AP Statistics – Summer 2022

Hello, 2022-2023 AP Statistics Student (and Parent/Guardian),

Congratulations on deciding to take the most applicable math class you have ever taken! It will be different than any other math course you have taken in high school.

There is no resource on this planet more valuable than data and through your experiences in AP Statistics, you will learn how to consume and analyze different types of this data. Our course includes four major themes: exploring data; planning a study; anticipating patterns; and producing models using probability, simulation, and statistical inference while extensively using technology. The AP Statistics course will best prepare you for the knowledge of statistical analysis required by most fields of study in college.

Over the summer, you should read the article after this letter and complete this Google Form: **https://tinyurl.com/MHSAPStat**

The assignment is due Friday, September 9th.

Have a great summer!

Mr. Doris

NOT THE WHOLE TRUTH: LYING WITH STATISTICS

Data analysis and the field of statistics are thought to be hard sciences: data is gathered, analyzed using certain processes – often precise mathematical ones – and findings are reported.

But this is not always the case. We do not always hear (or see) the whole truth about data findings; it is possible for them to be distorted. This is not the same as there being unintentional flaws in the data collection or data analysis methods. No, this is taking liberty with how the data is obtained or presented.

There are many different ways a researcher or statistician can lie or mislead with the statistics they report. This article will look at a few of the common ways that this is done. A good source of additional information is Darrell Huff's *How to Lie with Statistics* (1982).

Purposeful Bias

There are several types of bias that can be used to distort data. At the outset of data collection, bias can occur if the sample is not representative of the entire population to which the results are intended to be generalized. Choosing certain subsets of people or using a sample of convenience are two ways to have bias in the sample. For example, a survey being conducted on a college campus in which people in the same major or in certain classes will not represent the entire student body.

The way the research is conducted can also be biased. An example from survey research is having questions that lead the responder in a specific direction or leaving out legitimate choices as responses because the researcher does not want to have such responses. This can also happen during personal or group interviews as well as in experimental studies in which a researcher or research assistant is directly involved with the study's subjects.

At the other end is the reporting of findings. Bias can occur if only the statistics that support the researcher's claim are given. An example of this might be someone who is studying the number of lottery tickets the average person who plays the lottery buys at one time. Assume he observes the following number of tickets bought by 10 people in a half-hour period at one particular outlet:

1   1   1   1   4   4   5   8   15   200

This set of data has the following measures of central tendency:

Mean: 24

Median: 4

Mode: 1

If the researcher wants to support a claim that people are buying lots of tickets, he will use the mean; if he wants to say people are buying very few, he will use the mode (if he wants very few) or the median.

The Misleading Graph

Displaying data in a graph or some other pictorial manner is one of the most common ways to communicate statistical information to others, especially the public. But graphs can be manipulated to show a perspective on the data that the person who makes the graph prefers. One technique for doing this is "zooming" in or out: bringing the data close together to cause distortion or being far back from the data to blur details. The histogram below is a good representation of the results of a hypothetical survey on preferences for ice cream flavors, with the top five shown in the graph. It shows that vanilla is the most-preferred flavor and the others are close in preference.
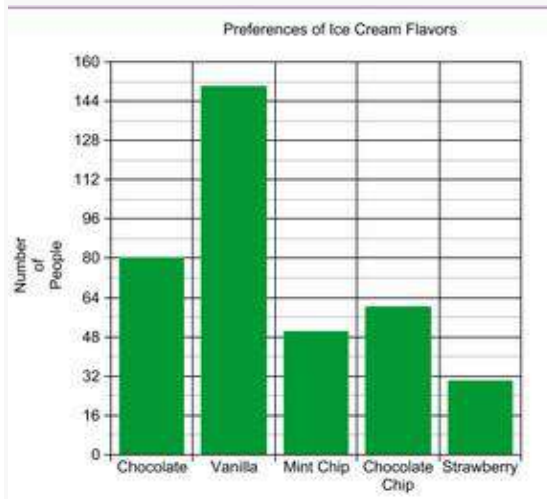
But the graph can be distorted to show that vanilla was much more highly favored and that strawberry had few takers. This is accomplished by zooming in so that the y-axis (number of people choosing that flavor) does not begin at zero and by making the highest value on the graph right at the level of the highest data value.
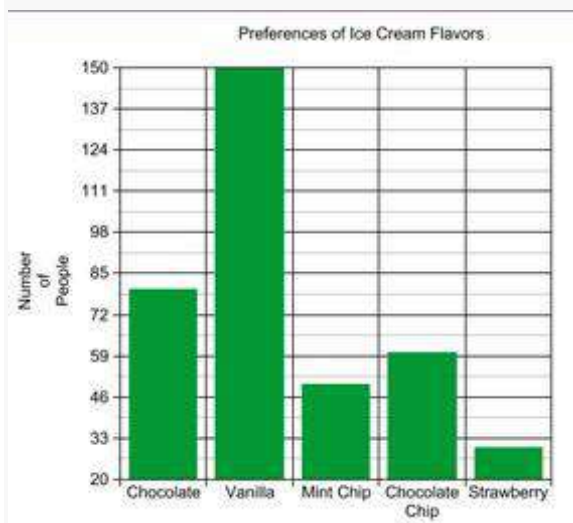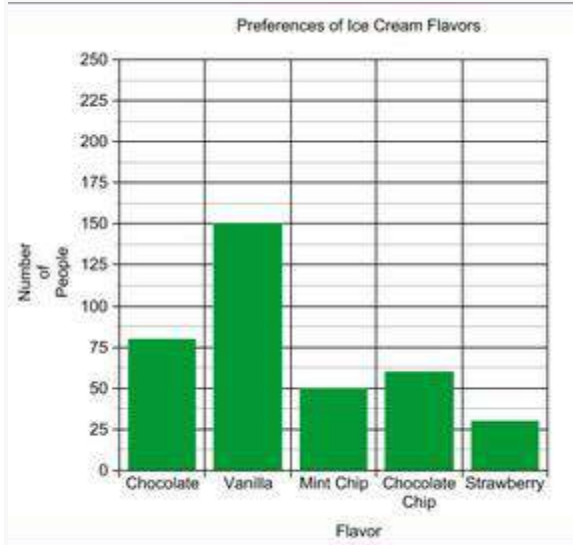
On the other hand, the differences can be shown to be less than they really are by zooming out. This is accomplished by raising the top value on the chart well beyond the actual highest data value.

With today's technology, it is relatively easy for one to distort a graph to make results look better or worse than they really are. This is why it is important for consumers of the data to study the details of the data carefully and not take charts and diagrams at face value, no matter how nice they look.

The Left Out Data

Another way to make a study's results look better is by leaving out data that does not support the hypothesis or goal of the study. This often involves values that the researcher classifies as outliers, possibly attributing them to mistakes in the way the data was collected. By eliminating data that does not suit the researcher's needs, he can make the data say what he wants it to say. This is an egregious, but not uncommon, way to lie with statistics.

As a hypothetical example, suppose a teacher wants to apply for a grant to implement a project he has been doing on a pilot basis. To get the grant, he must show evidence that the pilot was effective in, say, improving test scores. After running the pilot and giving the test, he finds that 16 of 20 students have improved test scores but four still have very low scores. Using just the 16 scores, the teacher says his new method is effective. The other scores, he explains, cannot be counted for one reason or another, such as the students missed school days while the pilot was being conducted or the students were fooling around while taking the test.

All of these ways of lying with statistics point out why independent review of research projects and full disclosure of data are necessary for legitimate research. It also provides a warning for those who use the results of studies to understand the processes involved in the collection and analysis of the data and the details of the results.