Statistics: A way to support or undermine one's arguments?

Miriam Seel
Nagoya University

Writing is essential when creating a thesis or article. While writing, theories are connected, the previous research is introduced, and in the end, one's own hypotheses or assumptions are presented. In order to test these hypotheses, many areas use data, either taken from experiments or surveys for instance. These data are then analyzed with the help of statistics, and based on the results, conclusions are drawn. During that process, students as well as researchers should rely on their knowledge of statistics to avoid potential traps and fallacies based on which wrong outcomes might result. However, often, mistakes happen as human reasoning is far from perfect. In order to better understand these fallacies and other problems associated with statistics, this article will describe issues of data presentation, data analysis as well as data interpretation, including the concepts behind statistics.

1. Introduction

In academia, writing is essential for conveying one's findings in a thesis or an article. During this process, various theories are presented and connected, previous research is analyzed, and in the end, the researcher or student examines his or her own hypotheses. In several scientific areas, hypotheses are tested with the help of (empirical) data, taken from experiments or surveys, for instance. This is where things often get complicated. Even though a lot of researchers and students are good at looking up theories, describing them and using them as a basis for their theoretical arguments, analyzing data is done rather sloppily. In some cases, knowledge about statistics is limited, leading to mistakes or faulty interpretations. But what is a good theory without even better (empirical) data to prove it?

In this article, I aim to give the readers a short idea about common problems associated with statistics, in particular with the so-called null hypothesis significance testing (NHST). Note that this article is neither a guide nor a complete overview. Instead, those not so familiar with statistics might catch a first glimpse of what not to do, and those who want to "dig deeper" are welcomed to refer to further readings about statistics.

Section two will shortly introduce the concept of statistics. The next section will deal with probability and the difficulty to understand "chance". Then, the following three sections will discuss common problems concerning data analysis, data presentation, and data interpretation, leading to a short summary in the conclusion.

2. Statistics or "What is this all about"

As with every keyword, one will find plenty of definitions of what *statistics* is. Citing Wikipedia, "Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data" (2015). In other words: everything connected to data. In fact, statistics can be found not only in sciences but also in everyday life: Whether we take the train or bus depends on how long each of them takes to get to our goal on average; whether we choose a certain hotel or airline depends on their ratings; and even how long we boil our eggs in order to have them hard but not too hard – all of these decisions depend on numbers, on statistics.

NU Ideas Volume 4.2. 2015. Nagoya University Institute of Liberal Arts & Sciences

Similarly, statistics are used in scientific fields as well to come to conclusions, and below I will briefly describe the processes involved.

At least in the area of psychology, we distinguish between two types of statistics: descriptive statistics and inferential statistics. Descriptive statistics, as the name implies, are used to describe samples and distributions. They include means, standard deviations, ranges and many more, and should help to give a first idea of what a sample or distribution looks like. Inferential statistics, on the other hand, are used for testing hypotheses. Here, we rely on *t*-tests, ANOVAS, or regression models to find out whether there are, for instance, differences between groups or relationships between variables. A probability (or *p*-value) is calculated when using one of these tests to decide on our hypothesis. If this value is lower than a certain threshold, we reject the null hypothesis (e.g., there is no difference between two groups) and assume the alternative hypothesis (e.g., two groups differ significantly from each other). Although this process might sound simple, in fact, it is complicated and, first of all, assumes that the student or researcher does already have a hypothesis before conducting any tests. What is more difficult, however, is the concept of probability. The next section explains why humans sometimes have problems with "chance".

3. Probability

Many events are based on chance or probability. For instance, imagine going to a casino to play roulette. The last 20 rounds, only red numbers came up, and now it is our turn to decide what number to take. Which one would you choose – a red or a black number? (To make it simple, let us ignore the green zero.) More than a few people would choose black, reasoning that it should finally be the turn for a black number to be chosen. However, these events are unrelated – chances are, at every turn, equal for both kinds of numbers. Thus, ignoring probabilities may result in misjudgments. In reality, whether we choose a red or black number does not matter, since in every turn the probability for either red or black is 50%.

This example from everyday life may not seem to have serious consequences (except that we may lose a lot of money), but there are some instances where life depends on a right decision. Eddy (1982) conducted a study in which he asked physicians to calculate the probability that a patient has breast cancer, giving following information to his subjects. In the population of 40-year-old women, 1% of all women has breast cancer. A mammography correctly detects that a women has breast cancer with a probability of 80%, but 10% of healthy women are falsely diagnosed with breast cancer. How high is the probability that a woman in her forties who received a positive diagnosis has breast cancer?

Let us try to calculate the probability by ourselves. The actual probability that a woman with a positive diagnosis has breast cancer is approximately 7.7%, as can be calculated using the Theorem of Bayes. Yet, nearly all of the physicians Eddy (1982) surveyed chose 75%, a number almost ten times higher. Eddy concluded that even doctors who have to judge illnesses based on their incident rates had problems doing so, as they were not able to interpret statistics properly. If even specialists make mistakes, how should novices be able to judge correctly based on probabilities? Although this finding seems alarming at first glance, there is hope: Other researchers (Gigerenzer and Hoffrage 1995) tried to replicate Eddy's findings by describing the incident rates in actual numbers instead of probabilities. As a result, participants were more likely to calculate the probabilities properly. Why? One explanation states that actual numbers are closer to how we use probabilities in real life. Simple percentages, on the other hand, are more abstract and therefore more difficult to work with.

Thus, although with the help of statistics we decide on certain events, there is always uncertainty involved. Yet, it is worse if people fail to understand how chance works. More training, in particular by replacing probabilities with actual numbers, is needed, so that

researchers as well as students are able to work with probabilities. Leaving chance behind, I will present more concrete problems regarding data analysis, data presentation, and data interpretation in the following sections.

4. Data analysis

Now that we have learned a little bit about probabilities, what could possibly go wrong? According to Murphy's Law, everything, but let us start with one crucial aspect of data analysis - choosing the right method. There is a variety of measures for descriptive as well as inferential statistics which students and researchers can chose from. Of course, the measures selected should fit to one's purpose and hypothesis, but often preconditions are ignored or simply not known. As I cannot describe all measures here, I will just give a few examples to illustrate the problem.

Students are often judged with the help of grades. In Germany grades range from 1 (very good) to 6 (very bad). Teachers like to calculate averages; for instance, the average for a mathematics test across all students was 2.4. Even though it is possible from a mathematical point of view to calculate the average, it is actually not an appropriate measure of central tendency, as psychologist (among others) would say. Grades measured in this way count as an ordinal variable; there is an order, ranging from the highest (1) to the lowest (6). For ordinal variables, you can only calculate the median, the value which divides your distribution into two halves so that 50% of all values are in each half. In order to calculate the average, any variable should be at least metric – there is supposed to be a "0" and the differences between numbers should be of same size. However, if "1" equals very good and "5" indicates that a student has failed, the interval between 4 and 5 is a bigger difference than that between 1 and 2. Besides, there is no absolute zero. Thus, from a statistical point of view, this calculation of average grade is invalid, although in reality very often it is found exactly as described here.

Maybe you still wonder why this is wrong, so let me give you another example. Biological sex is a nominal variable, with a person being either male or female (exceptions are not considered here for ease of understanding). Often, this variable is coded into "1" for male and "2" for female in a computer program. You could now calculate the average, the median and the range, but is this useful? Does it make sense? Certainly not. Yet, it happens that students or researchers (accidently) rely on descriptive statistics that are not suitable for their variable, in particular since the program will not complain about doing so. It will simply calculate. The user is the one who should know what he or she is doing.

Let us now have a look at inferential statistics. Unfortunately, the picture becomes even more complex as there are a lot of measures or methods that can be used. All of them have, of course, preconditions which should be checked before making use of them. Yet, I claim that a considerable number of researchers do not check these preconditions and just rely on the most common (or easiest) methods. For instance, the so-called *t*-test is a popular measure to test whether the means of two groups are statistically different from each other. In order to use that test, the variable needs to be metric, be normally distributed, have sample size smaller than n=30 and so on. When reading papers, then, we are often left in the dark whether these preconditions were met or not. Sometimes sample sizes are extremely small, with only eight subjects in one group. Very small groups are more likely to be influenced by extreme values; this is not good if an extreme value contradicts the hypothesis. The software for analyzing data will again not tell you that. Still running a test even though you know that you are violating the preconditions might have bad outcomes. One might think that bad outcomes are still better than no outcomes at all, but if those outcomes are then invalid, cannot be replicated, or lead to totally wrong conclusions, no outcomes might be indeed the better choice as you can at least start all over. Instead of then using parametric tests such as the *t*-

test, non-parametric tests should be the measures of choice, even though they might lack power.

In fact, the issue of power, that is, the ability of any test to detect a difference if there is one, is often ignored. Small sample sizes suffer from little power, meaning that in the worst case, the test will not show any significant difference even if there is a difference. That problem is avoidable by calculating power, sample size and the like in advance with software such as G-Power (Faul et al. 2007). Large samples, on the other hand, might become problematic, too, if tests such as the Kolmogorov-Smirnoff-test are chosen that are sensitive to sample size and instantly show a significant result. In any case, consider your sample size and think about whether it matches your needs and expectations.

All the issues illustrated here can strongly undermine the author's argument and ultimately invalidate the results. The good thing is, these issues can be avoided. If you are unsure, just grab some thorough books on statistics and ensure you know the preconditions for your statistical measures.

5. Data presentation

In general, all researchers dealing with statistics will need to decide what data to present. Normally, it is impossible to show all of the data to the audience as the flood of information will most likely overwhelm listeners. Thus, data need to be selected, and mostly those are chosen which support one's arguments. This is, of course, not wrong. It may become, however, misleading if data are not explained properly or if they are shown out of context so that readers have difficulties making sense out of them.

To illustrate this problem, let me give an example. In their news broadcast from October 18, 2014, the Japanese TV channel NHK explained that in those areas affected by the March 2011 tsunami, students in their 5th year of elementary school throw a ball a smaller distance compared to tests before the disaster. In the video clip, it was suggested that this trend is due to the fact that in those areas school sports fields had been changed to temporary housing areas for citizens who lost their houses in the disaster. Thus, students are not able to play outside anymore, which, in turn, leads to reduced physical abilities, including ball-throwing.

	2010 (mean)	**2013 (mean)**	**Difference**
Iwate	26.85	23.93	-2.92
Miyagi	25.18	23.56	-.162
Fukushima	25.32	22.45	-2.87
Tokyo	24.07	22.62	-1.45
Aichi	24.94	22.68	-2.26
Wakayama	26.12	23.64	-2.48

Table 1. Ball-throwing (in meters) in the prefectures affected by the March 2011 tsunami (Iwate, Miyagi, and Fukushima) compared to three prefectures which were not affected (Tokyo, Aichi, and Wakayama). (MEXT 2015)

However, looking carefully at the data taken from the homepage of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT 2015), it becomes obvious that this is not the whole story. As can be seen in Table 1, other prefectures which were not affected by the disaster actually suffer from the same problem. Comparing the data from various years, it seems that the decline in physical abilities among students is a general trend

across Japan. Yet, without these additional data, viewers or readers are misled to think that this decline is due to the disappearance of the sport fields, as the data presented in the news supports this argumentation. Instead of only presenting data taken from the areas affected by the tsunami without comparing them to data from other prefectures, it might be more meaningful to compare data within these prefectures. That is, for instance to compare schools in Iwate prefecture which can use outdoor sports fields with schools in the prefecture which cannot. By doing so, the argument that a lack of sport fields may lead to a decline in physical abilities can be evaluated within each prefecture and may give a first idea whether this hypothesis is supported. Note, however, that in this case randomization (i.e., assigning schools into two groups with and without outdoor sport fields by chance) is impossible and thus results might be distorted in any case.

In a thesis or journal article as well as in an oral presentation, data are typically shown in the form of graphs and/or tables. These graphs or tables aim to summarize the most important findings. Yet, sometimes people invest more time in choosing colors than in conveying their message. For instance, one common mistake is to show a bar chart that either does not start with "0" or in which the scale is distorted, making differences between bars seem larger than they really are. See Figure 1 and Figure 2 for a concrete example, using exactly the same data.

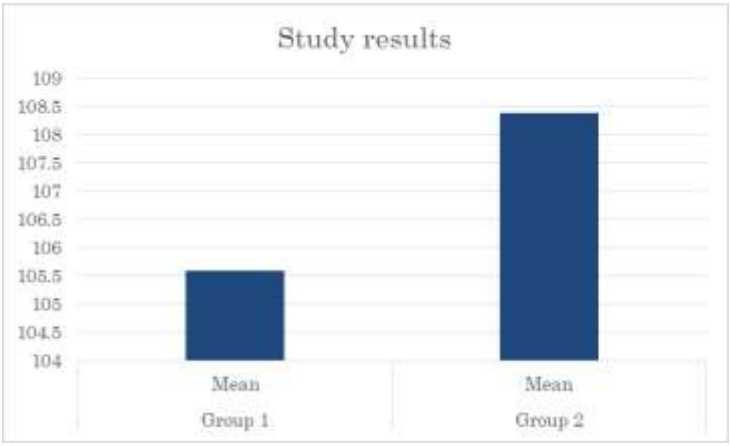

Figure 1. Scale does not start with 0. No exact values are given.

In Figure 1, the means of two groups are presented. However, the scale neither starts with "0" nor are the exact values given. Besides, the scale is distorted, making it seem that the difference between the two groups is actually larger than it really is. In Figure 2, the same two

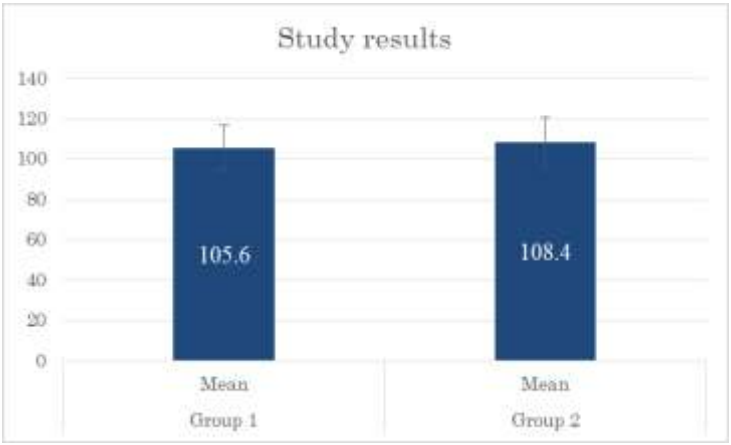

Figure 2. Same data as Figure 1. Values start with 0. Exact numbers given. Error bars represent standard deviation.

means are presented. Here the values start with "0", exact numbers are given, and error bars representing the standard deviation are included. By adding these elements, it should be easier to understand the graph, even without a thorough description. In some cases means are presented without the actual number, so that readers have to guess by themselves what the precise mean is. In addition, it is advisable to include error bars, either representing standard deviations or standard errors, in order to make a bar chart more easily understandable.

All of these examples apply to bar charts. In the end, what chart is most suitable needs to be determined by the presenter. It is thus recommended to realize what the most important message is: Do I want to present a mean? a distribution as a whole? or just percentages of limited groups? The easier the chart or graph is to understand, the more likely the audience will be able follow the presenter's argumentation and – hopefully – remember the findings.

6. Data interpretation

Another issue that is sometimes ignored is the fact that significance does not equal relevance. Assume that you would like to find out whether men and women differ in their IQ scores, a research topic that has been investigated with varying results (Neisser et al. 1996, 91-92). As a part of this study, 1000 men as well as 1000 women take a standard IQ test. The analysis shows that indeed, there is a significant difference between these two groups. However, a closer look at its size reveals that it is only a 0.5 point difference. Is half a point relevant? Most likely not, and this is why recently researchers are advised to present the so-called effect size along with their results (Cohen, 1988). Furthermore, results that are not significant might be relevant as well since for some diseases, for instance, only a few subjects can be recruited to take part in a study. Due to the resulting lack of statistical power, then, differences may fail to become statistically significant, even though the effect is practically relevant. Thus, it is crucial to know the difference between statistical significance – that is, what we are testing with the help of *t*-tests, correlations and so on – and practical relevance: the actual effect size compared to other effect sizes from the same area of interest. Even though not significant, a finding may be important for future research with larger statistical power to confirm the results. Not publishing or discussing so-called null findings, findings of no significant effect, might hamper research that is actually meaningful.

Besides the debate concerning significance versus relevance, it might happen that researchers wrongly interpret their data. Let me take Messerli (2012) as an example. In his study, Messerli found a strong positive and significant correlation between chocolate consumption and Nobel laureates in certain countries. In his conclusion, Messerli cautions that this relationship cannot be interpreted as causation, but reasons that chocolate consumption has been found to enhance cognitive abilities and thus might indeed help to win a Nobel Prize. He concludes, however, that this hypothesis has to be tested with the help of experiments. Note that the first part is correct: Whether chocolate consumption is responsible for Nobel prizes or the other way around remains unclear. Maybe researchers, who are busy with their work and therefore stressed, like to eat more chocolate for its positive effects? The last part, the need for experiments, is important, yet in this particular area it is impossible. There is no chance to randomly assign researchers to chocolate and non-chocolate eaters and then to see who wins more Nobel prizes. Messerli's argument that chocolate may enhance cognitive functions is an exaggeration, as chocolate in small doses does not have such a big effect.

In fact, Messerli's finding can be explained by the influence of a third variable, as Maurage and colleagues (2013) have demonstrated. They investigated whether an unknown variable was responsible for the correlation between chocolate consumption and Nobel laureates. They

found a strong correlation between the GDP of a country and the number of Nobel laureates, as well as correlation between GDP and consumption of chocolate. Therefore, it is not the chocolate itself, but rather the GDP that influences Nobel laureates and chocolate consumption. Why? It can be assumed that richer countries have more citizens that can afford and also like to eat chocolate. Besides, these countries also invest more money in education and research. This, in turn, leads to a seemingly strong correlation between unrelated variables. However, interpreting a correlation as cause and effect implies that we know which one is the cause and which one the effect, and this, in reality, is seldom clear. For this purpose, random experiments or very well-planned long-term studies are needed that help to gain insight into often complex relations between variables. Messerli's (2012) findings, although interesting, can be easily explained otherwise and thus, his argument is invalid and should not be cited to increase chocolate consumption in schools or universities. In general, defining cause and effect is a difficult task and thus faulty conclusions are unfortunately common.

7. Conclusion

In the process of making sense out of one's data, various potential traps as well as mistakes may lead to wrong results. Students as well as researchers are advised to consider these problems when working with data, and readers might want to look carefully at charts and results presented. Although statistics is a powerful tool, if misused it might undermine the arguments presented rather than support them. As mentioned above, recently alternatives for the traditional null hypothesis significance testing (NHST) have been presented (Giner, Leech, and Morgan 2002) that may circumvent some problems described in this article. However, as NHST is still the dominant way for testing hypotheses, it is crucial to at least have these problems in mind. In the end, in order to avoid mistakes, a sound statistical education is needed as well as regular "trainings" to not forget what one has learned and to keep up-to-date with developments in this particular area. Always remember: Without empirical data to prove it, a good theory is not worth that much. Do not stop critical thinking when it comes to literature review; also apply it to your own data.

References

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Erlbaum.

Eddy, David M. 1982. "Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities." In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment Under Uncertainty: Heuristics and Biases*, 249-267. Cambridge: Cambridge University Press.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." Behavior Research Methods 39: 175-191.

Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning without Instruction: Frequency Formats." Psychological Review 102: 684-704.

Gliner, Jeffrey A., Nancy L. Leech, and George A. Morgan. 2002. "Problems with Null Hypothesis Significance Testing (NHST): What do the Textbooks Say?" The Journal of Experimental Education 71: 83-92.

Maurage, Pierre, Alexandre Heeren, and Mauro Pesenti. 2013. "Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies." The Journal of Nutrition 143: 931-933.

Messerli, Franz H. 2012. “Chocolate Consumption, Cognitive Function, and Nobel Laureates.” New England Journal of Medicine 367: 1562–1564.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2015. “Zenkoku Tairyoku/Undo Noryoku, Undo Shukan Nado Chosa” [Results Concerning School Children’s Physical Abilities]. http://www.mext.go.jp/a_menu/sports/kodomo/zencyo/1266482.htm

Neisser, Ulric, Gwyneth Boodoo, Thomas L. Bouchard Jr, A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert Perloff, Robert J. Sternberg, and Susana Urbina. 1996. “Intelligence: Knowns and Unknowns.” American Psychologist 51: 77-101.

Wikipedia. “Statistics.” April 26, 2015. http://en.wikipedia.org/wiki/Wikipedia:Statistics