Methods For Statistical Data Analysis

CONF DR.CRISTINA SERBANESCU

AGENDA

I.Introduction to Statistical Methods of Data Analysis

II. Methods For Statistical Data Analysis

III.Course profile

IV.References

V.Examination

I.Introduction to Statistical Methods of Data Analysis

In the Information Age, data is no longer scarce – it's overpowering. The key is to sift through the overwhelming volume of data available to organizations and businesses and correctly interpret its implications. But to sort through all this information, you need the right statistical data analysis tools.

Methods of statistical processing

Statistics offers a range of methods, the choice of which will depend on four factors: (1) The type of variables: qualitative or quantitative; (2) The status of the variables: explanatory or dependent; (3) The number of variables: one, two or multiple and (4) the type of analysis: exploratory (descriptive) or confirmatory (inferential).

Scope and purpose

Data analysis is the process of developing answers to questions through the examination and interpretation of data. The basic steps in the analytic process consist of identifying issues, determining the availability of suitable data, deciding on which methods are appropriate for answering the questions of interest, applying the methods and evaluating, summarizing and communicating the results.

Analytical results underscore the usefulness of data sources by shedding light on relevant issues. Data analysis also plays a key role in data quality assessment by pointing to data quality problems in a given survey. Analysis can thus influence future improvements to the survey process.

II. Methods For Statistical Data Analysis

We suggest starting our data analysis efforts with the following five fundamentals – and learn to avoid their pitfalls – before advancing to more sophisticated techniques.

1. Mean

The arithmetic mean, more commonly known as "the average," is the sum of a list of numbers divided by the number of items on the list. The mean is useful in determining the overall trend of a data set or providing a rapid snapshot of your data. Another advantage of the mean is that it's very easy and quick to calculate.

Pitfall:

Taken alone, the mean is a dangerous tool. In some data sets, the mean is also closely related to the mode and the median (two other measurements near the average). However, in a data set with a high number of outliers or a skewed distribution, the mean simply doesn't provide the accuracy you need for a nuanced decision.

2. Standard Deviation

The standard deviation, often represented with the Greek letter sigma, is the measure of a spread of data around the mean. A high standard deviation signifies that data is spread more widely from the mean, where a low standard deviation signals that more data align with the mean. In a portfolio of data analysis methods, the standard deviation is useful for quickly determining dispersion of data points.

Pitfall:

Just like the mean, the standard deviation is deceptive if taken alone. For example, if the data have a very strange pattern such as a non-normal curve or a large amount of outliers, then the standard deviation won't give you all the information you need.

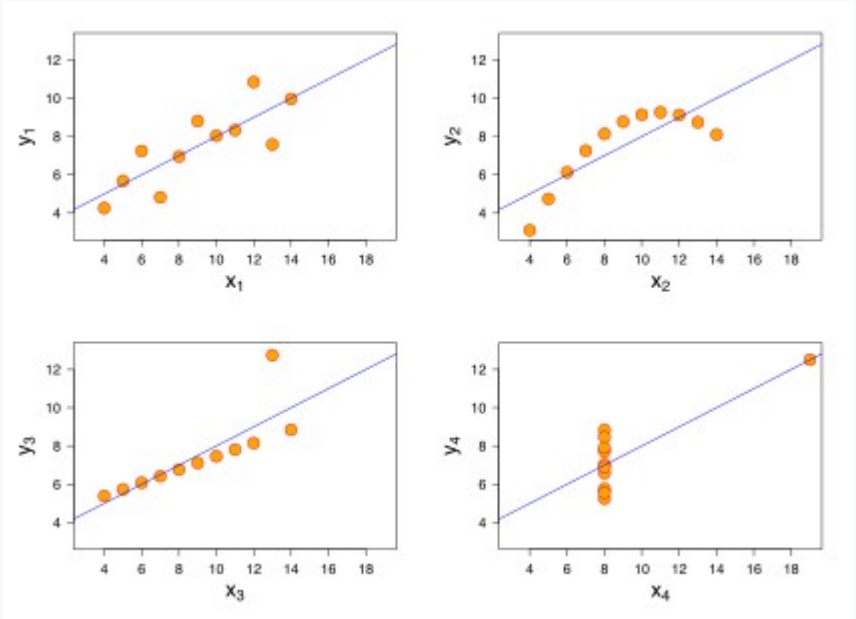
3. Regression

Regression models the relationships between dependent and explanatory variables, which are usually charted on a scatterplot. The regression line also designates whether those relationships are strong or weak. Regression is commonly taught in high school or college statistics courses with applications for science or business in determining trends over time.

Pitfall:

Regression is not very nuanced. Sometimes, the outliers on a scatterplot (and the reasons for them) matter significantly. For example, an outlying data point may represent the input from your most critical supplier or your highest selling product. The nature of a regression line, however, tempts you to ignore these outliers. As an illustration, examine a picture of **Anscombe's quartet** , in which the data sets have the exact same regression line but include widely different data points.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and

4. Sample Size Determination

When measuring a large data set or population, like a workforce, you don't always need to collect information from every member of that population – a sample does the job just as well. The trick is to determine the right size for a sample to be accurate. Using proportion and standard deviation methods, you are able to accurately determine the right sample size you need to make your data collection statistically significant.

Pitfall:

When studying a new, untested variable in a population, your proportion equations might need to rely on certain assumptions. However, these assumptions might be completely inaccurate. This error is then passed along to your sample size determination and then onto the rest of your statistical data analysis

5. Hypothesis Testing

Also commonly called *t* testing, hypothesis testing assesses if a certain premise is actually true for your data set or population. In data analysis and statistics, you consider the result of a hypothesis test *statistically significant* if the results couldn't have happened by random chance. Hypothesis tests are used in everything from science and research to business and economic

Pitfall:

To be rigorous, hypothesis tests need to watch out for common errors. For example, the placebo effect occurs when participants falsely expect a certain result and then perceive (or actually attain) that result. Another common error is the Hawthorne effect (or observer effect), which happens when participants skew results because they know they are being studied.

III.Course profile

1. Data and Probability Models

2. Parameters and “Statistics”

3. Bayesian Models

3. Statistical Inference as a Decision Problem

4. Prediction

5.Exponential Families of Probability Models

6.Hypothesis Testing and Confidence Regions

7.Neyman-Pearson Lemma

8.”Most Powerful” Confidence Tests

9.Confidence Bounds

10.Confidence Intervals/Regions

11.Gaussian Linear Models

12.Test Pearson’s Chi-squared Test (Discrete Models)

IV.References

1.https://www.bigskyassociates.com/blog/bid/356764/5-Most-Important-Methods-For-Statistical-Data-Analysis

2.James L. Hutter Department of Political Science, Department of Statistics , Iowa State University
„ INTRODUCTION TO STATISTICAL DATA PROCESSING” 2018

3. https://ocw.mit.edu/courses/mathematics/18-655-mathematical-statistics-spring-2016/lecture-notes/

V.Examination

– Homework project (theory & practice, with a deadline before 6 january 2021);

– Application project you can work in a team of max. 4 members;

– Seminar activity;.

– Final mark: H+A=80%,S=20%