Lesson 5
Sampling, Experiments, and Observational Studies

Outline of the Lesson

As the tax-filing deadline approached in early April of 2011, a national polling organization conducted a survey. They asked the question, "Do you think the amount of taxes you pay is fair?" News media reported the results as, "According to a recent survey, 54% of Americans age 18 and above believe that the amount of taxes they pay is fair, with a margin of error ±4.2%." This is an example of ***inferential statistics***: the polling organization certainly did not poll every person age 18 and above in the entire country, but the results of the poll were reported as a fact about the entire country. The pollsters made an *inference* about the entire country, based on the results of the poll.

In late May of 2011, the World Health Organization (WHO) reversed its previous stand, and announced its conclusion that radiation from cell phones can possibly cause cancer. This conclusion was based on a number of studies, one of which was a large international study released in 2010. That study showed that participants in the study who used a cell phone for 10 years or more had double the rate of a particular type of brain cancer. This is another example of inferential statistics. The international study included a large number of people, but it certainly did not include everyone in the world. Yet the WHO statement was a statement that applied to everyone. The WHO made an inference about the entire population of the world, based on the results of the study.

Beginning in Lesson 6, we will learn about the *mathematical* foundations that allow these types of inferences. In this lesson, we concentrate on the methodology used to gather the statistics on which the ultimate conclusions are based.

5.1 – Sampling

Both these reported studies were based on a procedure known as sampling. In each case, there was a **population** the researchers wanted to make a statement about. In the first case the population consisted of all Americans age 18 or above. In the second case the population consisted of all people in the world. In each case it was clearly impossible to base the study on an examination of the entire population. So the researchers chose a **sample**, a subset of the population. They gathered **statistics** for that sample using methods we learned about in Lessons 1-3.

- In the first example, they simply asked the question, "Do you think the amount of taxes you pay is fair?" and recorded the responses, then calculated the proportion of yes answers. This was the statistic reported: 54% answered yes.

- In the second example, they measured two things about the individuals in the study: "Have you used a cell phone for 10 years or more?" and "Do you have this particular type of brain cancer?" They set up a contingency table with *Cell phone use* as the explanatory variable and *Cancer* as the response variable. Just as we did in Lesson 3, they calculated conditional proportions for the two groups, then used division to compare these proportions. The reported statistic was this ratio, which was approximately 2 (as indicated by the phrase "double the rate.")

In both examples, the purpose of the study was to make a conclusion about the larger population. The statistic (the information about the sample) was used to make a statement about a **parameter** (the corresponding piece of information about the entire population). In the first example, the *statistic* was the proportion *of the sample* who believe the amount of taxes they pay is fair. The *parameter* is the proportion *of* ***all*** *Americans age 18 and above* who believe the amount of taxes they pay is fair.

Drawing a conclusion about the entire population based on measurements from a much smaller sample is valid only if the sample is **representative**. For a representative sample, the results from the sample will mirror those for the entire population quite well. As we will see in later lessons, choosing a sample completely at random is one way to obtain a representative sample. The mathematical methods presented in this course are all based on the assumption that the sample is what is called a **simple random sample** or **SRS**. More complicated mathematical methods can be used along with more complicated methods of gathering samples, but we will not consider these more complicated situations in this introductory course.

5.2 – Studying Association

The second example described above is an example of studying association. The researchers ask the question, "Is there an association between using a cell phone and getting brain cancer?" For this type of study, there are two possible strategies.

The first strategy is called an **observational study**. This is the strategy used by the researchers in the reported study. The participants in the study were selected in a random fashion, and the researchers *observed* the values for both the explanatory variable ("cell phone use for 10 years or more") and the response variable ("occurrence of this type of brain cancer.") In this study, they found there was an association. Of course, as we have noted before, this does not necessarily imply a **cause and effect** relationship between the variables. There could be **lurking variables**, that is other variables that explain all or part of the observed effect. A possible lurking variable might, for example, be the level of industrialization where the people live. A well-designed observational study will attempt to take into account the possibility of lurking variables, but the basic fact remains that in general association does not necessarily imply causation.

The second strategy is called an **experiment.** The key difference in an experiment is that the researchers themselves randomly assign the participants to the different groups for the explanatory variable. This removes the possibility of lurking variables affecting the results, and allows the researchers to justifiably claim a cause and effect relationship if they identify an association. However, time constraints and ethical issues may make an experiment difficult or impossible to carry out. In the reported study, to carry out an experiment the researchers would have had to start with people who had never used a cell phone, then split them randomly into two groups. They would have to force the first group to use a cell phone for 10 years starting now, and force the second group not to use one. They would have to wait 10 years to get their results. They would have to live with the possibility that the experiment itself had caused people in the first group to develop cancer. For all these reasons, the researchers used an observational study rather than an experiment.

There are a number of important issues that arise when conducting an experiment. For example, consider an experiment to judge the effectiveness of a new pill for controlling allergies. The experiment will be conducted by randomly assigning individuals to two groups. One group will receive treatment, and the other (called the **control group**) will not. For each group, the researchers measure how well the allergies are being controlled. If they find a difference between the groups, this will suggest an association between the variables ("receiving the treatment" and "having allergies controlled").

There is a problem, however. It has been found that if you give people a pill that does *not* contain the medicine, but they think it *does* contain the medicine, they will show improvement even though they received no medicine. This is called the **placebo effect** (a pill that claims to contain medicine but actually does not is called a **placebo**). Because of this, both groups need to think they are receiving treatment. One group will be given the actual medicine, the other a placebo. The experiment should be **blind**; participants should not know which group they are in. Moreover, if the measuring process involves a subjective judgment, then the persons measuring the result should also not know which group the participants are in (a **double blind** experiment).

Now suppose that the new pill is actually totally useless. The randomness of the situation will cause various levels of apparent control of the allergies among the participants, both in the treated group and in the control group. One would therefore expect to see small differences in the results for the two groups, based merely on the randomness of the situation. It might happen that the treated group has slightly better results than the control group, even though the pill is actually useless. Because of this, researchers must see a reasonably large difference before they will conclude that the pill is effective. When the difference is large enough that it cannot reasonably be attributed to the randomness of the situation, it is described as **statistically significant** or sometimes just as **significant**. What this means is that the difference observed *in the sample* is large enough that the researchers can conclude that it reflects an actual difference *in the entire population*. In later lessons we will learn about the mathematical methods that determine how large the difference must be in order to be statistically significant.

5.3 – What Can Go Wrong?

Statistical studies do not always report correct conclusions. For the question of cell phones and cancer, for example, some studies have reported no apparent association and others have found an association. What this means is that in some studies there was no association for that particular sample – or at least no association large enough to be statistically significant. For other samples and other studies there was a statistically significant association. Which studies are correct, and for the studies that are incorrect, what went wrong?

As we will see starting in Lesson 6, when you take a random sample and use it to make a statement about the population, there is a certain probability that your statement will be incorrect. For a very small sample, the probability of being incorrect is quite large. To see why, imagine estimating the probability of "heads" in a coin toss by taking a sample consisting of two tosses of the coin. There is a 25% chance that you will report a probability of 0 because you got two tails, and a 25% chance you will report a probability of 1 because you got two heads. The total probability of being incorrect is 50%. However, if you toss the coin more and more, the resulting proportion will become a better and better estimate of the actual probability. Likewise, taking a larger sample can make it far less likely that the reported conclusions are incorrect.

The point of the previous paragraph is that the probability of having an incorrect conclusion just by chance can be controlled by using large enough samples. It is possible to calculate just how likely this type of error is, and to make that likelihood as small as you wish to make it.

As a result, by far the most frequent causes for incorrect studies are a variety of ways in which a study can be poorly designed. This may involve a lack of randomness in the selection of the participants, or a flaw in the manner in which the study is carried out. For example, for a survey it is possible that poorly worded questions can influence the participants' answers. Your instructor may refer you to additional resources for more details on the ways in which **bias** may be introduced into a study. But we want to reiterate this: ***the mathematical probability of obtaining incorrect answers in a well-designed study can be made very small; in general, when studies are incorrect it is far more likely to be a flaw in the design of the study itself.*** (Sadly, some studies are intentionally designed to be flawed, in order to prove what the researchers want to prove; but that's a story for another day.)