4.3 How to collect and get data

Sampling plan: How many subjects / items should be sampled?

Depends on:

- Presence of the amount of uncertainty in the quantity one is looking at
- How much that uncertainty must be reduced for making a proper decision
- The degree of error contained in the responses of a subject / item

4.3 How to collect and get data

Sampling plan: Coming from simple statistics, what are the usual suspects?

- Width of the confidence interval
- Confidence level
- Power of a test

4.3 How to collect and get data

Sampling plan: Usual assumptions

- Independence
- Identically distributed
- No systematic error

Further nice assumption: Data are from the exponential family (e.g. normal, exponential, Poisson, Gamma, Bernoulli, Chi-squared, Beta, binomial and multinomial (if n is fixed), negative binomial (if x fixed),…)

Thus, the sample size n can be determined quite easily.

Rule of thumb: Quadrupling the number of subject / items sampled reduces the uncertainty by half

4.3 How to collect and get data

Sampling plan: And in practice?

- As a starting point, it is often assumed that one has independence (or if obvious not)

Example: The decision that a customer is buying a new car is independent from the choice of another customer who is buying a new car.

Example: A car accident of one customer is independent from the car accident of another customer (except in few cases).

4.3 How to collect and get data

Sampling plan: And in practice?

- Identically distributed: often not

Example: The decision which car is bought depends on the needs, the wants and the salary level

Thus, we need homogenous subclasses for the analysis, but then maybe the sample is too small for performing meaningful data analytics.

4.3 How to collect and get data

Sampling plan: And in practice? (cont'd)

- No systematic error: ?

Systematic errors are very difficult to detect in practice.

Obvious ones are drifts or if the subsample is very small.

Example: In a customer analytics project for a special subgroup with very special attributes an analysis is performed. One have found out that a subgroup of 7 people is mainly buying this product in St.Moritz.

4.3 How to collect and get data

Determination of the questions to be asked

We have seen that we have to types of data (see 4.2).

- *Primary data* are data which are not yet available and has to be measured and collected first.
- *Secondary data* are data already collected by someone else.

$\Rightarrow$ Thus, the questions may be different for each type of data

4.3 How to collect and get data

Determination of the questions to be asked: Primary data

- One have data collected out of systems, machines, by counting, including meta data.

  Examples:

  - Data collection of failures during the production of wash machines
  - Data of car traffic on certain roads
  - Data of people who are buying Nespresso capsules and which tastes

- Or one have data out of interviews and questionnaires.

  Example:

  - Evaluation of customer satisfaction with the IT support

4.3 How to collect and get data

Determination of the questions to be asked: Primary data

Some remarks to interview questions and questionnaires:

There are four approaches for collecting data

- Categorical
- Semantic differential
- Rank-order
- Multiple-choice

4.3 How to collect and get data

Determination of the questions to be asked: Primary data

Categorical

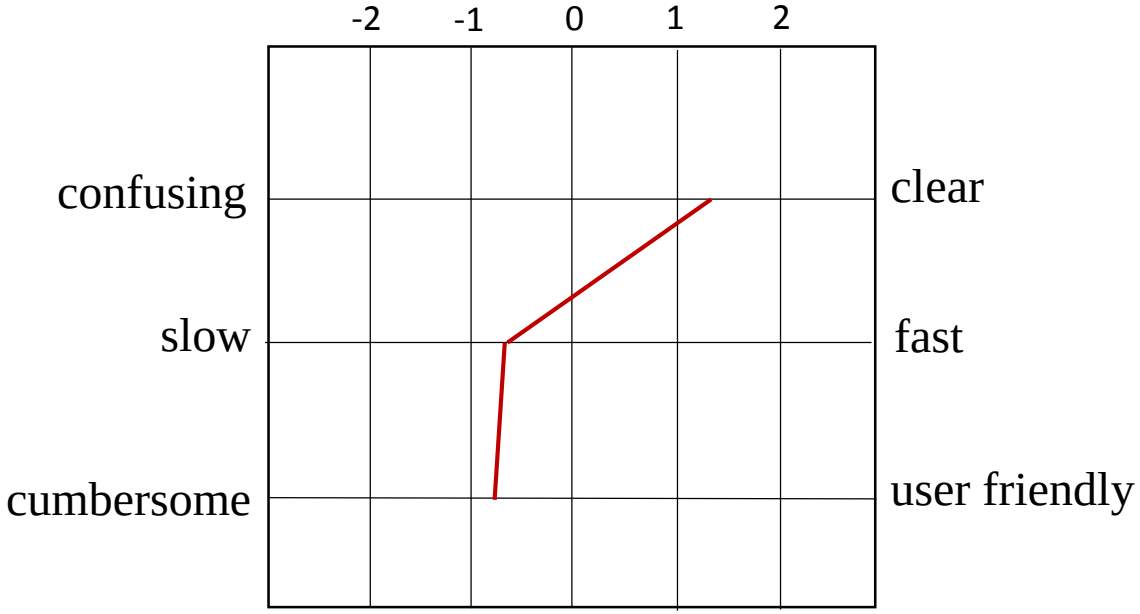
- What is your preferred car colour? [black, silver, red, white,…]
- Would you buy that product? [yes / no]
- Are you satisfied with the friendliness of our call center service? [fully agree, partially agree, are neutral, partially disagree, fully disagree]

4.3 How to collect and get data

Determination of the questions to be asked: Primary data

Semantic differential

What is your experience in navigating through our web page?



4.3 How to collect and get data

Determination of the questions to be asked: Primary data

Rank-order

From 1 to 5, please rank the most important benefits of our tablet:

[ ] weight

[ ] water resistance

[ ] screen

[ ] size

[ ] price

4.3 How to collect and get data

Determination of the questions to be asked: Primary data

Multiple-choice

What is the year of birth of Galileo Galilei?

- ☐ 1612
- ☐ 1514
- ☐ 1564
- ☐ 1642

4.3 How to collect and get data

Determination of the questions to be asked: Secondary data

Secondary data are already available.

Thus, the main task is to find them and to assess them for appropriateness.

4.3 How to collect and get data

Determination of the questions to be asked: Secondary data

Task: Think of about a airplane construction company and you have to collect data for the analytics project for optimizing a construction process.

You are the first day on premise and no further information about the data.

Discuss with your neighbour what would be the questions you would ask for getting data?

4.3 How to collect and get data

Determination of the questions to be asked: Secondary data

- What data are available?
- Do you have a data inventory?
- Are there references to the source of the data?
- What was the primary purpose for collecting the data?
- Who (what) has collected the data?
- When the data have been collected?
- Are these raw data or already cleaned data?

4.3 How to collect and get data

Determination of the questions to be asked: Secondary data

- Are meta data for the data available?
- What is the coverage of the data (time frame, business area, etc)?
- Are the data on an aggregated basis or a single basis?
- What is the type of data?
- What is the volume of these data?
- Who has the responsibility for that data?
- Are the data accessible in due time?

4.3 How to collect and get data

Determination of the questions to be asked: Secondary data

- From how many systems / sources are the data?
- What is the cost for accessing the data?
- What are the users of these data?
- Are there reports where these data are used?

4.3 How to collect and get data

Determination of the control group

- In some analytics project there is the need for a so-called control group: Tests / studies / treatments are applied on a test or treatment group.
- To compare the test results with the current or standard status, a *control group* out of the same population are chosen where still the current or standard test / treatment is applied.

4.3 How to collect and get data

Determination of the control group

Examples: New marketing strategies. If a person has visited the web page of the insurance company X and text is send to the mobile with an offer for a household insurance policy.

To test the efficiency of this marketing strategy, a sample is randomly selected where no offer per text is send but the usual standard marketing letter for this household insurance.

Excurse: Sentimental Analysis Introduction

Some background:

Some years ago, when a company wanted to know about the opinion or sentiment of its customer surveys have to be conducted.

But the world and the internet have significantly changed.

And today you have reviews for products, blogs, forums, groups – all together called *"user-generated content"*

Thus, instead conducting a survey, the corresponding web pages are full of the information we would need. We have "just" to mine and extract them.

Excurse: Sentimental Analysis Introduction

Some background (cont'd):

But what are the difficulties to find this information?

- Huge volume of data
- Unstructured data
- The information is often hidden in a certain blog
- Text and opinions are not straight-forward e.g. irony, reversing expressions

Excurse: Sentimental Analysis Introduction

Definition: Sentiment or opinion analysis is extracting subjective information out of data by the use of natural language processing or text mining.

Definition: Natural language processing is an area of computer science and artificial intelligence that deals with analyzing, understanding, generating nad interacting the language that humans use for interactions with computers in both written and spoken contexts using natural human languages instead of computer languages.

Definition: Text mining is analyzing text and gathering information out of it by using pattern analysis techniques.

Excurse: Sentimental Analysis Introduction

Example:

"(1) I bought an Samsung tablet last week. (2) It is a nice tablet (3) and has right and vividly colorful screen. (4) Its rail-thin design is comfortable and ultracompact. (5) The battery life is ok (6) but that's it (7) Given its average specs, its price feels too expensive.(8) The automated display optimizer did little more than needlessly adjust the screen's RGB level and brightness."

Excurse: Sentimental Analysis Introduction

Question: What do we want to extract and mine out of this review?

Answer: If the buyer is satisfied or not with the purchase.

Question: What do you notice?

Answer: There are several opinions in this review

Excurse: Sentimental Analysis Introduction

"(1) I bought an Samsung tablet last week. (2) It is a nice tablet (3) and has right and vividly colorful screen. (4) Its rail-thin design is comfortable and ultracompact. (5) The battery life is ok (6) but that's it. (7) Given its average specs, its price feels too expensive.(8) The automated display optimizer did little more than needlessly adjust the screen's RGB level and brightness."

neutral *positive* *negative*

Excurse: Sentimental Analysis Introduction

One have objects and targets like the “tablet” as a whole, the “screen”, “design”, “battery” and “price”.

Source is a review: “I”

There are opinions or emotions expressed like “It is a nice tablet” and “its price feels too expensive”

Excurse: Sentimental Analysis Introduction

How to formalize this a bit more?

We are following the formalization of *Sentiment Analysis and Subjectivity* by *Bing Liu*

Excurse: Sentimental Analysis Introduction

Definition (object): An object o is an entity which can be a product, person, event, organization, or topic.

It is associated with a pair, o: (T, A), where T is a hierarchy of components (or parts), sub-components, and so on, and A is a set of attributes (properties) of o. Each component has its own set of sub-components and attributes.

Definition: An *opinionated document d*, is a product review, a forum post or a blog that evaluates a set of objects. In the most general case, d consists of a sequence of sentences $d = \langle s_1, s_2, \ldots, s_m \rangle$.

Excurse: Sentimental Analysis Introduction

Definition (opinion passage on a feature): An opinion passage on a feature f of an object o evaluated in d is a group of consecutive sentences in d that expresses a positive or negative opinion on f.

Example: *(5) The battery life is ok (6) but that's it.*

Excurse: Sentimental Analysis Introduction

Definition (explicit and implicit feature): If a feature f or any of its synonyms appears in a sentence s, f is called an explicit feature in s. If neither f nor any of its synonyms appear in s but f is implied, then f is called an implicit feature in s.

Example:

Explicit: It is a nice tablet

Implicit: Given its average specs

Excurse: Sentimental Analysis Introduction

Definition (opinion holder): The holder of an opinion is the person or organization that expresses the opinion.

Definition (opinion): An opinion on a feature f is a positive or negative view, attitude, emotion or appraisal on f from an opinion holder.

Definition (opinion orientation): The orientation of an opinion on a feature f indicates whether the opinion is positive, negative or neutral.

Excurse: Sentimental Analysis Introduction

Model of an object: An object o is represented with a finite set of features, $F = \{f_1, f_2, \ldots, f_n\}$, which includes the object itself as a special feature.
Each feature $f_i \in F$ can be expressed with any one of a finite set of words or phrases $W_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$, which are synonyms of the feature, or indicated by any one of a finite set of feature indicators $I_i = \{i_{i1}, i_{i2}, \ldots, i_{iq}\}$ of the feature.

Excurse: Sentimental Analysis Introduction

Model of an opinionated document: A general opinionated document d contains opinions on a set of objects $\{o_1, o_2, \ldots, o_q\}$ from a set of opinion holders $\{h_1, h_2, \ldots, h_p\}$. The opinions on each object o_j are expressed on a subset F_j of features of o_j. An opinion can be any one of the following two types:

1. *Direct opinion:* A direct opinion is a quintuple $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, where o_j is an object, f_{jk} is a feature of the object o_j, oo_{ijkl} is the orientation or polarity of the opinion on feature f_{jk} of object o_j, h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i. The opinion orientation ooijkl can be positive, negative or neutral. For feature f_{jk} that opinion holder h_j comments on, he/she chooses a word or phrase from the corresponding synonym set W_{jk}, or a word or phrase from the corresponding feature indicator set I_{jk} to describe the feature, and then expresses a positive, negative or neutral opinion on the feature.

Excurse: Sentimental Analysis Introduction

2. *Comparative opinion:* A comparative opinion expresses a relation of similarities or differences between two or more objects, and/or object preferences of the opinion holder based on some of the shared features of the objects. A comparative opinion is usually expressed using the *comparative or superlative* form of an adjective or adverb, although not always.

Excurse: Sentimental Analysis Introduction

Objective of mining direct opinions: Given an opinionated document d,

1. discover all opinion quintuples (o_j, f_{jk}, oo_{ijkl}, h_i, t_l) in d, and

2. identify all the synonyms (W_{jk}) and feature indicators I_{jk} of each feature f_{jk} in d.

Excurse: Sentimental Analysis Introduction

Why have we develop this theoretical framework? (and not just run data analytics examples?

Sentiment analysis is very complex. Even if humans are receiving an e-mail there is often some misinterpretation of the meaning.

Thus, to implement and analyses text and mining out the opinions and feelings requires a sufficient de-composition of text structure under the constraints that text or word relationships are contained.

And this requires an understanding how to structure text for data analytics methods.

Excurse: Sentimental Analysis Introduction

Thus, we can start with methods: