# Crooked Numbers

## *Using Opinions to Shape Statistics*

MICHAEL LENYO

"Every year since 1950, the number of American children gunned down has doubled."[1]

This statistic may shock you and make you question the security we provide for our children, but there is one problem: it is not true. If you carefully analyze this statement you would realize that if this statement was accurate, the number of American children killed by guns in 1995 would be 35 trillion even if the number killed in 1950 was one. Clearly, you would then realize that the number of children killed each year could not be nearly 4400 times the size of the world's population. If you simply looked at this statement and did not think deeply about what it was literally saying, you may accept it as the truth. This appeared in a newspaper article and was quite different from the original data. The original statistics, provided by the Children's Defense Fund, stated the number of American children killed by guns each year has doubled since 1950.[1] This simple difference in wording provides a much more powerful statement that is actually completely false.

These statistics shape our perceptions and alter our decision making process. Statistics like these appear everywhere in our world. You can find them in newspapers, magazines, reports, and many other places. This global display of false numerical claims raises questions about the ethics of statistics and what is expected of the people providing these numbers.

In this paper, I will explain the various ways in which statistics can be flawed, the ethics involved in these misrepresentations, and possible ways to limit the power of misleading statistics. I will examine the ethics from duty-based perspective, and discuss the obligations of both the presenters of the information and the audience using that information to make decisions.

**Human Perceptions**

Prior to discussing anything further on statistical manipulation, we must first examine the human response to statistics. For almost any argument, statistics are employed to represent facts. An exploration into the definition of statistics, as expressed by early statisticians, returns a common idea that statistics are "numerical statements of fact".[2] When people see statistics, they tend to believe the numbers are sound and were reliably collected.[3] It is clearly seen that people trust statistics simply by looking at the wide range of areas in which they are applied. Statistics are used to make decisions in government, the economy, science, medicine, and even in our own personal lives.[4] If we continuously use statistics to make important decisions in our society, should we not question whether statistics represent concrete facts? Since statistics help drive critical decision making, it is important to not be making conclusions

from anything other than actual facts. A careful examination of all the processes involved in collecting and compiling statistics is required in order to fully understand why this data should not be merely accepted as fact.

**Uses of Statistics**

We must first establish why the ethics of statistics is a relevant topic. Even by viewing the material of an introductory statistics class, one can understand the wide array of uses for statistics. Surveys are a major source of statistical data. They are used in dozens of different areas. Government agencies use surveys to determine the amount of unemployment and to establish the Consumer Price Index (CPI). The CPI is a major economic indicator, which can be used to evaluate the effectiveness of an administration's economic policies.[5] Statistics are also commonly used to determine to what extent a company is financially stable. This data helps guide if an individual should invest in particular company's stock.

Statistical studies are used as the basis for making decisions in several other areas. Companies use surveys to conduct marketing research to help determine what customers want to buy. Sociological research can be done by surveying people to understand the way people live and the way society is constructed.[6]

In education, schools are judged based on the statistics of their students test scores. Funding can be largely based on these numbers. Thousands of schools can be affected if false data is used. Studies are also often used in medicine to determine the effectiveness of a drug and its possible side effects. The implications flawed data can have on the users

of these drugs are enormous. An unsafe drug could reach the market if incorrect data showed it was safe. They also use statistics to determine the important risk factors that can lead to contracting a certain disease. Accurate numbers can help guide doctors into testing for diseases for which a patient is more likely to contract.

From these few examples taken from the thousands of uses for statistics, it is clear that the accuracy of statistics used in decision making is truly an issue. With so many key choices being dependent on what the numbers tell us, it is important to consider the ethical uses of these numbers and the effects misleading statistics can have.

**Lying with Statistics**

The term lying leads many to think that someone is purposely attempting to deceive them. In order to fully understand how statistics can be misleading, one must realize that people can accidentally present statistics that misrepresent the actual data.[7] Before considering the ethics associated with statistics, we must understand these two broad areas of statistical misinformation. It also must be noted that since accidental and intentional misrepresentations are quite different, the ethical examinations should be done separately.

**Statistical Errors**

When evaluating the accuracy of statistics, one must look at the creation process. Data is collected, interpreted, and then presented. Mistakes can be made at all levels of construction. It is important to

understand the many types of errors that can occur in order to realize how easily statistics can be flawed.

The most common errors in statistics are known as sampling errors.[8] When statistics are collected, the entire population is not usually included in the test. Instead, a sample is taken which is supposed to be indicative of the entire population. It is clear that errors can arise because the characteristics of a smaller portion of a whole do not always exactly represent the characteristics of that whole. By using probability sampling methods, it is possible to estimate the amount of sampling errors, but probabilities are not perfect. Sampling error can be reduced by using larger samples and also by using effective methods to select the samples. In all studies, it is accepted that the information has a certain amount of unreliability. The main problem, though, is that this amount of unreliability is not usually stated when the statistics are displayed. The audience is unaware of the possibility of inaccurate data.

It is important to understand that numbers must be examined in context. Many make the mistake of comparing just raw numbers and not really taking into account the specific situations surrounding those numbers.[9] If one was to compare the number of police officers in Los Angeles to the number in Anchorage without considering the population difference, it is easy to make the claim that Los Angeles is safer than Anchorage. The raw numbers simply do not tell the entire story and thus any conclusions drawn from just those numbers could be faulty.

One of the main concerns involved during the collection stage of statistical analysis is the likelihood of bias.[10] Again, we look at the idea of samples. The people conducting the surveys are in control of selecting what to include or who to include in the collection. If this group intends to prove a particular point, it is clearly seen that they might be drawn to

samples that would be more likely to present figures to support their position. It is important to note that this is not always the intent, but it is almost impossible to conduct a survey without some bias.

Another problem with the people in charge of the surveys is that many times these people do not have a strong knowledge of the subject matter.[11] If someone does not understand what they are working on, it is hard to be certain that the important data is being collected and displayed. You would not want someone who does not have a strong understanding of medicine and diseases to being testing drugs to help cure diseases.

**Ethics of Flawed Data**

As stated before, I will analyze this from a deontological or duty-based perspective. Mostly based on the ideas of Immanuel Kant, this theory suggests that our morals are based on the obligations we have to each other.[12] Every action is basically subject to a universal moral law which promotes treating individuals as ends in themselves rather as means to an end. By this logic, everyone would treat everyone else based on standard rules, and thus all could be treated fairly. This theory does not look at consequences because one cannot always be sure what could result or the implications of those results.

I must mention in my discussion of mistakes that my assertion is that making mistakes is an ethical matter. Even though one can claim that an accidental error is not an ethical matter because it is not a purposeful attempt to deceive, I suggest that acknowledging that errors will occur and taking action to limit those errors is subject to ethical scrutiny. It is widely accepted that no survey can be absolutely accurate and that there

will always be mistakes. The ethical issue arises in the idea that people have some obligations to realize the possible errors and take measures to reduce their impact on the overall survey. In other words, I will discuss whether it ethical to simply ignore the possibility of errors and take no action to attempt to eliminate them. When analyzing the ethics involved in making mistakes in statistical studies, we must first look at what would be involved in ensuring these mistakes are limited. First, exploring errors made in displaying the statistics, it certainly seems that more effort could be made to have data shown accurately. From a duty-based perspective, statisticians and those using statistics have a duty to attempt to provide the most accurate information. Based on this duty, it is unethical to cut corners and not attempt to display information in the most accurate way possible. In displaying, it is simply a matter of selecting an accurate way to show the particular statistics. It may take a little more time and research to find the best way to display the data, but it is necessary to uphold the duty to the audience.

Now it is important to consider the problems with data collection differently than the display of that data. The data collectors still have the same duty to attempt to produce the most accurate information, but they are limited in terms of collection. They should attempt to get the largest possible sample and the one that is most comparable to the population.[13] They may be limited by the cost of collecting larger samples, but again I emphasis they must attempt, to the best of their ability, to reduce errors. I will not claim that there should be a specified size of a sample, but based on the constraints of each individual situation, one could determine the largest possible sample. Even with this though, it would be impossible to eliminate all bias and sampling errors. To act ethically, their goal should just be to limit the amount of errors. They clearly have a duty to select

qualified people to conduct the studies. It may cost more to hire experts, but that is the only way to ensure the most important data is being given the most attention and possible mistakes made out of ignorance are limited. If they cannot afford to qualified surveyors, the ethical thing to do may be just not to conduct the survey at all. If the accurateness of the survey is compromised, it will really not serve its purpose and will not really be useful in decision making.

**Bending the Numbers**

Ethics in statistics becomes even more of a concern when you examine the ways people use statistics to intentionally deceive their audience. People want the statistics they provide to help prove the argument they are trying to make.[14] This desire may drive some to manipulate their data to lead to certain conclusions. I will break intentional misuse of statistics into three basic categories: manipulation of raw data, displaying data in a misleading way, and completely omitting data.

*Data Manipulation*

There are many ways people can lie with statistics. One obvious way would be to alter the data or simply make up the data all together.[15] Someone can do a study or claim to do a study and then present some numbers. These numbers could have no truth behind them, but since people often accept statistics as hard facts without question[5],[16] this type of practice has a good chance of actually obtaining its desired results.

When we first discussed bias, we only looked at people accidentally being influenced. People can also purposely attempt to steer data. If someone believes deeply in a cause, that person may purposely select data

that would best support that cause. The data itself is correct for the specific sample, but most likely does not accurately represent the entire population.

In the case of surveys, the interviewer could manipulate the data by manipulating the test subject. They could change the way they form their questions or the way they ask questions in order to lead the subject to a certain conclusion. In this way, the data is actually the data that was collected, but still would not present the truth.

It is clear that data can easily be manipulated and there are a really no sure ways to prevent it. Laws really can't do much to stop people from fabricating data, so ethical thought is the only way to really look at this manipulation.

*Displaying Techniques*

Statistics can be misleading because of the way they are displayed. The data may be accurate, but there is still an attempt to be deceptive. Display methods can easily deceive the audience. In Figure 1 below, simply changing the scale from 0-250 to 100-215 gives a completely different idea about the number of police in each city.

Fig. 1

There are several other ways to make data appear to say different things. The way things appear visually to us has a real impact on our opinions. In the second graph it appears that City 2 has twice as many police officers as City 1, but looking at the first graph, you can tell that is incorrect. Deception can occur based on how the information is displayed.

*Data Omission*

Another way people can lie with statistics is by omitting them.[17] A prime example would be with the drug company Merck.[18] They developed a drug for treating arthritis called Vioxx. Prior to submitting this drug to the public, they conducted a series of tests that showed that it had a strong correlation to increased heart attack risk. Merck proceeded to rush the drug to market and never included any information to the public about the heart attack risk. The data Merck obtained in their study would have had a negative effect on selling their drug and thus they simply chose to omit some very important information. This is misleading because Merck had the data showing the heart attack risk, but since they omitted that data, people have a false sense that there is no increased heart attack risk.

**Ethics**

There are many reasons why people choose to manipulate statistics. They may feel pressure from outside sources, they may be constrained by budget, or they may have personal objectives or values they wish to support.[19]

To understand the dilemma of statistical manipulation it is important to see that good can come out of statistical manipulation. In the case of the drug Vioxx, many people could benefit from the drug's positive effects. By omitting some data, Merck was able to get the drug into the market for the people who needed it. They may have believed that the studies they did were valid, but the positives of the drug outweighed the increased heart attack risk. Thus, by misleading their audience, they were able help thousands of people.

I would contend though, that even if the drug caused no heart attacks in the population, Merck did act unethically. Merck conducted a study and had a duty to provide the results of that study. Ethically, no one should ever manipulate data, even if they have good intentions and it could lead to a favorable outcome. First, you have to consider who would decide what a good outcome is. A good outcome for one group of people is not always a good outcome for another. Even if you ignore that fact, and we hypothetically say that everyone agrees on what is a good outcome, you could still not argue for manipulation based on good intentions. You could not universally apply some sort of maxim that suggested that you should manipulate data if it can lead to good results. The very practice of statistical analysis would be useless. If everyone simply changed factual data to support their cause or even just left out the data against it, all data would loss any meaning it could have had. In order for statistics to actually be a useful, factual tool, statisticians must consistently present the genuine data.

Data manipulation can in no way be ethical. Statisticians have a duty to present the accurate and complete results of their studies. Accepting any sort of manipulation means accepting all manipulation, good and bad.

**Conclusion**

Proper use of statistics and making decisions based on statistics is the responsibility of both the people supplying the statistics and those using them. The statisticians have a responsibility to provide the most accurate information possible. We have already discussed their ethical responsibility pertaining to intentional manipulation and also that they must be required to reduce their mistakes. It is their responsibility to study the mistakes they made and improve their statistical system.[20] There are many ways to ensure that data is accurate as it can be. They can look at other similar studies that have been successful in using appropriate methods and learn how best to conduct their own study. They can do the research multiple times with different samples. Some methods may not be cost effective, but there is always a way to better your data.

One must know that it is also the viewer's job to determine how to use the statistics he is presented. Audiences need to be aware of the possibility that the statistics they are seeing may not actual be representative of the truth. It is important to analyze data rather than just accepting it as the truth. The audience also has a duty to make well-informed and conscientious decisions. To do this, they must take certain steps to analyze the validity of statistics. First, when viewing statistics, they must step back and look at the data as impartial observers. Humans tend to confront statistics that question our beliefs and are more forgiving of those which support them.[21] We need to consider the issue of construction. Before we accept a statistic, we must analyze who is giving us this statistic, where they got their information, and what methods they used to get it.[22] Understanding these things will help us understand

possible biases and other problems associated with the stats. The main thing to remember is not to simply accept statistics as fact.

If we simply look deeper into statistics, we can weed out false and misleading information and be able to have a more informed perspective. So remember this the next time you see something like this:

> "In a study in 1685 of the ages and professions of deceased men, it was found that the profession with the lowest average age of death was 'student.' Being a student seems like a very dangerous occupation."[23]

Examine this statement closely and now you should be able to understand why we should think about what we read or we might just jump to some absurd conclusions.

## Works Cited


1. Best, Joel. (2001). *Damned Lies and Statistics.* Berkeley: University of California Press.

2. Annadurai, B. (2007). *A Textbook of Biostatistics.* New York: New Age International.

3. Best, Joel. (2005). Lies, Calculations and Constructions: Beyond How to Lie with Statistics. *Statistical Science,* 20, 210-214.

*4. Ibid* 2

5. U.S. Bureau of Labor Statistics. (2001). Consumer Price Index. *BLS.* Retrieved November 19, 2007, from http://www.bls.gov/CPI

6. Wild, Chris and George Seber. (1999). *Chance Encounters: A First Course in Data Analysis and Inference.* New York: John Wiley and Sons.

7. Gelman, Andrew and Deborah Nolan. (2002) *Teaching Statistics: A Bag of Tricks*. New York: Oxford University Press.

8. Groves, Robert. (2004). *Survey Errors and Survey Costs.* New York: Wiley-Interscience.

9. *Ibid* 7

10. *Ibid* 7

11. Jaffe, A.J. (1987). *Misused Statistics: Straight Talk for Twisted Numbers.* New York: Marcel Decker Inc.

12. Kant, Immanuel. (1997). *Groundwork of the Metaphysics of Morals.* New York: Cambridge University Press.

13. Arnold, Margery. (1996). *The Effects of Two Types of Sampling Error on Common Statistical Analyses.* New Orleans: Southwestern Educational Research Association. (ERIC Document Reproduction Service ED395952)

14. Huff, Darrell. (1954). *How to Lie with Statistics.* New York: Norton.

15. *Ibid* 7

16. *Ibid* 5

17. *Ibid* 10

18. Testimony of David J. Graham, MD, MPH, November 18, 2004.
Associate director for Science and Medicine in the Office of Drug Safety, FDA.
19. Seltzer, William. (5 Feb. 2001). *U.S. Federal Statistics and Statistical Ethics: The Role of the American Statistical Association's Ethical Guidelines for Statistical Practice*. Washington Statistical Society Seminar: Washington.
20. Seltzer, William. (2005). *Official Statistics and Statistical Ethics: Selected Issues.* New York: International Statistical Institute.
21. *Ibid* 3
22. *Ibid* 3
23. *Ibid* 7