Topic 1: Design Concepts

Motivation

Is there an association between smoking and lung cancer**?** Does smoking cause lung cancer? Are these questions the same?

Consider a similar example. Is there an association between shoe size (for children ages 6 – 13) and reading ability? Does a larger shoe size cause an increase in reading ability? Are these questions the same? Is there perhaps another reason why those with larger shoe sizes exhibit a higher reading ability?

A sample of 100 male high school seniors competing in a local basketball tournament showed that their average height is 6 feet tall. Does this suggest that if we measured the heights of all male high school seniors that their average height would be around 6 feet tall? Think about where the observations came from. Is the result (from the sample of 100 individuals competing at the basketball tournament) misleading or biased for the population of interest (all male high school seniors)?

Statistics is the science of collecting, organizing, and interpreting sample data. When interpreting the data, there is a fine line between a valid conclusion and a misleading conclusion. In statistics (and science in general) it is crucial that we understand how to interpret the sample data properly, otherwise we may end up with misleading conclusions. The interpretation part of the analysis and the conclusions we can or cannot make depend on how and where the sample of observations came from. First, we must accept that in most scientific studies we are dealing with incomplete information (observations gathered in a sample from a larger population). It is rare that we can ever collect ALL the information needed to make any conclusion with 100% certainty. For example, is it possible to measure the heights of ALL individuals in the population? Most certainly not! When dealing with samples, we have to ask ourselves, where did the sample come from? Also, what method was used to obtain this sample? This first topic will focus solely on the design of the experiment and the resulting conclusions that can and cannot be made depending on how the observations were collected.

Note: The results and conclusions discussed in this chapter are based solely on the study design. They are independent of the numerical outcomes of the study.

Two Questions (or Inferences)

In most scientific studies there are usually two questions we want to address. Can we generalize our results (based on a sample) to a larger population? Can we isolate the cause of the difference in responses between treatments?

Population Inference

The first question is whether or not we can generalize the (observed) results from the study, which are based on a sample, to a larger population (population inference). A study of 100 right-handed adult males showed that their upper right-arm strength was 35% stronger than their upper left-arm strength. Is this true for all individuals (right-handed adult males) in the population?

1. **Population Inference:** A population inference implies that the results obtained from the sample can be ***generalized*** to an entire population.

A population inference would imply that the results obtained from the 100 individuals in the study would apply to all right-handed adult males. In other words, can we conclude that if all right-handed adult males in the population were to be measure, that we would see their right-arm strength to be roughly 35% stronger than their left-arm strength?

When can we make generalizations (population inferences)?

Causal Inference

Another important question is whether or not we can determine or isolate the cause of the observed differences (in the responses) between treatments (causal inference). A study showed that those who had attended regular church services tended to live longer. Does this imply that attending church caused one to live longer? Another study for a new migraine medication compared the decrease in pain for those taking the new drug to the decrease in pain for those taking a placebo. The study showed a significantly higher decrease in pain for those taking the new drug. Does this imply that it was the drug that caused the higher pain relief? Does smoking cause lung cancer?

2. **Causal Inferences:** Consider the two-sample case when we are comparing the responses from two treatments. A causal (or ***cause and effect***) inference implies that the difference in responses is caused by the difference in treatments.

A causal inference in the drug-testing example would imply that the difference in pain relief was caused strictly because of the difference in the treatments (drug vs. placebo). A causal inference in the church study would be the conclusion that one would live longer if they attended church. Could there be another explanation?

When can we make causal inferences?

To answer these questions we must consider how and where the observations were obtained.

Random Mechanism: A random mechanism has two components.

1. Random Selection of Units (Random sampling): The units in the sample are selected randomly from the population. The most basic form of random sampling is called a simple random sample (SRS). An SRS of size *n* from a population is a subset of the population consisting of *n* members selected in such a way that every subset of size *n* is equally likely to be selected. There are other random sampling techniques you may have heard of, such as stratified random sampling, cluster sampling, etc…

Inferences: Inferences to the population (generalizations) can be drawn only if we have random selection from the population of interest.

Discussion: When we have random selection we can extend the results from the sample to the entire population because by randomly selecting the units, we assume the sample will carry the same properties as the population (only in a smaller context).

When we do not have random sampling from a population, conclusions should be restricted to the sample and generalizations to an entire population should be made with caution.

Note: In some cases, population inferences can be made if it can be reasonably assumed that the individuals in the sample are representative of the population. It would be up to the experimenter to decide if the sample is representative.

2. Random Allocation of Units to Groups: Units are randomly assigned to treatment groups. Again, look at comparing our drug to a placebo. If we have a sample of *n* individuals and we want to randomly assign each individual to either the drug or the placebo, we could simply flip a coin and assign each individual to the drug if the flip is heads and to the placebo if the flip is tails. Theoretically, for a large sample size this randomization would result in *close* to half the individuals taking the drug and the remaining (*close to*) half taking the placebo.

Inferences: Statistical inferences of cause-and-effect (causal) relationships can be drawn only if we have random allocation.

Discussion: Since the units are randomly allocated to treatment groups we can conclude that any differences in the response must be due to (caused by) the difference in treatments. Why? The influences of other factors are assumed to have an equal effect on each group equally and thus "cancel out".

When there is no random allocation, causal inferences "should" not be made. The difference in responses could be caused by other (confounding) variables. These variables are known as **confounding** variables, which are related to both group membership and to the outcome. In other words, a variable is confounded if it could have been used to define the treatment groups in a very similar manner.

Note: In some cases, causal conclusions can be reached, when there is no random assignment, if all reasonable confounding factors can be eliminated as possible explanations. This can be achieved by measuring these other variables and excluding them as possible factors, or upon repetition of the study (with similar results) in different situations. See "Does smoking cause lung cancer" at then end of these notes.

Study Designs and Inferences:

1. **Observational Study:** A study in which there is no attempt to influence the response. The response is simply observed on each individual. When comparing responses over different levels of some treatment we can identify observational studies as those where the units are pre-classified into treatment groups.

   For example, if we want to study the difference in some characteristic between males and females, can we randomly assign individuals to either the "male" group or the "female" group? NO! Individuals are already pre-classified into the treatment groups, "male" or "female". This would be considered an observational study where we simply observe responses for males and females.

   For observational studies, **causal inferences** cannot be immediately made because there is no random allocation to treatment groups.

2. **Randomized Experiment:** A study in which treatment(s) are purposely imposed on the experimental units (individuals) to try and influence the response is called an experiment.

   When there are two (or more) treatments and units are randomly assigned to one treatment group we have a randomized experiment. The drug example is an illustration of a randomized experiment. From the selected migraine sufferers in the study, we purposely try to influence their response by giving some of them a drug or a placebo.

   In a randomized experiment causal inferences CAN be made because we have random allocation.

Does smoking cause lung cancer?

Many studies have shown a significant **association** between smoking and lung cancer. There is factual evidence finding that smokers have a higher rate of lung cancer than non-smokers do. However, does this mean that smoking **causes** lung cancer?

Is it possible to design an experiment to prove that smoking causes lung cancer?

Fisher's Hypothesis: The following hypothesis was first introduced by R.A. Fisher

Reference: Fisher R.A.. Cigarettes, cancer and statistics. *Centennial Review* 1958;**2**:151–66 http://www.library.adelaide.edu.au/digitised/fisher/274.pdf

1. Suppose it is found that a certain gene **causes** lung cancer. Thus, if an individual has this gene they are more likely to develop this disease (regardless if they smoke or not).

2. Suppose, it is found that individuals possessing this gene are also much more likely to find smoking a pleasurable experience. As a result, people who have this gene are much more likely to smoke than those who do not have the gene are.

Fisher's hypothesis introduces a possible ***confounding*** variable (the gene) as the explanation for the result. Thus, if this hypothesis were true, it is the gene that causes lung cancer, NOT smoking.

Disclaimer: I am in no way saying that this hypothesis is true, and there is no evidence to indicate that it is. It is only a "what if" example. I am also in no way advocating smoking.

Other Notes:

In many cases randomized designs are impossible (observational studies). In order to prove that smoking causes lung cancer, one could design an experiment starting with non-smokers. He/she could then randomly assign individuals to smoking/non-smoking groups and observe the rate of lung cancer over time. Is this possible? Ethical? So, does this mean causation can never be established? Not necessarily…

Cause and effect conclusions **may** be considered if all other possible causes can be excluded.

There have been a few studies that appear to refute Fisher's hypothesis. One (observational) study was done on a number of monozygotic twins where one smoked and one didn't, and at least one was dead. It was found that of the ones who died first 80% of them were the smokers. This would seem to invalidate Fisher's hypothesis that it is a gene that causes lung cancer.

Causal conclusions are also possible by observing the same result over and over again in potentially different settings. There is indisputable evidence of a strong association between smoking and lung cancer. This result has been discovered over and over again from many different observational studies in many different settings. The combination of all these studies seem to indicate a causal relationship.

Just to make sure we are all on the same track:

Is there an association between shoe size and reading ability (for children ages 6 – 13)?

Yes, those with larger shoe sizes tend to have a higher reading ability.

Does a larger shoe size cause a higher reading ability?

Since this is an observational study we have no random allocation, thus no causal conclusions can be made. Age is a possible confounding variable. That is, age explains why those with larger shoe sizes have higher reading ability.

A sample of 100 male high school seniors competing in a local basketball tournament showed their average height to 6 feet tall.

Does this imply that the average height of all male high school seniors is around 6 feet tall?

Since we do not have random selection from the population of interest we cannot generalize our results.