Lecture 9

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

Levels of Abstraction

- Theory - The key interactions expressed in abstract terms
- Concept - The basic concepts. The theory often express connection between concepts
- Indicators - Something measurably which is related to concepts or demonstrate the existence of them
- Variables - The measurable component of the indicators
- Values - the results of measurements of the variables

Scientific data and observations

- The previous observations have been such that their values are true / false.
- In other contexts we assign observations a numeric value: observations E1, E2, ..., get the values f (E1), f (E2), ...
- The function f defines the type of scale we use.
- What kind of scale you use defines the type of information one can objectively be read out of the observations.

Types of scales

- Ordinal scale : Relations of type f (E1) < f (E2) are significant. An ordinal is just a ranking of the observations.
- Quota Scale: The value f (E2) / f (E1) is significant.
- Interval Scale: value (f (E2) -f (E0)) / (f (E1) -f (E0)) is significant. Typical examples are temperature scales (except Kelvin's).

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

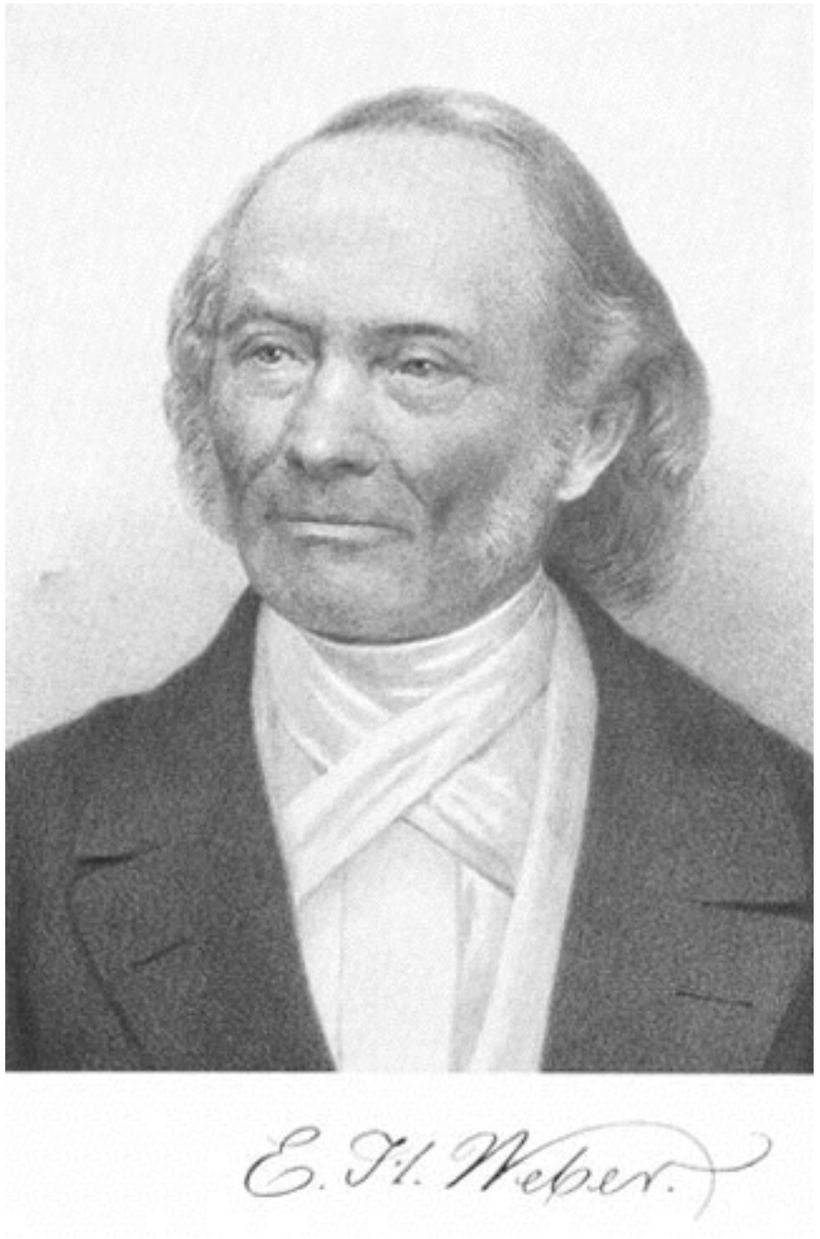
- Statistical surveys. Sampling.
- Interview Methods.
- Experiment. Some general methods for experiments is that in addition to the real experimental group we also have a control group. We should also, if possible, use so-called double-blind tests.

The 'mathematization' of the world

- One of the reasons for the success of mathematics in science is the possibility of measuring things and then doing mathematical processing of the data.
- This fact can lead to the opinion that only measurable facts can be the subject of science.
- But in Social Sciences it is often claimed that *qualitative* data are as important as *quantitative.*
- We will illustrate how it is possible to use how it is possible to use mathematics to define measures on seemingly qualitative and subjective observations.

Weber's law

- Does 10 kg feel twice as 'heavy' as 5 kg?
- If we have a body with of weight m, can we find a function f(m) which measures the subjective 'heaviness' we experience?
- Yes, it is possible.

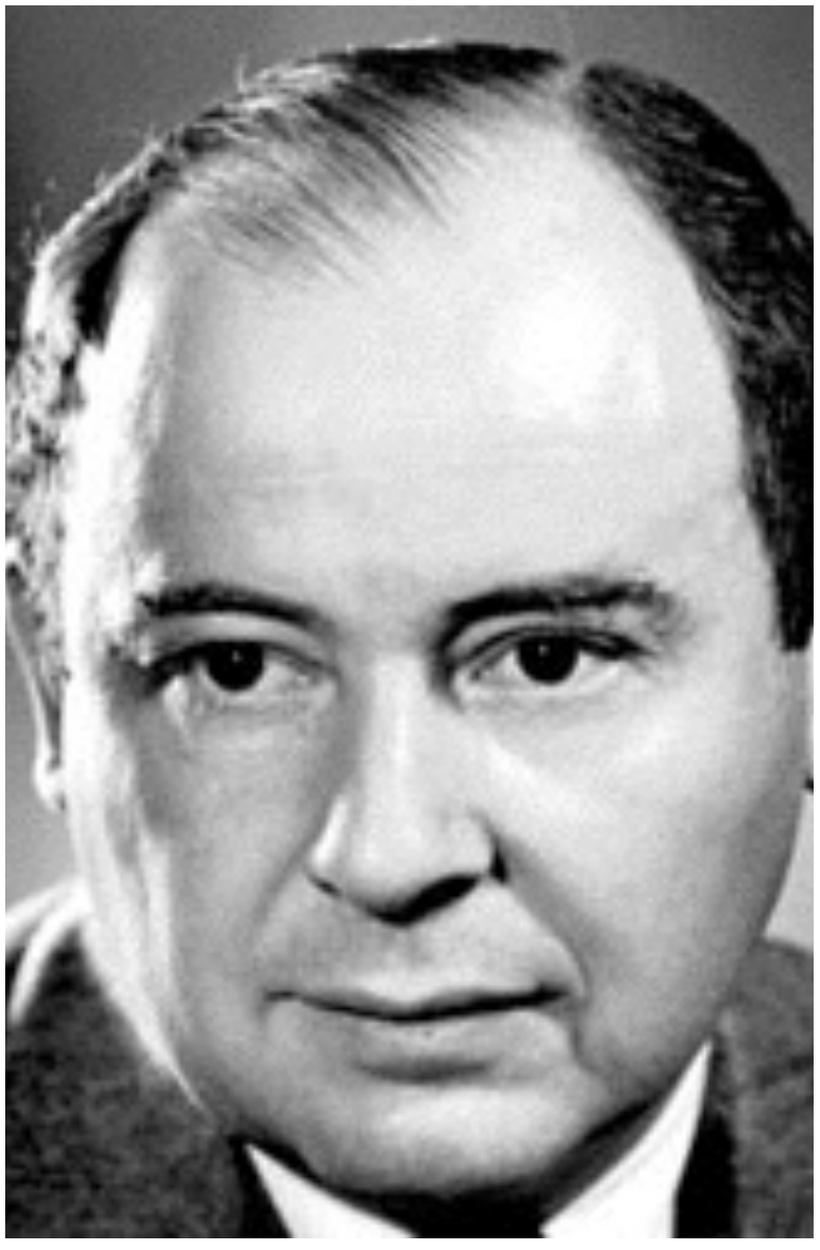


Weber's law II

- Let dm be the smallest change in mass that we (subjectively) can detect with our senses.
- In the beginning of the 19th century Weber showed that dm is linearly dependent on m, i.e. dm = cm for some constant m. (In an appropriate interval.)
- From this it is not hard to see that a natural definition of the subjective experience of 'heaviness' has the form k log m + w0 for some constants k, w0.

Utility theory

- Would you like to go to an interesting early morning lecture at KTH?
- Or would you rather sleep some hours more?
- Can you measure how much you want different things?
- John von Neumann suggested a way of measuring subjective preferences in an exact way.



Utility theory II

- In Game Theory and Mathematical Economics we make the assumption the we can personally order different things a, b, c ,... In preference order.
- We also make the assumption that we can measure how much we want different things by a utility function u.
- This means that we prefer e to q if and only if u(e) > u(q).
- How is it possible to define such a function?

"What do you chose?
5000 $ or the secret box?"

- The idea von Neumann had was to imagine a virtual lottery. Let a be the thing we prefer most of everything and z the thing we prefer least of everything.
- Now, let L1 be a lottery with two possible outcomes: You get a or you get z. You get a with probability p and z with probability (1-p).
- And then, we take a thing k and imagine a trivial lottery L2 where you get k with certainty.
- Which one do you prefer? L1 or L2? It must depend on p.
- There should be a number p such that you are *indifferent* between L1 and L2. This p is the utility of k, i.e. u(k) = p.
- This means that u(a) =1 and u(z) = 0.

Is it a realistic measure?

- To put it a little extreme, let us say that we want to compare *everything.*
- Let us then assume that a is "Happiness beyond all imagination" and z is "A horrible and painful death". Let k be "Attending a lecture in The Theory of Science".
- So L1 is the lottery [P(Happiness beyond all imagination) = p, P(A horrible and painful death) = (1-p)]
- And L2 is the lottery [Attending a lecture in The Theory of Science with certainty].
- For what p are you indifferent between L1 and L2?

Statistics

KTH/ICT
IX1501:F7
Göran Andersson
goeran@kth.se

Statistics

- We shall now study the observations of random variables. What values did we get?
- These observations are called one sample (sample).

Within the statistics, there are essentially four main areas

point estimation

interval estimation

hypothesis testing

Decision Problems

Point Estimation

- A sample $x = \{x_1, x_2, x_3, ...\}$ is an observation of stochastic variables $\xi = \{\xi_1, \xi_2, \xi_3, \ldots\}$.

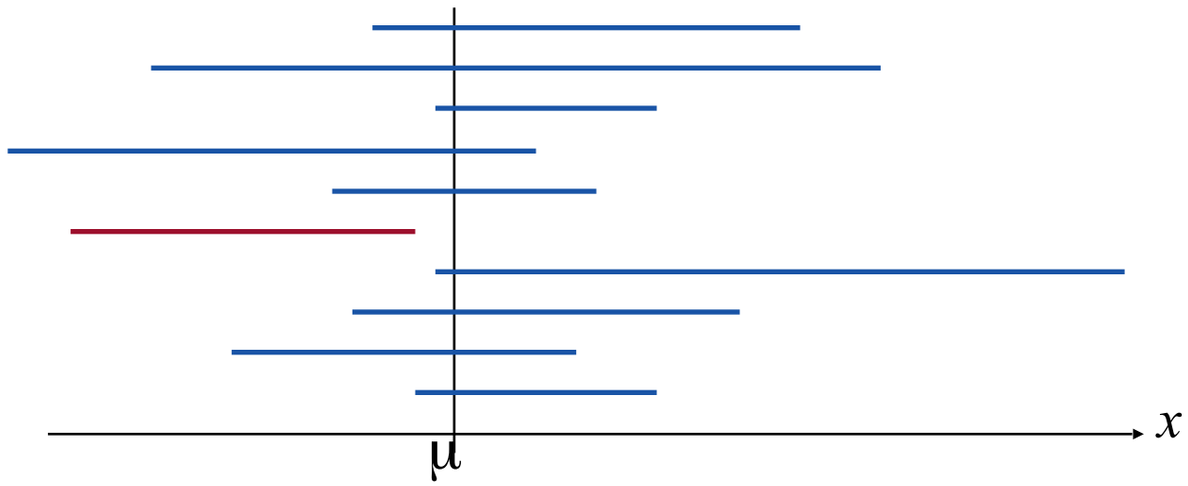
- An estimate is a function $f(x)$ and can be regarded as an observation of $f(\xi)$.

- An estimate of a parameter, μ, is said to be unbiased if the expected value of the estimate is the value of the parameter.

- An estimate is said to be consistent if the difference to the real value goes to 0 as the sample size grows.

Confidence Intervals

- A confidence interval for a parameter is an interval with a given confidence level (probability).
- The confidence level is the probability that the intervall covers the true value of the parameter.
- Different samples give different intervals.



- We try to find an intervall with confidence level 95% for μ based on independent observations of $\xi \in N(\mu, \sigma)$.
  - μ is unknown and is estimated with the mean value. We set

$$\eta = \frac{\bar{\xi} - \mu}{\sigma / \sqrt{n}} \Rightarrow \eta \in N(0,1)$$

- Then $P(q_{0.025} < \eta < q_{0.975}) = 0.95$

$$\Leftrightarrow P\left(q_{0.025} < \frac{\bar{\xi} - \mu}{\sigma / \sqrt{n}} < q_{0.975} \right) = 0.95$$

$$\Leftrightarrow P\left(\bar{\xi} - q_{0.975} \frac{\sigma}{\sqrt{n}} < \mu < \bar{\xi} - q_{0.025} \frac{\sigma}{\sqrt{n}} \right) = 0.95 \quad \Leftrightarrow P\left(\bar{\xi} - q_{0.975} \frac{\sigma}{\sqrt{n}} < \mu < \bar{\xi} + q_{0.975} \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

- We get the interval

$$\mathrm{CI}_{\mu}: \quad \bar{x} \pm q_{0.975} \frac{\sigma}{\sqrt{n}} \quad (95\%)$$

- One problem occurs when we have an unknown distribution. In many cases we assume that we have a normal distribution. The famous Central Limit Theorem (CLT) gives us good reasons to assume this. The problem is that the standard deviation might be unknown. Then we use the previous estimate for it. In that case we must use percentiles coming from the t-distribution.

σ known $$\mathrm{CI}_\mu : \ \bar{x} \pm q^N_{\frac{1}{2}(1+\alpha)} \frac{\sigma}{\sqrt{n}} \quad (\text{konfidensgrad} \approx \alpha)$$

otherwise $$\mathrm{CI}_\mu : \ \bar{x} \pm q^t_{\frac{1}{2}(1+\alpha)}(n-1) \frac{s}{\sqrt{n}} \quad (\text{konfidensgrad} \approx \alpha)$$

n	$q^t_{0.975}(n-1)$
10	2.26216
100	1.98422
1000	1.96234
∞	$1.95996 = q^N_{0.975}$

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:

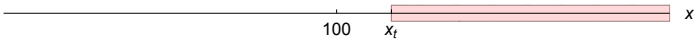
$H_0: \ \mu = \mu_0 = 100$ h (as in the old method)

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:

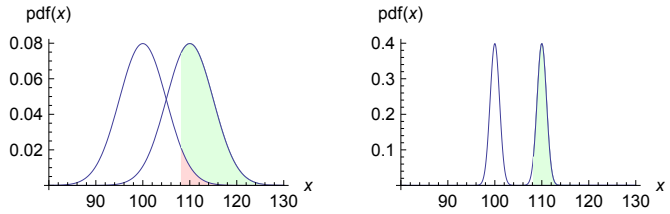
$$P(H_0 \text{ rejected} \mid H_1 \text{ true}) = P\left(\overline{X} > x_t \mid \overline{X} \approx N(110, 5)\right) = 1 - \text{cdf}_1(x_t)$$

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?