Lecture 6

Methodological questions concerning data, simulations and statistics

Some methodology

We will discuss the following:

- Some terminology and facts about data
- Some standard methods in statistics
- Some methods in computer simulations

Data

A part of research is to handle data of different types. It appears that one can describe data in several ways.

- Description after level of abstraction
- Split into primary and secondary data
- Quantitative and qualitative data
- Measurement for different types of scale

Primary and Secondary Data

- Primary data - direct measurements or observations of something. Can also be reports of people who themselves have experienced something
- Secondary data - is often compilation of primary data that has been processed in any form. They can e.g. occur in reports, articles or books

Quantitative and qualitative data

- Quantitative data - data given in the form of numbers
- Qualitative data - data that can not easily be given in the form of numbers. It may be opinions, stories or descriptions of situations

Collect and analyze secondary data

- Typical examples can be written materials novels, reports, biographies, newspapers, etc.
- It could be television programs, films, interviews, etc.
- It may be statistical surveys made for some other purpose

Problems with the analysis

- To find data
- To authenticate the sources
- Assessing credibility
- To determine how representative the data is
- Choosing methods for interpretation of the data

Three methods of analysis

- Content analysis - we simply count the occurrences of something such as words or types of images in the document and use occurrences as indicators
- Data mining - We use software to find patterns in the data
- Meta-analysis - We make an analysis of several other analyzes simultaneously and try to see patterns in them

To collect primary data

The amount of methods is great. However, we can mention three main areas

- Statistical surveys. Sampling.
- Interview Methods.
- Experiment. Some general methods for experiments is that in addition to the real experimental group we also have a control group. We should also, if possible, use so-called double-blind tests.

Some methods in statistics

Given a set D of data we want to find the best model for dscribing the data. What best means is of course a difficult question. Normally, we think of a *simple* structure that in some sense fits the data well.

We some parameter A that we want to find the value of it. We have measured w. How sure can we be if it is the correct value or how close it is to the correct value?

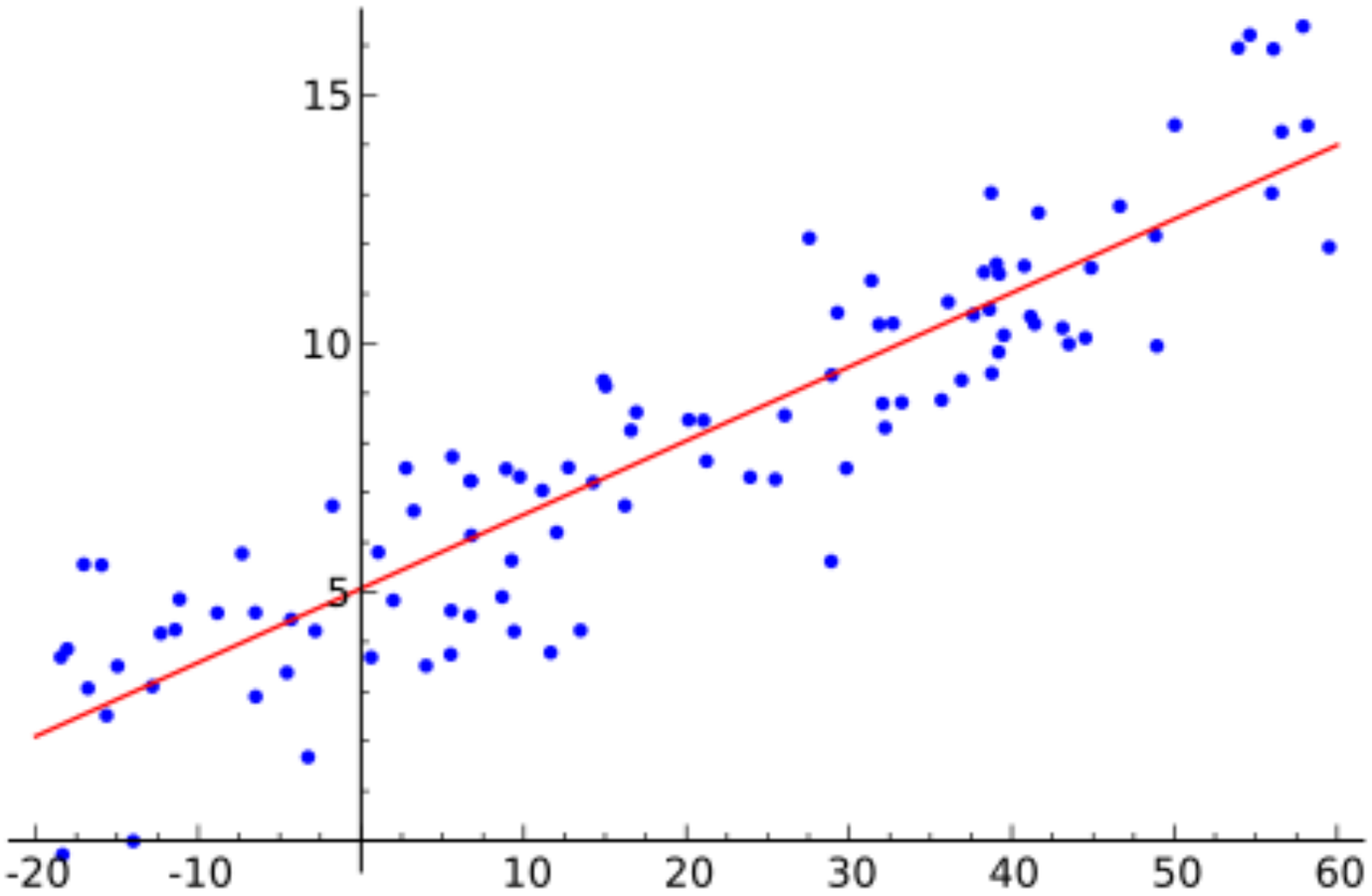
We have an hypothesis H. What tests can we perform to tell if it is true or not?

Least-square optimization

Let us assume that we have a set of data in the form $(x_1, y_1), x_2, y_2), ..., (x_n, y_n)$. Let us assume that the connection between x and y is $y = f(x)$ and that f should belong to a set M. Then we assume that f is the $f \in M$ that minimizes $\sum (y_i - f(x_i))^2$.

This method can be used in several different forms.

Linear regression



Confidence intervalls

Let us assume that we have measured a parameter Θ. We have got a measured value v. Can we estimate how far from v the *true* value Θ is?
In some cases we can assume that the stochastic variable v has normal distribution $N(\Theta, \sigma)$. We can even assume that we know the standard deviation σ. We want to find an intervall $CI = [v - \epsilon, v + \epsilon]$ such that the probability that Θ belongs to CI is α where α is a chosen confidence level. (We could take $\alpha = 0.95$.)
How do we find ϵ? The standard way is to compute like this: $\frac{v-\Theta}{\sigma}$ is $N(0,1)$-distributed.

$$P(-q_{0.0925} \leq \frac{v-\Theta}{\sigma} \leq q_{0.0925}) = 0.95$$

$$P(-q_{0.0925} \leq \frac{v-\Theta}{\sigma} \leq q_{0.0925}) =$$

$$P(v - q_{0.0925}\sigma \leq \Theta \leq v + q_{0.0925}\sigma)$$

So we can set $\epsilon = q_{0.925}$ where $q_{0.925}$ is the so called 0.925-quantile.

Hypothesis testing

Let us assume that we have a hypothesis H that we want to test. We compare it to a zero-hypothesis H_0. We design a test which gives us a value t. We define a set C such that we can reject H_0 if t is a member of C. (That means that we accept H.) In that case we say that the test is significant at level α if the probability that t belongs to C is less than or equal to α , given the assumption that H_0 is true. If we get $t \in C$ we say that the test is significant on the $(1-\alpha)$-level and that H passes the test on significance level α. The probability α is usually small, like 0.05, 0.01 or 0.001.

An example

Let us assume that we have a parameter Ψ and the null-hypothesis H_0 is that the value of $\Psi = \psi_0$. But now we have a rival hypothesis that the value of Ψ is some value larger than ψ_0. We try to measure Ψ. We get a value $\psi_1 > \psi_0$. Can we reject H_0?

We must have some assumptions on the distribution on measurements of Ψ. Let us assume that, *if* H_0 *is true*, the measured values are $N(\psi_0, \sigma)$-distributed.
Then $P(\frac{\psi - \psi_0}{\sigma} \leq q_{0.95}) = 0.05$. This means that
$P(\psi \leq \psi_0 + q_{0.95}\sigma) = 0.95$. So if
$\psi > \psi_0 + q_{0.95}\sigma$ we can reject H_0 on 5% significance level.

Some more examples:

Test of Hypotheses

Level of significance

A company makes components with a mean life span of 100 h. Some researchers claim taht they have found a way to increase the life span to 110 h. We know that the life span is normal distributed with $N(\mu, 15)$. We measure the life spans of components produced with the new method and get the results

{115.3, 106.5, 110.6, 106.9, 95.4, 115.1, 112.9, 107.7, 109.7}

To test if the new method really works we state a zero hypothesis:

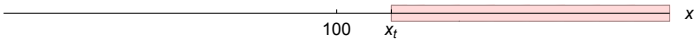
$H_0: \ \mu = \mu_0 = 100$ h (as in the old method)

Against this hypothesis we state a counter hypothesis:

$H_1: \ \mu = \mu_1 = 110$ h (new method)

We now want to reject H_0 (and accept H_1) with a certain risk of error α which we call the significance level. Let us try to get the level 5 %.

We reason like this: We estimate μ with $\bar{x}$. If $\bar{x}$ falls into the red area, i.e. $\bar{x} > x_t$ we reject H_0 and accepts the counter hypothesis H_1.



We now chose x_t such that so that the error margin is less than 5 %, i.e., the probability that $\bar{x} > x_t$ *if* H_0 is true is less than 5 %. The treshhold value x_t will be the percentile $x_{0.95}$ for the distribution $\overline{X} \approx N\left(100, 15/\sqrt{9}\right) = N(100, 5)$. The probability is then 95 % that $\bar{x}$ is less than x_t. Let $q_{0.95}$ be the percentile for $N(0, 1)$.

$$\frac{x_t - 100}{5} = q_{0.95} \Leftrightarrow x_t = 100 + 5\, q_{0.95} \approx 100 + 5 \times 1.64485 \approx 108.224$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 108.9$$

Since $\bar{x} > x_t$ we reject H_0. Our error risk is $\alpha = 1 - 0.95 = 0.05$ and is called the significance level.

$$\alpha = P(H_0 \text{ rejected} \mid H_0 \text{ true}) = P\left(\overline{X} > x_t \mid \overline{X} \approx N(100, 5)\right) = 1 - 0.95 = 0.05$$

We can note that if we want the significance level to be 1 % we can not reject H_0 since

$$x_{0.99} = 100 + 5\, q_{0.99} \approx 100 + 5 \times 2.32635 \approx 111.632$$

We can not say that the researchers are wrong. We can just say we can not reject $H_0$with significance level 1%.

The P-value Method

Instead of dtermining the significance level we can estimate the probability of our measured value. We call this method the p-value method. In our example we get

$$p = P\left[\overline{X} > \bar{x} \mid \overline{X} \approx N(100, 5)\right] = 1 - \text{cdf}_0(\bar{x}) = 1 - \Phi\left(\frac{\bar{x} - 100}{5}\right)$$

$$= 1 - \Phi\left(\frac{108.9 - 100}{5}\right) = 1 - \Phi(1.78) \approx 0.037538$$

We then see that the test “passes” the significance level 3.8 %, i.e. , we can reject H_0 with significance level 3.8 %.

Strenght Function

We continue our example. We will try to define the "strength" of the test. We assume that the life span of the components always have normal distribution with standard deviation 5. We measure:

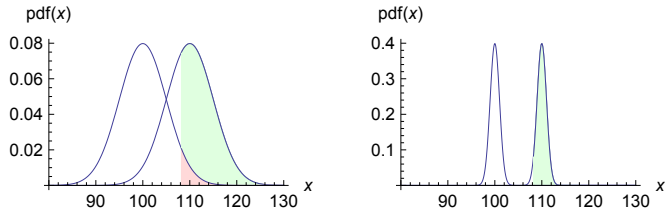
$$P(H_0 \text{ rejected} \mid H_1 \text{ true}) = P\left(\overline{X} > x_t \mid \overline{X} \approx N(110, 5)\right) = 1 - \text{cdf}_1(x_t)$$

In our case we get

$$1 - \text{cdf}_1(x_t) = 1 - \Phi\left(\frac{x_t - 110}{5}\right) \approx 1 - \Phi\left(\frac{108.224 - 110}{5}\right) \approx 1 - \Phi(-0.355146) = \Phi(0.355146) \approx 0.639$$

The strenght, that H_0 is rejected if H_1 is true is just 64 %.

What we want is a low significance level an a high strength. This is possible if we take a large sample so that the standard deviation is low. In the figure to the right we have $n = 25 \times 9$ and the strength of the test is 96 %.



We can define the strenght function of the test as:

$$h(\theta) = P(H_0 \text{ rejected} \mid \theta \text{ is the true value for the parameter})$$

In our example we get

$$h(\mu) = P\left(\overline{X} > x_t \mid \overline{X} \approx N(\mu, 5)\right) = 1 - \text{cdf}_\mu(x_t) = 1 - \Phi\left(\frac{x_t - \mu}{5}\right)$$

Computer Simulations

There is one area of science that is probably relatively unexplored: The Philosophy of Computer Simulation.

Normally we observe reality and make observations. We can use computer methods to analyze data.

But we can make computer simulations of ”reality” instead.

A formal aspect of computer methods

Let us assume that we have a process P that we want to simulate

- We might believe that we understand the process
- We the try to write a program simulating the process
- When we write the program we might run into difficulties which forces us to rethink our understanding of the process
- So even without running the program we might gain understanding of the model in which the process is living in

Example:The Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors:
Behind one door is a car; behind the others, goats.
You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
He then says to you, "Do you want to pick door No. 2?"
Is it to your advantage to switch your choice?

Solution?

It is said that it is to your advantage to alter your choice. Do you believe it. Some think it is impossible. But there are formal proofs that it is.

What to believe? Can we test it in ”reality”?

Computer Simulation

Let us try to write a program that tests if it is a good strategy to alter your choice. How do you design a program.

- There are some random components.
- Then there are some assumptions of the choices the actors (two) make.

Some details

Here are some steps:

- You distribute the car and the goats randomly
- The guest choses randomly
- Then you have to simulate the host's acting. What does the host know. What strategy does he use? - This is a critical point
- Then the simulation of the guest is simple - just change door

- Run this many times. Count the number of times the guest wins the car.
- Make the same simulation when the guest does not change door. Result?