}$	$\lambda_{\hat{\beta}}$	$\lambda_{\hat{\sigma}}$
Norris (l)	12.1	13.8	12.0	14.4
Pontius (l)	11.2	14.3	12.0	12.8
Origin1 (a)	14.7	15	14.7	15
Origin2 (a)	15	15	15	14.8
Filip (h)	0	0	7.2	7.2
Longley (h)	7.4	8.6	13.3	14.7
Wampler1 (h)	6.6	7.2	9.9	10.4
Wampler2 (h)	9.7	11.8	13.4	15
Wampler3 (h)	6.6	11.2	10.1	11.4
Wampler4 (h)	6.6	11.2	8.1	11.8
Wampler5 (h)	6.6	11.2	6.1	12.0

Figure 45: Table 10: (McCullough & Wilson, 1999, S. 32)

46.2.4 Non-Linear Regression

When solving nonlinear regression using Excel, it is possible to make choices about:

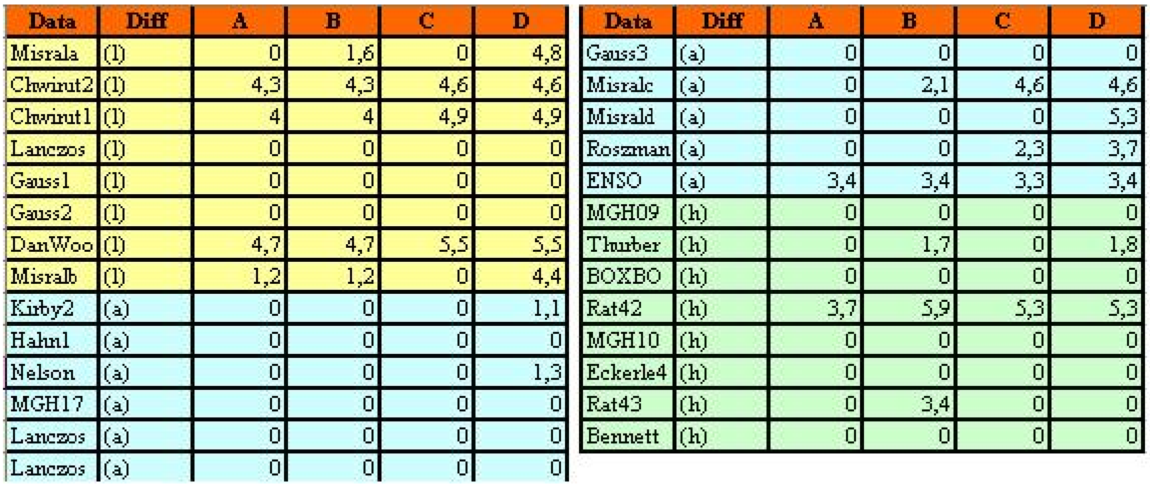
1. method of derivative calculation: forward (default) or central numerical derivatives
2. convergence tolerance (default=1.E-3)
3. scaling (recentering) the variables
4. method of solution (default – GRG2 quasi-Newton method)

Excel's default parameters don't always produce the best solutions always (like all other solvers). Therefore one needs to give different parameters and test the Excel-Solver for nonlinear regression. In table 10 the columns A-B-C-D are combinations of different non-linear options. Because changing the 1st and 4th option doesn't affect the result, only 2nd and 3rd parameters are changed for testing:

- A: Default estimation
- B: Convergence Tolerance = 1E -7
- C: Automatic Scaling
- D: Convergence Tolerance = 1E -7 & Automatic Scaling

In Table 11, the lowest LRE principle is applied to simplify the assessment. (like in linear reg.)

Results in table 11 are same for each Excel version (Excel 97, 2000, XP, 2003)



Data	Diff	A	B	C	D	Data	Diff	A	B	C	D
Misrala	(l)	0	1,6	0	4,8	Gauss3	(a)	0	0	0	0
Chwirut2	(l)	4,3	4,3	4,6	4,6	Misralc	(a)	0	2,1	4,6	4,6
Chwirut1	(l)	4	4	4,9	4,9	Misrald	(a)	0	0	0	5,3
Lanczos	(l)	0	0	0	0	Roszman	(a)	0	0	2,3	3,7
Gauss1	(l)	0	0	0	0	ENSO	(a)	3,4	3,4	3,3	3,4
Gauss2	(l)	0	0	0	0	MGH09	(h)	0	0	0	0
DanWoo	(l)	4,7	4,7	5,5	5,5	Thurber	(h)	0	1,7	0	1,8
Misralb	(l)	1,2	1,2	0	4,4	BOXBO	(h)	0	0	0	0
Kirby2	(a)	0	0	0	1,1	Rat42	(h)	3,7	5,9	5,3	5,3
Hahn1	(a)	0	0	0	0	MGH10	(h)	0	0	0	0
Nelson	(a)	0	0	0	1,3	Eckerle4	(h)	0	0	0	0
MGH17	(a)	0	0	0	0	Rat43	(h)	0	3,4	0	0
Lanczos	(a)	0	0	0	0	Bennett	(h)	0	0	0	0
Lanczos	(a)	0	0	0	0						

Figure 46: Table 11: (McCullough & Wilson, 1999, S. 34)

As we see in table 11, the non-linear option combination A produces 21 times, B 17 times, C 20 times and D 14 times "0" accurate digits. which indicates that the performance of Excel in this area is inadequate. Expecting to find all exact solutions for all problems with Excel is not fair, but if it is not able to find the result, it is expected to warn user and commit that the solution can not be calculated. Furthermore, one should emphasize that other statistical packages like SPSS, S-PLUS and SAS exhibit zero digit accuracy only few times (0 to 3) in these tests (McCullough & Wilson, 1999, S. 34).

46.3 Assessing Random Number Generator of Excel

Many statistical procedures employ random numbers and it is expected that the generated random numbers are really random. Only random number generators should be used that have solid theoretical properties. Additionally, statistical tests should be applied on samples generated and only generators whose output has successfuly passed a battery of statistical tests should be used. (Gentle, 2003)

Based on the facts explained above we should assess the quality of Random Number Generation by:

- analysing the underlying algorithm for Random Number Generation.
- analysing the generators output stream. There are many alternatives to test the output of a RNG. One can evaluate the generated output using static tests in which the generation order is not important. These tests are goodness of fit tests. The second way of evaluating the output stream is running a dynamic test on generator, whereas the generation order of the numbers is important.

46.3.1 Excel's RNG – Underlying algorithm

The objective of random number generation is to produce samples any given size that are indistinguishable from samples of the same size from a U(0,1) distribution. (Gentle, 2003) For this purpose there are different algorithms to use. Excel's algorithm for random number generation is Wichmann–Hill algorithm. Wichmann–Hill is a useful RNG algorithm for common applications, but it is obsolete for modern needs (McCullough & Wilson, 2005, S. 1250). The formula for this random number generator is defined as follows:

$X_i = 171.X_i - 1 mod 30269$

$Y_i = 172.Y_i - 1 mod 30307$

$Z_i = 170.Z_i - 1 mod 30323$

$U_i = \frac{X_i}{30269} + \frac{Y_i}{30307} + \frac{Z_i}{30323} mod 1$

Wichmann–Hill is a congruential generator which means that it is a recursive aritmethical RNG as we see in the formula above. It is a combination of three other linear congruential generator and requires three seeds: $X_0 Y_0 Z_0$.

Period, in terms of random number generation, is the number of calls that can be made to the RNG before it begins to repeat. For that reason, having a long period is a quality measure for random number generators. It is essential that the period of the generator be larger than the number of random numbers to be used. Modern applications are increasingly demanding longer and longer sequences of random numbers (i.e for using in Monte-Carlo simulations) (Gentle, 2003)

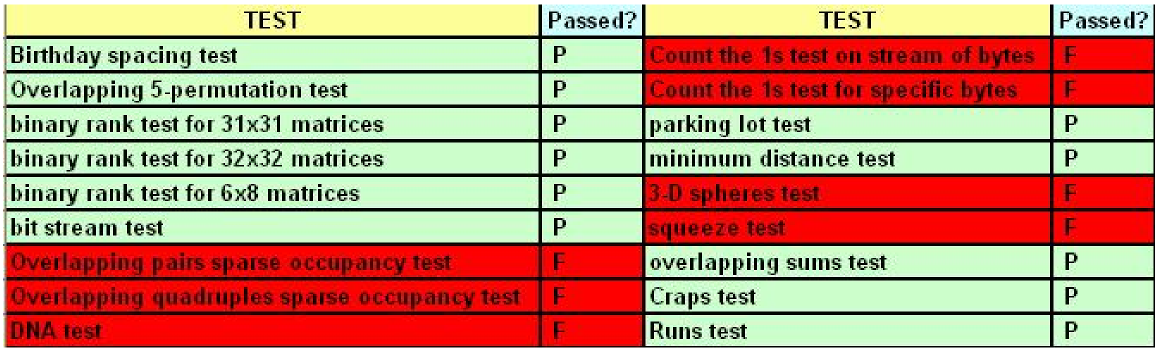
The lowest acceptable period for a good RNG is 2^{60} and the period of Wichmann-Hill RNG is 6.95E+12 ($\approx 2^{43}$). In addition to this unacceptable performance, Microsoft claims that the period of Wichmann-Hill RNG is 10E+13 Even if Excel's RNG has a period of 10E+13, it is still not sufficient to be an acceptable random number generator because this value is also less than 2^{60}. (McCullough & Wilson, 2005, S. 1250)

Furthermore it is known that RNG of Excel produces negative values after the RNG executed many times. However a correct implementation of a Wichmann-Hill Random Number Generator should produce only values between 0 and 1. (McCullough & Wilson, 2005, S. 1249)

46.3.2 Excel's RNG – The Output Stream

As we discussed above, it is not sufficient to discuss only the underlying algorithm of a random number generation. One needs also some tests on output stream of a random number generator while assessing the quality of this random number generator. So a Random Number Generator should produce output which passes some tests for randomness. Such a battery of tests, called DIEHARD, has been prepared by Marsaglia. A good RNG should pass almost all of the tests but as we can see in table 12 Excel can pass only 11 of them (7 failure), although Microsoft has declaired Wichmann–Hill Algorithm is implemented for Excel's RNG. However, we know that Wichmann-Hill is able to pass 16 tests from DIEHARD (McCullough & Wilson, 1999, S. 35).

Due to reasons explained in previous and this section we can say that Excel's performance is inadequate (because of period length, incorrect implementation Wichmann Hill Algorithm, which is already obsolete, DIEHARD test results)



TEST	Passed?	TEST	Passed?
Birthday spacing test	P	Count the 1s test on stream of bytes	F
Overlapping 5-permutation test	P	Count the 1s test for specific bytes	F
binary rank test for 31x31 matrices	P	parking lot test	P
binary rank test for 32x32 matrices	P	minimum distance test	P
binary rank test for 6x8 matrices	P	3-D spheres test	F
bit stream test	P	squeeze test	F
Overlapping pairs sparse occupancy test	F	overlapping sums test	P
Overlapping quadruples sparse occupancy test	F	Craps test	P
DNA test	F	Runs test	P

Figure 47: Table 12: (McCullough & Wilson, 1999, S. 35)

46.4 Conclusion

Old versions of Excel (Excel 97, 2000, XP) :

- shows poor performance on following distributions: Normal, F, t, Chi Square, Binomial, Poisson, Hypergeometric
- retrieves inadequate results on following calculations: Univariate statistics, ANOVA, linear regression, non-linear regression
- has an unacceptable random number generator

For those reasons, we can say that use of Excel 97, 2000, XP for (statistical) scientific purposes should be avoided.

Although several bugs are fixed in Excel 2003, still use of Excel for (statistical) scientific purposes should be avoided because it:

- has a poor performance on following distributions: Binomial, Poisson, Gamma, Beta
- retrieves inadequate results for non-linear regression
- has an obsolete random number generator.

46.5 References

- Gentle J.E. (2003) Random number generation and Monte Carlo methods 2nd edition. New York Springer Verlag
- Knüsel, L. (2003) Computation of Statistical Distributions Documentation of the Program ELV Second Edition. HTTP://WWW.STAT.UNI-MUENCHEN.DE/˜KNUESEL/ELV/ELV_DOCU.PDF[1] Retrieved [13 November 2005]
- Knüsel, L. (1998). On the Accuracy of the Statistical Distributions in Microsoft Excel 97. Computational Statistics and Data Analysis (CSDA), Vol. 26, 375-377.
- Knüsel, L. (2002). On the Reliability of Microsoft Excel XP for statistical purposes. Computational Statistics and Data Analysis (CSDA), Vol. 39, 109-110.
- Knüsel, L. (2005). On the Accuracy of Statistical Distributions in Microsoft Excel 2003. Computational Statistics and Data Analysis (CSDA), Vol. 48, 445-449.
- McCullough, B.D. & Wilson B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics & Data Analysis (CSDA), Vol. 49, 1244 – 1252.
- McCullough, B.D. & Wilson B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. Computational Statistics & Data Analysis (CSDA), Vol. 31, 27– 37.
- PC Magazin, April 6, 2004, p.71*

[1] HTTP://WWW.STAT.UNI-MUENCHEN.DE/~{}KNUESEL/ELV/ELV_DOCU.PDF

47 Authors

Authors and contributors to this book include:

- Cronian[1]
- Llywelyn[2]
- Murraytodd[3]
- Sigbert[4]
- Urimeir[5]
- Zginder[6]

[1] http://en.wikibooks.org/wiki/User%3ACronian
[2] http://en.wikibooks.org/wiki/User%3ALlywelyn
[3] http://en.wikibooks.org/wiki/User%3AMurraytodd
[4] http://en.wikibooks.org/wiki/User%3ASigbert
[5] http://en.wikibooks.org/wiki/User%3AUrimeir
[6] http://en.wikibooks.org/wiki/User%3AZginder

48 Glossary

This is a glossary of the book.

48.1 P

primary data

Original data that have been collected specially for the purpose in mind.

48.2 S

secondary data

Data that have been collected for another purpose and where we will use Statistical Method with the Primary Data.

49 Contributors

Edits	User
1	ACW[1]
2	Abigor[2]
3	AdRiley[3]
2	AdamRetchless[4]
76	Adrignola[5]
1	Albron[6]
1	Aldenrw[7]
13	Alicegop[8]
1	Alsocal[9]
1	Anonymous Dissident[10]
5	Antonw[11]
14	Artinger[12]
1	Avicennasis[13]
2	Az1568[14]
1	Azizmanva[15]
2	Baby jane[16]
2	Benjaminong[17]
16	Bequw[18]
1	Bioprogrammer[19]
5	Blaisorblade[20]
4	Bnielsen[21]

[1] http://en.wikibooks.org/w/index.php?title=User:ACW

[2] http://en.wikibooks.org/w/index.php?title=User:Abigor

[3] http://en.wikibooks.org/w/index.php?title=User:AdRiley

[4] http://en.wikibooks.org/w/index.php?title=User:AdamRetchless

[5] http://en.wikibooks.org/w/index.php?title=User:Adrignola

[6] http://en.wikibooks.org/w/index.php?title=User:Albron

[7] http://en.wikibooks.org/w/index.php?title=User:Aldenrw

[8] http://en.wikibooks.org/w/index.php?title=User:Alicegop

[9] http://en.wikibooks.org/w/index.php?title=User:Alsocal

[10] http://en.wikibooks.org/w/index.php?title=User:Anonymous_Dissident

[11] http://en.wikibooks.org/w/index.php?title=User:Antonw

[12] http://en.wikibooks.org/w/index.php?title=User:Artinger

[13] http://en.wikibooks.org/w/index.php?title=User:Avicennasis

[14] http://en.wikibooks.org/w/index.php?title=User:Az1568

[15] http://en.wikibooks.org/w/index.php?title=User:Azizmanva

[16] http://en.wikibooks.org/w/index.php?title=User:Baby_jane

[17] http://en.wikibooks.org/w/index.php?title=User:Benjaminong

[18] http://en.wikibooks.org/w/index.php?title=User:Bequw

[19] http://en.wikibooks.org/w/index.php?title=User:Bioprogrammer

[20] http://en.wikibooks.org/w/index.php?title=User:Blaisorblade

[21] http://en.wikibooks.org/w/index.php?title=User:Bnielsen

9	Boit[22]
1	Burgershirt[23]
4	Cavemanf16[24]
1	Cboxgo[25]
8	Chrispounds[26]
1	Chuckhoffmann[27]
1	Cronian[28]
4	Dan Polansky[29]
1	DavidCary[30]
7	Derbeth[31]
28	Dirk Hünniger[32]
1	Ede[33]
1	Edgester[34]
5	ElectroThompson[35]
11	Emperion[36]
1	Fadethree[37]
1	Flexxelf[38]
1	Frigotoni[39]
2	Ftdjw[40]
1	Gandalf1491[41]
3	GargantuChet[42]
1	Gary Cziko[43]
3	Guanabot[44]
4	Herbythyme[45]
1	HethrirBot[46]

[22] http://en.wikibooks.org/w/index.php?title=User:Boit
[23] http://en.wikibooks.org/w/index.php?title=User:Burgershirt
[24] http://en.wikibooks.org/w/index.php?title=User:Cavemanf16
[25] http://en.wikibooks.org/w/index.php?title=User:Cboxgo
[26] http://en.wikibooks.org/w/index.php?title=User:Chrispounds
[27] http://en.wikibooks.org/w/index.php?title=User:Chuckhoffmann
[28] http://en.wikibooks.org/w/index.php?title=User:Cronian
[29] http://en.wikibooks.org/w/index.php?title=User:Dan_Polansky
[30] http://en.wikibooks.org/w/index.php?title=User:DavidCary
[31] http://en.wikibooks.org/w/index.php?title=User:Derbeth
[32] http://en.wikibooks.org/w/index.php?title=User:Dirk_H%C3%BCnniger
[33] http://en.wikibooks.org/w/index.php?title=User:Ede
[34] http://en.wikibooks.org/w/index.php?title=User:Edgester
[35] http://en.wikibooks.org/w/index.php?title=User:ElectroThompson
[36] http://en.wikibooks.org/w/index.php?title=User:Emperion
[37] http://en.wikibooks.org/w/index.php?title=User:Fadethree
[38] http://en.wikibooks.org/w/index.php?title=User:Flexxelf
[39] http://en.wikibooks.org/w/index.php?title=User:Frigotoni
[40] http://en.wikibooks.org/w/index.php?title=User:Ftdjw
[41] http://en.wikibooks.org/w/index.php?title=User:Gandalf1491
[42] http://en.wikibooks.org/w/index.php?title=User:GargantuChet
[43] http://en.wikibooks.org/w/index.php?title=User:Gary_Cziko
[44] http://en.wikibooks.org/w/index.php?title=User:Guanabot
[45] http://en.wikibooks.org/w/index.php?title=User:Herbythyme
[46] http://en.wikibooks.org/w/index.php?title=User:HethrirBot

3	Hirak 99[47]
2	Iamunknown[48]
1	Ifa205[49]
1	Isarl[50]
1	Jaimeastorga2000[51]
2	Jakirkham[52]
62	Jguk[53]
3	Jimbotyson[54]
1	Jjjjjjjjjj[55]
3	John Cross[56]
1	John H, Morgan[57]
7	Jomegat[58]
2	Justplainuncool[59]
1	Kayau[60]
2	Krcilk[61]
25	Kthejoker[62]
1	Kurt Verkest[63]
3	Landroni[64]
1	Lazyquasar[65]
6	Littenberg[66]
35	Llywelyn[67]
1	Matt73[68]
71	Mattb112885[69]
2	Matthias Heuer[70]
3	Melikamp[71]

[47] http://en.wikibooks.org/w/index.php?title=User:Hirak_99
[48] http://en.wikibooks.org/w/index.php?title=User:Iamunknown
[49] http://en.wikibooks.org/w/index.php?title=User:Ifa205
[50] http://en.wikibooks.org/w/index.php?title=User:Isarl
[51] http://en.wikibooks.org/w/index.php?title=User:Jaimeastorga2000
[52] http://en.wikibooks.org/w/index.php?title=User:Jakirkham
[53] http://en.wikibooks.org/w/index.php?title=User:Jguk
[54] http://en.wikibooks.org/w/index.php?title=User:Jimbotyson
[55] http://en.wikibooks.org/w/index.php?title=User:Jjjjjjjjjj
[56] http://en.wikibooks.org/w/index.php?title=User:John_Cross
[57] http://en.wikibooks.org/w/index.php?title=User:John_H%2C_Morgan
[58] http://en.wikibooks.org/w/index.php?title=User:Jomegat
[59] http://en.wikibooks.org/w/index.php?title=User:Justplainuncool
[60] http://en.wikibooks.org/w/index.php?title=User:Kayau
[61] http://en.wikibooks.org/w/index.php?title=User:Krcilk
[62] http://en.wikibooks.org/w/index.php?title=User:Kthejoker
[63] http://en.wikibooks.org/w/index.php?title=User:Kurt_Verkest
[64] http://en.wikibooks.org/w/index.php?title=User:Landroni
[65] http://en.wikibooks.org/w/index.php?title=User:Lazyquasar
[66] http://en.wikibooks.org/w/index.php?title=User:Littenberg
[67] http://en.wikibooks.org/w/index.php?title=User:Llywelyn
[68] http://en.wikibooks.org/w/index.php?title=User:Matt73
[69] http://en.wikibooks.org/w/index.php?title=User:Mattb112885
[70] http://en.wikibooks.org/w/index.php?title=User:Matthias_Heuer
[71] http://en.wikibooks.org/w/index.php?title=User:Melikamp

1	Metuk[72]
5	Michael.edna[73]
119	Mike's bot account[74]
7	Mike.lifeguard[75]
10	Mobius[76]
10	Mrholloman[77]
9	Murraytodd[78]
11	Nijdam[79]
23	PAC2[80]
5	Panic2k4[81]
67	Pi zero[82]
1	Pinkie closes[83]
1	Preslethe[84]
1	PyrrhicVegetable[85]
9	QuiteUnusual[86]
1	Ramac[87]
1	Rammamet[88]
1	Ranger2006[89]
1	Ravichandar84[90]
12	Recent Runes[91]
1	Remi Arntzen[92]
1	Robbyjo[93]
32	Saki[94]
1	Sean Heron[95]
10	Sebastian Goll[96]

[72] http://en.wikibooks.org/w/index.php?title=User:Metuk

[73] http://en.wikibooks.org/w/index.php?title=User:Michael.edna

[74] http://en.wikibooks.org/w/index.php?title=User:Mike%27s_bot_account

[75] http://en.wikibooks.org/w/index.php?title=User:Mike.lifeguard

[76] http://en.wikibooks.org/w/index.php?title=User:Mobius

[77] http://en.wikibooks.org/w/index.php?title=User:Mrholloman

[78] http://en.wikibooks.org/w/index.php?title=User:Murraytodd

[79] http://en.wikibooks.org/w/index.php?title=User:Nijdam

[80] http://en.wikibooks.org/w/index.php?title=User:PAC2

[81] http://en.wikibooks.org/w/index.php?title=User:Panic2k4

[82] http://en.wikibooks.org/w/index.php?title=User:Pi_zero

[83] http://en.wikibooks.org/w/index.php?title=User:Pinkie_closes

[84] http://en.wikibooks.org/w/index.php?title=User:Preslethe

[85] http://en.wikibooks.org/w/index.php?title=User:PyrrhicVegetable

[86] http://en.wikibooks.org/w/index.php?title=User:QuiteUnusual

[87] http://en.wikibooks.org/w/index.php?title=User:Ramac

[88] http://en.wikibooks.org/w/index.php?title=User:Rammamet

[89] http://en.wikibooks.org/w/index.php?title=User:Ranger2006

[90] http://en.wikibooks.org/w/index.php?title=User:Ravichandar84

[91] http://en.wikibooks.org/w/index.php?title=User:Recent_Runes

[92] http://en.wikibooks.org/w/index.php?title=User:Remi_Arntzen

[93] http://en.wikibooks.org/w/index.php?title=User:Robbyjo

[94] http://en.wikibooks.org/w/index.php?title=User:Saki

[95] http://en.wikibooks.org/w/index.php?title=User:Sean_Heron

[96] http://en.wikibooks.org/w/index.php?title=User:Sebastian_Goll

4	Senguner[97]
1	Shruti14[98]
113	Sigbert[99]
6	Sigma 7[100]
20	Slipperyweasel[101]
1	Someonewhoisntme[102]
1	Spoon![103]
1	Stradenko[104]
16	Synto2[105]
1	Techman224[106]
1	Technotaoist[107]
1	Timyeh[108]
2	Tk[109]
5	Tolstoy[110]
4	Urimeir[111]
2	Urzumph[112]
1	Waxmop[113]
4	Webaware[114]
2	Whisky brewer[115]
1	Winfree[116]
5	WithYouInRockland[117]
5	WolfVanZandt[118]
1	Wxhor[119]
3	Xania[120]
1	Xerol[121]

[97] http://en.wikibooks.org/w/index.php?title=User:Senguner
[98] http://en.wikibooks.org/w/index.php?title=User:Shruti14
[99] http://en.wikibooks.org/w/index.php?title=User:Sigbert
[100] http://en.wikibooks.org/w/index.php?title=User:Sigma_7
[101] http://en.wikibooks.org/w/index.php?title=User:Slipperyweasel
[102] http://en.wikibooks.org/w/index.php?title=User:Someonewhoisntme
[103] http://en.wikibooks.org/w/index.php?title=User:Spoon%21
[104] http://en.wikibooks.org/w/index.php?title=User:Stradenko
[105] http://en.wikibooks.org/w/index.php?title=User:Synto2
[106] http://en.wikibooks.org/w/index.php?title=User:Techman224
[107] http://en.wikibooks.org/w/index.php?title=User:Technotaoist
[108] http://en.wikibooks.org/w/index.php?title=User:Timyeh
[109] http://en.wikibooks.org/w/index.php?title=User:Tk
[110] http://en.wikibooks.org/w/index.php?title=User:Tolstoy
[111] http://en.wikibooks.org/w/index.php?title=User:Urimeir
[112] http://en.wikibooks.org/w/index.php?title=User:Urzumph
[113] http://en.wikibooks.org/w/index.php?title=User:Waxmop
[114] http://en.wikibooks.org/w/index.php?title=User:Webaware
[115] http://en.wikibooks.org/w/index.php?title=User:Whisky_brewer
[116] http://en.wikibooks.org/w/index.php?title=User:Winfree
[117] http://en.wikibooks.org/w/index.php?title=User:WithYouInRockland
[118] http://en.wikibooks.org/w/index.php?title=User:WolfVanZandt
[119] http://en.wikibooks.org/w/index.php?title=User:Wxhor
[120] http://en.wikibooks.org/w/index.php?title=User:Xania
[121] http://en.wikibooks.org/w/index.php?title=User:Xerol

1	YanWong[122]
1	Youssefa[123]
7	ZeroOne[124]
11	Zginder[125]

[122] http://en.wikibooks.org/w/index.php?title=User:YanWong

[123] http://en.wikibooks.org/w/index.php?title=User:Youssefa

[124] http://en.wikibooks.org/w/index.php?title=User:ZeroOne

[125] http://en.wikibooks.org/w/index.php?title=User:Zginder

List of Figures

- GFDL: Gnu Free Documentation License. http://www.gnu.org/licenses/fdl.html
- cc-by-sa-3.0: Creative Commons Attribution ShareAlike 3.0 License. http://creativecommons.org/licenses/by-sa/3.0/
- cc-by-sa-2.5: Creative Commons Attribution ShareAlike 2.5 License. http://creativecommons.org/licenses/by-sa/2.5/
- cc-by-sa-2.0: Creative Commons Attribution ShareAlike 2.0 License. http://creativecommons.org/licenses/by-sa/2.0/
- cc-by-sa-1.0: Creative Commons Attribution ShareAlike 1.0 License. http://creativecommons.org/licenses/by-sa/1.0/
- cc-by-2.0: Creative Commons Attribution 2.0 License. http://creativecommons.org/licenses/by/2.0/
- cc-by-2.0: Creative Commons Attribution 2.0 License. http://creativecommons.org/licenses/by/2.0/deed.en
- cc-by-2.5: Creative Commons Attribution 2.5 License. http://creativecommons.org/licenses/by/2.5/deed.en
- cc-by-3.0: Creative Commons Attribution 3.0 License. http://creativecommons.org/licenses/by/3.0/deed.en
- GPL: GNU General Public License. http://www.gnu.org/licenses/gpl-2.0.txt
- PD: This image is in the public domain.
- ATTR: The copyright holder of this file allows anyone to use it for any purpose, provided that the copyright holder is properly attributed. Redistribution, derivative work, commercial use, and all other use is permitted.
- EURO: This is the common (reverse) face of a euro coin. The copyright on the design of the common face of the euro coins belongs to the European Commission. Authorised is reproduction in a format without relief (drawings, paintings, films) provided they are not detrimental to the image of the euro.
- LFK: Lizenz Freie Kunst. http://artlibre.org/licence/lal/de
- CFR: Copyright free use.
- EPL: Eclipse Public License. http://www.eclipse.org/org/documents/epl-v10.php

1		GPL
2	User:Webaware[126]	PD
3	User:Webaware[127]	PD
4		GFDL
5		GFDL
6		PD
7		PD
8	Ryan Cragun	PD
9		GFDL
10		PD
11		PD
12		PD
13		GFDL
14		GFDL
15	Alicegop[128]	PD
16	Alicegop[129]	PD
17	Winfree[130]	cc-by-sa-3.0
18		GFDL
19		PD
20		PD
21		PD
22		PD
23		cc-by-sa-2.5
24		cc-by-sa-2.5
25		cc-by-sa-2.5
26		cc-by-sa-2.5
27		cc-by-sa-2.5
28		cc-by-sa-2.5
29		cc-by-sa-2.5
30		cc-by-sa-2.5
31		cc-by-sa-2.5
32		cc-by-sa-2.5
33		cc-by-sa-2.5
34		cc-by-sa-2.5
35		cc-by-sa-2.5
36		cc-by-sa-2.5
37		cc-by-sa-2.5
38		cc-by-sa-2.5
39		cc-by-sa-2.5
40		cc-by-sa-2.5
41		cc-by-sa-2.5
42		cc-by-sa-2.5
43		cc-by-sa-2.5
44		cc-by-sa-2.5
45		cc-by-sa-2.5

[126] http://en.wikibooks.org/wiki/User%3AWebaware

[127] http://en.wikibooks.org/wiki/User%3AWebaware

[128] http://en.wikibooks.org/wiki/User%3AAlicegop

[129] http://en.wikibooks.org/wiki/User%3AAlicegop

[130] http://en.wikibooks.org/wiki/User%3AWinfree

46		cc-by-sa-2.5
47		cc-by-sa-2.5