Avoiding Statistical Mistakes

Nora Strasser, (E-mail: Strasser@friends.edu), Friends University

ABSTRACT

Avoiding statistical mistakes is important for educators at all levels. Basic concepts will help you to avoid making mistakes using statistics and to look at data with a critical eye. Statistical data is used at educational institutions for many purposes. It can be used to support budget requests, changes in educational philosophy, changes to programs, and addition of new programs. Therefore, it is important to be sure that decisions based on statistical data are valid. Also, it is important to be certain that the data has been presented in a fair and unbiased way. Otherwise, the changes and decisions may not reflect the best interests of the educational institution or its constituents. Many faculty, staff, and administrators at colleges and universities use statistics to support their findings. Most conclusions are based on the data collected and the analysis. If either the data collected or the analysis is faulty, the conclusions may be invalid. Unfortunately, misuses and abuses are common. This paper attempts to present basic concepts that will help you to avoid making mistakes when using statistics and to look at data with a critical eye. No extensive knowledge of mathematics or statistics is needed to be able to judge the validity of the data. The following questions will be answered: How can administrators avoid making invalid decisions when presented with statistical data? And, how can researchers avoid making mistakes that undermine the validity of their data? First, we will look at sources of bias. Experiments and the collection of data can be influenced by many different sources of bias. These sources include the population used for the study, the sample size and method of selection, the funding of the experiment, statistical assumptions, the publication source, and the biases of the researcher. Examples from within education will be used to gain an understanding of the biases that can occur. Second, we will look at the methods used to analyze and display the statistical data. This can include descriptive statistics and more advanced analysis as well. The characteristics of good graphic presentations of data will be compared to those that are flawed. Examples of graphs that present data fairly will be shown as well as graphs that distort the data. The differences between ordinal and ratio data types will be emphasized. Appropriate uses of ordinal data will be discussed. Also for ratio data a discussion of the meaning of average and importance of measures of variance will be examined. We will also look at the crafting of surveys and the choosing of samples. Finally, we will look at how the results can be interpreted. Valid data and analysis can still result in incorrect interpretation. We will examine examples of cases where misinterpretation is common and results in invalid conclusions. It is important to realize that statistical data and analysis is subject to many forms of error and misinterpretation. All consumers of statistical data need to look carefully at the data and the analysis. Numbers do lie! However, the use of statistics and data is very important in making decisions. By using statistics correctly, we can significantly impact education for the better. This goal can be achieved by those who are knowledgeable about statistics and its use. This paper presents the tools for using statistics correctly and identifying problems with statistical data.

INTRODUCTION

Data is a powerful tool that can be used or misused in various ways. Many people tend to believe anything that is supported by statistical analysis. Since they do not understand the data or the analysis, they assume that it is true. In fact, math classes experienced by most of the population were ruled by the "correct answer". This has led many people to believe that numbers presented in statistical analysis must be right! Unfortunately statistics is ruled by subtleties and various interpretations of data can be both right and wrong. Even the title of books such as "How to lie with Statistics" can imply that statistics can be manipulated.

Data can be used to mislead intentionally, but it can also be used in such a way that the mistakes are unintentional. Even the most scrupulous researcher can make mistakes that render the data invalid. However, a few basic concepts will help to avoid and recognize the statistical misuses.

We will consider three main categories of statistical error. First, we will look at sources of bias. Second we will consider methods of displaying and analyzing data. Finally we will look at the interpretation of the results.

BIAS

Data analysis and statistical results can be significantly influenced by many things. These things can include the population of the study, the sample size and how it was selected, the funding of the study, the biases of the researcher, and the statistical assumptions.

For example, a researcher wants to answer the question: "At my institution, are the students satisfied with the food service?" How does the researcher define the population of interest? The researcher might define it as "the students who eat in the cafeteria" and then decides to choose a sample by selecting random students leaving the cafeteria to fill out a survey. However, this definition and sampling method demonstrates a bias of the researcher. It was assumed that only students who actually eat at the cafeteria should be considered. When the reason many students do not eat there might be because of the quality of the food. The population should be defined as "all students at my institution".

In the previous example, the biases represented included the definition of the population, the methods for obtaining a sample, and the researcher's biases. The researcher showed his bias by defining the population as he did.

The sampling method can also demonstrate bias in many ways. It is easy to introduce bias into a sampling method. In fact, it is almost impossible not to. An unbiased sample would have to represent a fully random subset of the correct population. Fully random means that every person in the population has an equal chance of being chosen and all those who are picked agree to be surveyed.

Mailing a survey results in a large non-response rate (bias). Using a telephone interview results in a sample that may not represent the population since some people may not have phones (bias). Using in-person interviews are difficult to organize and you will still have the non-response rate (bias). The researcher, even one with the best intentions, usually must compromise the randomness of her sample. So any sampling method must be considered to be a source of possible bias and should be reported as such.

Another example is of a researcher conducting a study to determine if there is enough interest to offer a new major. The researcher decides to survey the students taking an introductory course in the subject. The researcher reports that the 60% of the students are very interested in the major and 40% of the students are interested in the major. The researcher in question, not only teaches this introductory course, but would be the primary faculty member for this new major.

The biases here are many and no decision should be based on the results. In fact, the results shouldn't even be considered in the decision making process. First of all, the population is not defined properly. The sample is highly suspect and the reporting of the results is most likely biased as well. You might think that this is an extreme example, but I have seen this type of thing happen many times. At my institution, we have a non-traditional Associates degree program. To decide whether or not to offer a new Bachelor's degree in a similar format, a couple of courses taught by the researcher were surveyed to determine interest in the program. This researcher again would be the principle faculty member for the new program and the courses were introductory courses in the subject. Based on a similar analysis, the program was started. The program was a success, but it was due more to luck than to adequate research. In fact, hearsay would probably be as accurate as the statistics generated.

One more example is a common one that most people are exposed to if they teach in a college or university. It has to do with student evaluations. First student evaluations are given to a class that is small in size, say nine students. Those nine students could be looked at as a sample of students who will take that same class over time. However, this sample is not random and it is very small. The assumption is that the responses from this class represent a normal population (One were the values cluster around the mean). This is a form of bias. Not only is the size of the sample far too small, we don't know that it represents a normal population at all.

When looking at research, it is always important to ask yourself several questions: "What is the population?", "How was the sample chosen and is it of adequate size?", "How is the researcher related to the outcome of the project?", and "What statistical assumptions were used?" If the answers to all of these questions are satisfactory, then you can assume that the results of the research are relatively unbiased and it will be valid to use these results to make decisions.

METHODS OF DISPLAYING AND ANALYZING DATA

The way data is presented can also be misleading. From the graphs to the descriptive statistics used, the presentation of the data can add to the understanding or detract from it.

One of the most common mistakes made when analyzing data is to use the wrong statistical tests. This is especially true when analyzing survey data. Most survey data is based on a Likert scale (those that have you mark categories such as strongly agree, agree, neutral, disagree, strongly disagree, etc.). Most statisticians consider the data obtained from a Likert scale to be ordinal, meaning that it has an inherent order, but it is not continuous data. Since the data is not continuous, the numbers associated with the responses do not necessarily occur in proportion to each other. For example, a two (disagree) is not necessarily half as good as a four (agree). In fact the numbers are arbitrary and really only represent the order. Therefore, it doesn't make sense to calculate a mean. Means should only be used for ratio data.

This is an error that is pervasive in almost every aspect of education. There is some argument that a mean is acceptable in the Social Sciences, but to be sure that your analysis actually has meaning, it is safer to use other statistical measures. You should use a median or a mode (not a mean) as a measure of central tendency. Again the variability should be expressed in terms of a range (not a standard deviation). Any graphs should be bar charts (Gaps between the bars) not histograms (no gaps between the bars).

One very common example that typically uses Likert data inappropriately is the student evaluations most instructors have to give out to their classes. It isn't the evaluations themselves that are suspect, but the way they are typically tallied and the results reported. It is very likely that a mean will be reported for each question and then those means are used to find a number that is your overall score. This is an inappropriate use of Likert data. The number generated is highly suspect and should never be used for anything. Unfortunately, this number is typically used by administrators to gauge the instructor's ability to teach. Decisions about tenure, promotion, salary increases, and terminations are based on it.

One final problem with this blatant misuse of statistics is that when the problem is pointed out to the researcher and/or the users of the data, it is dismissed by them as minor. Just because this mistake is made a lot, doesn't mean that it is insignificant. From a mathematical standpoint, it is a mistake equal to adding two and two and getting five. In this example, we know the answer is close, but it is still wrong. When we use a mean with Likert data, we don't even know if the answer is close.

This is also compounded by small data sets. Let's say that we have a typical question given during a course evaluation: The instructor explains clearly. 5 – strongly agree, 4 – agree, 3- neutral, 2- disagree, 1 – strongly disagree. A class with six students answers the question in the following way: Three students strongly agree, two students agree, and one student strongly disagrees. So the values are 5, 5, 5, 4, 4, 1. If you calculate a mean you get 4. Does this make sense? Doesn't it seem that the one student who is dissatisfied is given enormous power? The median of

the data set is 4.5 and the mode is 5. The best representation of the data is provided by the median. It more fairly distributes the scores based on the students' answers.

Let's consider the graphs used to represent data. In Figure 1, the graph is distorted so that the differences in the bar heights are emphasized. This graph was created in Excel and it automatically set up the bars this way. All graphs should start at 0 so as to not emphasize the slight differences in the bar heights. Figure 2 shows the same graph in an unbiased way. Notice that the first graph looks like the number of majors is steadily increasing, but the second graph seems to indicate that the number of majors is in a steady state pattern.

Figure 1
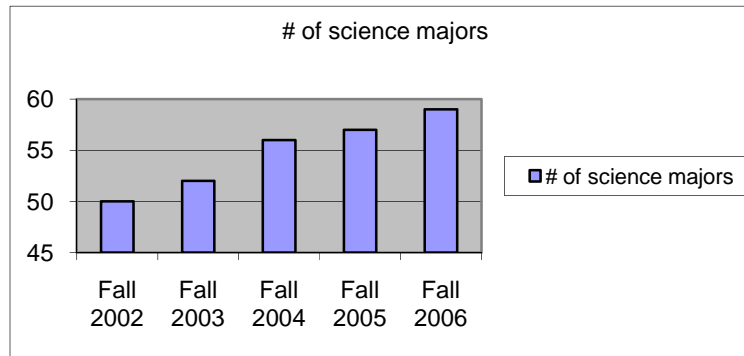
Figure 2
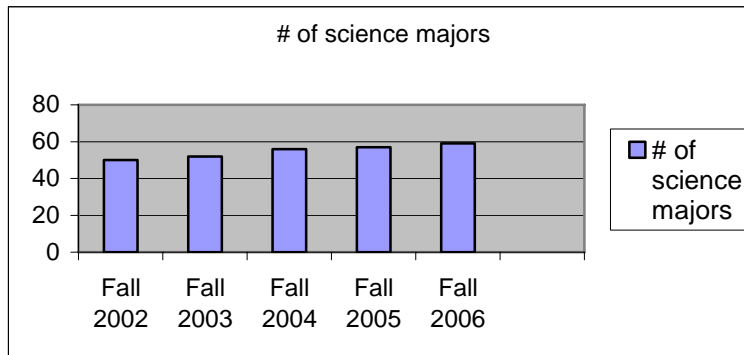
Another very common graph problem is when something other than a two dimensional bar is used to represent the data. It is common to use pictures or three dimensional representations as a graph. These types of representations can be misleading. A three dimensional bar graph is never a good idea, it just makes the data harder to understand and easier to distort. In figure 3, the graph indicates that the number of graduates in 2005 was 5 and the number of graduates in 2006 was 15. However, the larger picture takes up about 8 times the area of the smaller image. The size of the images should always be in proportion to the actual numbers. In this example, the larger image should be only 3 times the size of the smaller image.

Figure 3

What should you look for when evaluating graphs? First, the simplest graph should always be the first choice for any researcher. The idea of a graph is to make the data easy to understand. The more complicated the graph the less likely you will get the point across. The first choice is a bar graph or histogram. Almost all statistical data commonly encountered will be best represented by either a bar graph or a histogram. Use a bar graph (gaps between the bars) when the data is not continuous.   Use a histogram (no gaps between the bars) when the data is continuous. Never use a three dimensional bar graph or histogram! It just makes the data harder to understand.

Second, you should always start your scale at zero. Never start your scale at any value other than zero since this distorts the data. If you start your scale at a number higher than zero, it emphasizes the difference in the heights of your bars. The proportions of the bars will not match that of the data. Also you should be sure to keep the width of the bars the same. This is true when pictures are used instead of bars as well. See Figure 4 for an example of how to correctly use pictures in a bar graph.

Sometimes it seems necessary to emphasize small differences and researchers use truncated scales. This is almost always a problem with bar graphs and you should consider using a line graph instead. It is more acceptable to use a line graph with a truncated scale. However, another valid approach is to show the bar graph without the truncated scale (starting at 0) and with the truncated scale. That way you will not be guilty of trying to distort the data.

Figure 4
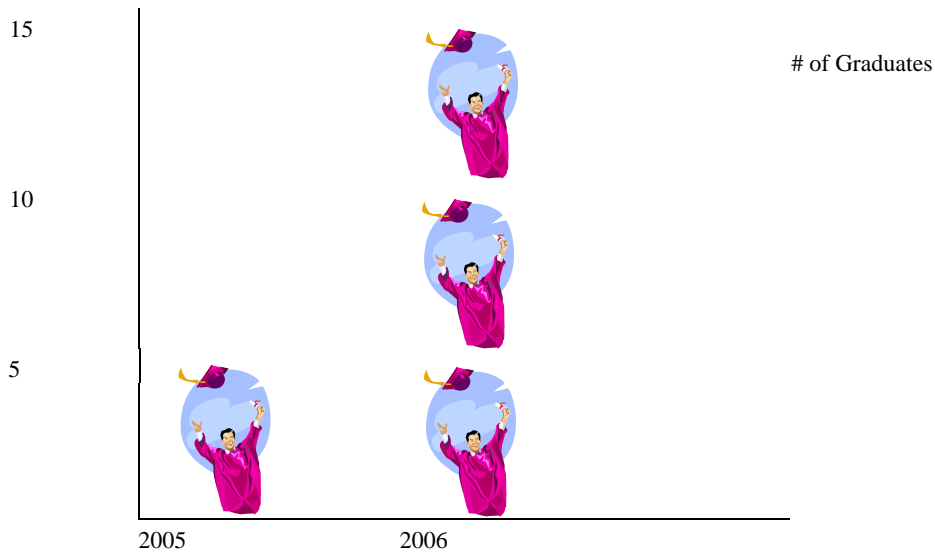
INTERPRETATION

How data is interpreted is also subject to flaws. It is important to understand the meaning of the results and present the interpretation fairly. Using survey results can be tricky depending upon how the survey results are crafted and the questions that need to be answered.

For example, if you wanted to decide whether to offer a course at 7am, 5pm, or 10pm, you might survey those students who had to take the course. Let's say that you asked the students to rank their preference of times. What if 13 students have to take the course and none of the times appeal to any of the students? The results are given in the table:

7am	5pm	10pm
1	3	2
1	2	3
1	2	3
1	2	3
3	1	2
3	1	2
3	1	2
3	1	2
3	2	1
2	3	1
2	3	1
2	3	1
2	1	3

This is the raw data. The researcher decides that the best way to decide what time the class should be scheduled is to add the columns and choose the number that is smallest. The sums are as follows: 24 for 7am, 25 for 5pm, and 26 for 10pm. Based on this data, the researcher decides to schedule the course at 7 am. Now let's consider whether this is a reasonable interpretation of the data. Since all three of the times might be viewed by the students as equally unacceptable, it's not surprising that the sums are almost equal. Notice that even though the total was smallest for 7 am, there are less ones in the column than there are in the 5pm column. The largest number of students picked 5pm as their preferred time. Also notice that the 7 am column had the most students who put it as their least preferred time. By looking at it this way, the researcher should have picked 5pm as the time that was most liked and least disliked by the students. The interpretation of the data was flawed. However, this problem could have been prevented in the first place by changing the way the question was asked to begin with. Please note that this analysis is really taking ordinal data and assuming that it is ratio. What this analysis assumed is that there were equal differences between the answers of 1, 2, or 3. That is not necessarily true. The students might have felt that their number one answer was 5 times better than their number two answer.

Another problem with the interpretation of data is establishing cause and effect. It is a common problem to assume that two correlated events are related. It is often the case that there is no cause and effect. For example, if it is found that the percent of students who graduate is larger for Math majors than for all other majors, does this mean that we have a cause and effect situation? Does being a math major mean that you are more likely to graduate? With cause and effect there are 4 possibilities: first event A may cause event B, or Event B may cause event A, or Event A and B are caused by a third event C, or the occurrence may be random. Even if we eliminate the last possibility by statistical tests, we still have 3 other possible relationships. The most likely one is where a third event is causing the other two. In our example, the Math majors may have better advisors and therefore are more likely to graduate.

CONCLUSION

Using statistics correctly is very important to maintaining the integrity of data. Checking your statistical data for bias is one of the first things that you should do. Trying to eliminate as much bias as possible will result in better data and more accurate decisions. The second thing that should be looked at is the methods used to display and analyze data. Accurate graphs and using the correct statistical measures goes a long way to improve the validity of the data. Finally, you should be careful in your interpretation of the results. Even if all the statistics is done correctly, the results can be interpreted incorrectly.

By following these simple guidelines, you can use statistics properly and get accurate data. Always look at data and the analysis of data with a critical eye. Do not assume that just because statistical data is presented that it is correct or that it has been interpreted correctly. If we all follow these guidelines, statistics can be viewed as the wonderful tool it is.

REFERENCES

1. Helberg, C. (1995) *Pitfalls of Data Analysis*. http://www.execpc.com/~helberg/pitfalls/
2. Huff, D. (1954) *How to Lie with Statistics*. New York: W. W. Norton & Company.
3. King, G. (August, 1986) How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*, 30 (3), 666-687.
4. Utts, J.M. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, 57(2), 74-79.

NOTES