WHAT EDUCATED CITIZENS SHOULD KNOW ABOUT STATISTICS AND PROBABILITY

Jessica Utts
University of California
USA

Much has changed since the widespread introduction of statistics courses into the curriculum in the 1960s and 1970s, but the way introductory statistics courses are taught has not kept up with those changes. This paper discusses the changes, and the way the introductory syllabus should change to reflect them. In particular, seven ideas are discussed that every student who takes elementary statistics should learn and understand in order to be an educated citizen. Misunderstanding these topics leads to cynicism among the public at best, and misuse of study results by physicians and others at worst.

INTRODUCTION

Statistical studies are prominently featured in most major newspapers on a daily or weekly basis, yet most citizens, and even many reporters, do not have the knowledge required to read them critically. When statistics courses were first introduced in the 1960s and 1970s, they were primarily taken by students who intended to pursue their own research. The focus of those courses was on computation, and little emphasis was placed on how to integrate information from study design to final conclusions in a meaningful way. Much has changed since then, in three ways: the audience, the tools available to students, and the world around us.

The Audience:

- Broader set of majors represented, many will never “do” statistics
- Greater age mix – more likely to have “returning” students

The Tools For Students:

- Universal use of calculators, most have keys for mean, standard deviation, etc.
- Universal access to computers
- Programs like *Excel* have standard statistical features
- Programs like *Minitab* and *SPSS* are now menu-driven

The World Around Us:

- Many more studies reported in the news
- Abundance of examples available on the Internet through sites like Gallup, USA Today, Bureau of Labor Statistics, etc.
- Journal articles available on-line

The consequence of all of these changes is that students have less need to do calculations, and more need to understand how statistical studies are conducted and interpreted.

SEVEN IMPORTANT TOPICS

There are of course many important topics that need to be discussed in an elementary statistics course. For this paper, I have selected seven topics that I have found to be commonly misunderstood by citizens, including the journalists who present statistical studies to the public. In fact the researchers themselves, who present their results at the scientific meetings from which the journalists cull their stories, misunderstand many of these topics. If all students of introductory statistics understood them, there would be far less misunderstanding and confusion in the public eye. In fact the public is often cynical about statistical studies, because these misunderstandings lead to the appearance of a stream of studies with conflicting results. This is particularly true of medical studies, where the misunderstandings can have serious consequences when neither physicians nor patients can properly interpret the statistical results.

A summary of the seven topics is presented first, followed by a more in-depth explanation with examples of each topic:

1. When it can be concluded that a relationship is one of cause and effect, and when it cannot, including the difference between randomized experiments and observational studies.
2. The difference between statistical significance and practical importance, especially when using *large* sample sizes.
3. The difference between finding "no effect" or "no difference" and finding no *statistically significant* effect or difference, especially when using *small* sample sizes.
4. Common sources of bias in surveys and experiments, such as poor wording of questions, volunteer response and socially desirable answers.
5. The idea that coincidences and seemingly very improbable events are not uncommon because there are so many possibilities.
6. Understanding that what may appear to be a trend is just a part of a cycle in a time series.
7. Understanding that variability in natural, and that “normal” is not the same as “average”.

CAUSE AND EFFECT

Probably the most common misinterpretation of statistical studies in the news is to conclude that when a relationship is statistically significant, a change in an explanatory variable is the *cause* of a change in the response variable. This conclusion is only appropriate under very restricted conditions, such as for large randomised experiments. For single observational studies, it is rarely appropriate to conclude that one variable caused a change in another. Therefore, it is important for students of statistics to understand the distinction between randomised experiments and observational studies, and to understand how the potential for confounding variables limits the conclusions that can be made from observational studies.

As an example of this problem, an article appeared in *USA Today* titled "Prayer can lower blood pressure (Davis, 1998)." The article reported on an observational study funded by the United States National Institutes of Health, which followed 2,391 people aged 65 or over for six years. One of the conclusions reported in the article read:

> *“Attending religious services lowers blood pressure more than tuning into religious TV or radio, a new study says. People who attended a religious service once a week and prayed or studied the Bible once a day were 40% less likely to have high blood pressure than those who don’t go to church every week and prayed and studied the Bible less"* (Davis, 1998).

The headline and the displayed quote both indicate that praying and attending religious services actually *causes* blood pressure to be lower. But there is no way to determine a causal relationship based on this study. It could be that people who are healthier are more able to attend religious services, so the causal relationship is the reverse of what is attributed. Or, it could be that people who are more socially inclined are less stressed and thus have lower blood pressure, and are more likely to attend church. There are many other possible confounding variables in this study that could account for the observed relationship. The problem is that readers may mistakenly think that if they alter their behavior with more prayer and church attendance, it will cause their blood pressure to lower.

Another example illustrates that even researchers can make this mistake. An article in *The Sacramento Bee* (Perkins, 1999) reported on an observational study that observed a random sample of over 6,000 individuals with an average age of 70 when the study began. The study followed them over time and found that a majority, over 70%, of the participants did not lose cognitive functioning over time. One result was quoted as *"Those who have diabetes or high levels of arteriosclerosis in combination with a gene for Alzheimer's disease are eight times more likely to show a decline in cognitive function (Perkins, 1999)."* So far, so good, because the reporter is not implying that the increased risk is causal. However, one of the original researchers was not as careful. The researcher was quoted as follows: *"That has implications for prevention, which is good news. If we can prevent arteriosclerosis, we can prevent memory loss over time, and we know how to do that with behavior changes - low-fat diets, weight control, exercise, not smoking, and drug treatments"* (Perkins, 1999).

In other words, the researcher is assuming that high levels of arteriosclerosis are *causing* the decline in cognitive functioning. But there are many possible confounding variables that may cause both high levels of arteriosclerosis and decline in cognitive functioning, such as genetic disposition, certain viruses, lifestyle choices, and so on. Resisting the temptation to make a causal conclusion is particularly difficult when a causal conclusion is logical, or when one can think of reasons for how the cause and effect mechanism may work. Therefore, it is important to give many examples and to discuss how confounding variables may account for the relationship.

STATISTICAL SIGNFICANCE AND PRACTICAL IMPORTANCE

Students need to understand that a statistically significant finding may not have much *practical* importance. This is especially likely to be a problem when the sample size is large, so it's easy to reject H_0 even if there is a very small effect. It is also a common problem when multiple comparisons are done, but only those that achieve statistical significance are reported. As an example, the *New York Times* ran an article with the title "Sad, Lonely World Discovered in Cyberspace (Harmon, 1998)." It said, in part:

> *" People who spend even a few hours a week online have higher levels of depression and loneliness than they would if they used the computer network less frequently... it raises troubling questions about the nature of "virtual" communication and the disembodied relationships that are often formed in cyberspace"* (Harmon, 1998).

It sounds like the research uncovered a major problem for people who use the Internet frequently. But on closer inspection, the magnitude of the difference was very small. On a scale from 1(more lonely) to 5, self-reported loneliness decreased from an average of 1.99 to 1.89, and on a scale from 0 (more) to 3 (less), self-reported depression decreased from an average of 0.73 to 0.62. Here is another example of how a very large sample size resulted in a highly statistically significant difference that seems to be of little practical importance. The original report was in *Nature* (Weber et al, 1998), and a *Reuters* article on the Yahoo Health News website ran a headline “Spring Birthday Confers Height Advantage (Feb 18, 1998).” The article described an Austrian study of the heights of 507,125 military recruits, in which a highly significant difference was found between recruits born in the spring and the fall. The difference in average heights was all of 0.6 centimetres, or about ¼ inch.

LOW POWER VERSUS NO EFFECT

It is also important for students to understand that sample size plays a large role in whether or not a relationship or difference is statistically significant, and that a finding of “no difference” may simply mean that the study had low power. For instance, suppose a study is done to determine whether more than a majority of a population has a certain opinion, so the test considers H_0:p=.5 versus H_a:p>.5. If in fact as much as 60% of the population has that opinion, a sample size of 100 will only have power of .64. In other words, there is still a 36% chance that the null hypothesis will not be rejected. Yet, reporters often make a big deal of the fact that a study has “failed to replicate” an earlier finding, when in reality the magnitude of the effect mimics that of the original study, but the power of the study was too low to detect it as statistically significant.

As an example with important consequences, a February 1993 conference sponsored by the United States National Cancer Institute (NCI) conducted a meta-analysis of eight studies on the effectiveness of mammography as a screening device. The conclusion about women aged 40-49 years was *“For this age group it is clear that in the first 5-7 years after study entry, there is no reduction in mortality from breast cancer that can be attributed to screening (Fletcher et al, 1993).”* The problematic words are that there *is no reduction*. A debate ensued between the NCI and American Cancer Society. Here are two additional quotes that illustrate the problem:

> *"A spokeswoman for the American Cancer Society’s national office said Tuesday that the ...study would not change the group’s recommendation because it was not big enough to draw definite conclusions. The study would have to screen 1 million women to get a certain answer because breast cancer is so uncommon in young women".* San Jose Mercury News, Nov 24, 1993.”

> *"Even pooling the data from all eight randomized controlled trials produces insufficient statistical power to indicate presence or absence of benefit from screening. In the eight trials, there were only 167,000 women (30% of the participants) aged 40-49, a number too small to provide a statistically significant result"* (Sickles & Kopans, 1993).

The confidence interval for the relative risk after 7 years of follow-up was 0.85 to 1.39, with a point estimate of 1.08, indicating that there may be a small reduction in mortality for women in this age group, or there may be a slight increase. (See Utts, 1999, p.433). The original statement that there was “no reduction in mortality” is dangerously misleading.

BIASES IN SURVEYS

There are many different sources through which bias can be introduced into surveys. Some of the more egregious are difficult to detect unless all of the details are understood. For example, a Gallup Poll released on July 9, 1999, based on a random sample of 1016 U.S. adults, asked two different questions in random order, each of which could be used to report the percentage of people who think creationism should be taught in public schools in the United State. The two questions, and the proportion that answered, “Favor” were:

Question 1: *Do you favor or oppose teaching creationism ALONG WITH evolution in public schools?* (68% Favor).

Question 2: *Do you favor or oppose teaching creationism INSTEAD OF evolution in public schools?* (40% Favor).

Notice that depending on one’s own opinion, these results could be misused to advantage. Someone in favor of creationism could report that 68% think it should be taught, while someone opposed to creationism could report that only 40% think it should be taught. There are many examples of how question wording, question order, method of sample selection and many other issues can bias survey results. See Utts (1999) or Utts and Heckard (2002) for lengthy discussion and examples.

PROBABLE COINCIDENCES

Most people have experienced one or more events in their lives that seem to be improbable coincidences. Some such events are so surprising that they attract media attention, often with estimates of how improbable they are. For instance, Plous (1993) reported a story in which a Mr. and Mrs. Richard Baker left a shopping mall, found what they thought was their car in the parking lot, and drove away. A few minutes later they realized that they had the wrong car. They returned to the parking lot to find the police waiting for them. It turned out that the car they were driving belonged to *another* Mr. Baker, who had the same car, with an identical key! Plous reported that the police estimated the odds at a million to one.

The problem with such stories and computations, is that are based on asking the wrong question. The computation most likely applies to that exact event happening. A more logical question is, what is the probability of that or a similar event happening sometime, somewhere, to someone. In most cases, that probability would be very large. For instance, I was once on a television talk show about luck, with a man who had won the million dollar New York State lottery twice, and the host of the show thought this demonstrated extraordinary luck. While it may have been wonderful for that individual, Diaconis and Mosteller (1989) report that there is about an even chance of the same person winning a state lottery in the United States in a seven year period. That was precisely the interval between the two wins for this person.

Remember that there are over six billion people in the world, with many circumstances occurring to each one daily. Therefore, there are surely going to be some that seem incredible. In fact if something has only a one in a million probability of happening in a given day, it will happen, on average, to over 6000 people in the world, each day. When the media reports an incredible coincidence, psychic prediction, and so on, it should be viewed from this perspective.

CYCLES VERSUS UPWARD OR DOWNWARD TRENDS

Short-sighted, short-changed is a good way to express this problem. Most economic and social indicators follow cycles, but an upward or downward part of a cycle can be misinterpreted as an ongoing upward or downward trend if the time-span examined is too short. As examples, unemployment rates and interest rates tend to go up over several years and then down over several years, but the cycle of ups and downs tends to repeat itself every few decades.

AVERAGE VERSUS NORMAL

The final lesson students need to understand is that of natural variability and its role in interpreting what is "normal." Here is a humorous example, described by Utts and Heckard (2002). A company near Davis, California was having an odor problem in its wastewater facility, which they tried to blame on "abnormal" rainfall:

> *"Last year's severe odor problems were due in part to the extreme weather conditions created in the Woodland area by El Nino [according to a company official]. She said Woodland saw 170 to 180 percent of its normal rainfall. 'Excessive rain means the water in the holding ponds takes longer to exit for irrigation, giving it more time to develop an odor'* (Goldwitz, 1998)*."*

The problem with this reasoning is that yearly rainfall is extremely variable. In the Davis, California area, a five-number summary for rainfall in inches, from 1951 to 1997, is 6.1, 12.1, 16.7, 25.4, 37.4. The rainfall for the year in question was 29.7 inches, well within the "normal" range. The company official, and the reporter, confused "average" with "normal." This mistake is very common in reports of temperature and rainfall data, as well as in other contexts. The concept of natural variability is so crucial to the understanding of statistical results that it should be reinforced throughout the introductory course.

CONCLUSION

The issues listed in this paper constitute one list of common mistakes in understanding statistics and probability. There are others, but I have found these to be dangerous in the sense that millions of people can be mislead by these misunderstandings. It is the responsibility of those of us teaching introductory statistics to make sure that our students are not among them.

REFERENCES

Davis, R. (1998). Prayer can lower blood pressure. *USA Today*, August 11, 1998, 1D.

Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association, 84*, 853-861.

Fletcher, S.W., Black, B., Harris, R., Rimer, B.K., & Shapiro, S. (1993). Report on the international workshop on screening for breast cancer. *Journal of the National Cancer Institute, 85(20),* 1644-1656.

Goldwitz, A. (1998). *The Davis Enterprise*, March 4, 1998, A1.

Harmon, A. (1998). Sad, lonely world discovered in cyberspace. *New York Times,* August 30, 1998, A3.

Perkins, K.D. (1999). Study: age doesn't sap memory. *Sacramento Bee*, July 7, 1999, A1, A10.

Plous, S. (1993). *The psychology of judgment and decision making.* New York: McGraw Hill.

Stickles, E.A., & D.B. Kopans (1993). Deficiencies in the analysis of breast cancer screening data. *Journal of the National Cancer Institute, 85(20),* 1621-1624.

Utts, J.M. (1999). *Seeing Through Statistics, 2nd Edition*. Pacific Grove, CA: Duxbury Press.

Utts, J.M., & Heckard, R.F. (2002). *Mind on statistics*. Pacific Grove, CA: Duxbury Press.

Weber, G.W., Prossinger, H., & Seidler, H. (1998). Height depends on month of birth. *Nature, 391(6669),* Feb 19, 754-755.