Information Analysis Assignment 1 Janna Leguijt studentnummer 11268638: Statistics

What is statistical Inference?

Statistical Inference is the process of drawing conclusions from a sample of the population.
For example: On the base of a sample we try to predict how people will vote in the Dutch elections on March the 15.
Statistical Inference is based on the following assumptions:

1. A sample is likely to be a good representation of the population.
2. There is an element of uncertainty at how well the sample represents the population. Usually a margin of error is used , the so called confidence intervals.
3. The way the sample is taken matters. Is it representative? This means that every element in the population has the same change to be selected. In the case of humans this is difficult. We can see that in the example of the elections. The past few years has shown us that it is very difficult to predict the outcome of elections. Somehow the researchers missed a part of the population. For example people who don't want to answer questions from a researcher. In the case of humans (and elections) it is also possible that they don't tell the truth, but give a more social accepted answer or change their mind. There are limitations to statistical inference. There is always a factor of uncertainty. Therefore we have to discover if our data are statistical significant.

Before I discuss this, I will go back and talk about the role of statistics in research. Because using statistics is not a goal on its own. It is part of a research process. This process always starts with a question. For example: "How many people are using archives in the Netherlands and what kind of people are they? " The next step will be to check the literature to see if your question is not already answered. The literature also helps you to formulate your original question better. More exact questions could be "How many visits were there on websites of Dutch archives in 2016? Is this a growing number since 2006? What kind of interests have the people who visit the websites of Dutch archives? What is the level of education of these people? What is their age?"

If I want to answer these questions I have to make several steps. To find out how many times the websites are visited I have to ask all the archives to send me their results from google analytics (I am aware that working with google analytics has it flows, but for my goal it is sufficient). Then I know what the population is. This is a form of descriptive statistics, which is not further discussed in this paper, which deals with probabilistic statistics. If I want to compare the data of 2016 with the data of 2006 I have to collect data from older researches.

If I want to investigate the last three questions I can conduct this research by interviewing people, a qualitative approach, or I can choose a more quantitative approach, or combine both methods. At this point probabilistic statistics is starting to play a role. I will try to explain what I have to consider when I use statistics to answer my research questions.

First I have to formulate a hypothese : 70 % of the visitors of websites of Dutch archives are interested in genealogy. My O-hypothese is that 30% of the visitors has another interest. This means that I have made two categories for my research: genealogist or non- genealogist. This type of question is typical categorial. The population is the visitors of sites of Dutch Archives in 2016. With Dutch Archives I mean the public archives . You can see that I have to define the terms I use. I have to decide how big my sample must be. As method to collect my data I can make a short questionnaire and ask all the archives to put it on their website during one month. This means that everyone who visits a site of a Dutch archive has the same change of answering the question. I have to run a test to see if my I can reject my 0-hypothese.

In the questionnaire I have also put questions about education and age so I can also try to answer my other two questions about education and age. I want to know if there is a relation between the level of education and the interest in genealogy. Beforehand I have to make categories of education (low – middle – high). I can present my data in a scatter graph to establish if there is a positive correlation between the level of education and interest in genealogy. I can work out the correlation coefficient and this has to be o<1.

In the case of the ages of the genealogists I also have to formulate a hypothese etcetera. This are typical numerical data. In this case the results will probably have the form of a normal distribution. This means that I can use the STN (Z) test to calculate the probabilities of my results.

Although I didn't describe my research above completely , you can see that I have to consider several things when I am using statistics during my research:

1. What is the purpose of the test (to establish if my hypothese can be rejected or not, to compare data, to find a relationship?)
2. What is my population?
3. What kind of data do I have to collect? This can be categorical (genealogist or non-genealogist), or quantitative (the numbers of visitors, age, prices).
4. How I am taking a sample? By what method? How big? What kind of categories I am going to use and define? Must I divide my data in classes(for example low – middle – high education)?
5. What kind of test must I run to establish whether:
   - My data are representative for the population?
   - I can reject my hypothesis or not?
   - I can compare results?
   - Find out if there is a relation between the data or not?
   - My results are statistical significant and for what level?
6. How do I present my data (diagram, scatter graph etc)?

A key concept is the question whether my results are statistical significant or not. This means that I can reject my H0 within certain boundaries. (95 %, 99% certainty, the level of significance). I have to be aware of the factors which make my research more powerful: the size of the sample, the differences between groups, the preciseness of my data, my standard deviation (the smaller the better). And of course I can also make an error en falsely reject my HO.

In the end we have to be aware that using statistics is not a brainless trick. We cannot "prove" anything with it. It is the task of the researcher to make the results also interpretive significant and give a good explanation.