S 10

Data Analytics for Data Scientists

Design of Experiments (DoE)

Suggested solutions for Exercise 10: Large Data Quantities

2024

Prof. Dr. Jürg Schwarz
MSc Adrian Bieri

Suggested solution 01

Properties of data / examples

There is a distinction between *made data* and *found data.*

Develop examples and explain your choices based on the distinction *made data* vs. *found data*. Estimate the size and complexity of the data set (e.g. number of variables, sub-populations).

Suggested answers to the questions

Made data	**Size / Complexity**
Data from the European Social Survey (ESS) (www.europeansocialsurvey.org) Partial data set for Switzerland from *Round 8* The ESS dataset is designed to answer research questions on topics such as " ... attitudes, beliefs and behavior patterns of diverse populations ...". The ESS data set serves this purpose exclusively.	The data set is quite small, with a sample size of about 1,500. There are no "hidden" sub-populations, all subgroups can be described with variables – for example, region, gender, etc. Although the dataset contains around 500 variables, these variables are neither nested nor exhibit complex interdependencies.

Found data	**Size / Complexity**
The Swiss population census produced two basic statistics with data from registers: ◦ Statistics on the population and households (STATPOP) and ◦ Statistics on buildings and housing (GWS). The register logs cover about 99% of the population and households: See also below under "**f large**"	The data set's size is about 99% of the size of the Swiss population → 0.99 x 9,000,000 = 8,910,000 Thus, the dataset is large relative to the ESS data but small compared to social media datasets, which can reach petabyte scale. The complexity is considered to be low, roughly comparable to that of the ESS data set.

Suggested solution 02

Statistical paradises and paradoxes in big data

When considering *administrative data sets*, it is necessary to take into account the proportion f.

$f = \frac{n}{N}$ Proportion of sample size of the *administrative dataset* in the population

Develop three typical examples that differ significantly in terms of their f proportion.

Suggested answers to the questions

Example for **very small f**

Tweets collected during a specific period cover only a small fraction of the X user base.
For example, when a sample is taken to investigate a shitstorm.
In this case, the f proportion is approximately 1%.

Example for **around medium f**

The registration of long-term unemployed individuals (LTU: Long-term unemployed) in the EU member states, for example 48.3% in Italy, captures about half of the long-term unemployed population. (ec.europa.eu).
In this case, the f proportion is approximately 50%.

Example for **very large f**

The Swiss population census, which has been conducted since 2011, is based on a register survey that evaluates existing administrative data. For this purpose, the Federal Statistical Office uses the cantonal and communal population registers, the federal population registers and the federal building and housing register.
The registers cover more than 99% of the resident population in Switzerland.
In this case, the f proportion is approximately 99%.