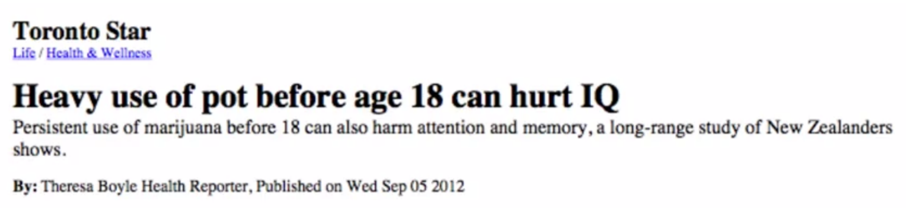
Data Collection

Observational Studies

If we observe an interesting relationship between two variables and we have concluded that it isn't due to chance, we might want to know if changing one of the variables *causes* the other to change. In this section we will begin to investigate whether or not we can answer this question.

Example 1

Figure 1 shows a recent headline that reported on the negative effects of smoking marijuana on the IQ of teenagers.



Toronto Star

Life / Health & Wellness

Heavy use of pot before age 18 can hurt IQ

Persistent use of marijuana before 18 can also harm attention and memory, a long-range study of New Zealanders shows.

By: Theresa Boyle Health Reporter, Published on Wed Sep 05 2012

Figure 1: Toronto Star headline. Source: http://www.thestar.com/life/health_wellness/2012/09/05/heavy_use_of_pot_before_age_18_can_hurt_iq.html

But maybe it says more about those who use marijuana than the actual effects of the plant. For example, socioeconomic status has been blamed as a potential alternate factor contributing to the drop in IQ. If we want to be able to make causal conclusions about one variable causing another we need to control how the data were collected.

We will begin with a general paradigm for a study. For simplicity, we will assume that we are interested in comparing two different groups. There is a particular variable called the **response variable** or **outcome**, that we are interested in comparing between the two groups. Examples of outcome variables include:

- cardiovascular disease risk depending on whether or not one eats egg yolks
- IQ in pot smokers and non pot smokers
- error in age estimation between males and females
- life expectancy across two geographic regions
- infection rates for a vaccinated versus a non vaccinated group

An **explanatory variable** is, as its name suggests, any variable that might explain differences that we observe in the response between the groups. An important concern when planning the study and analyzing the resulting data is the possibility of **confounding variables**. Confounding variables affect the response and have different values in the different groups we are comparing, making it hard to determine what causes the differences between the groups. To avoid having confounding we need to collect the data carefully!

One way we could collect some data is through anecdotes. We often hear anecdotes about how a change in diet or environment has resulted in miraculous improvements to an acquaintance's health. However, due to weak evidence and small sample size, this method is not useful for making any general conclusions.

Two other better methods of data collection are observational studies and experiments. The advantage of an experiment over an observational study is the strength of the causal conclusions we can make. In **observational studies** the data are measurements of existing characteristics of a group or groups of individuals. The goal is either to draw conclusions about the population or about differences between two or more groups or about the relationships between variables. Because existing characteristics are being observed, the investigator has no control over which group one individual belongs to.

In contrast, in an experiment, the investigator imposes an intervention on the individuals being studied, perhaps assigning some to one group and some to another. Experiments are the gold standard for making causal conclusions, but for now we will focus more on potential issues with observational studies.

Example 2
The study we will consider was published in the New England Journal of Medicine in May, 2012 and it concerned the relationship between coffee drinking and mortality. Headlines in the New York Times and the Toronto Star reported that coffee may help us live longer. The Washington Post, in contrast, was a little more reserved. So should we be pausing now for a coffee break?

Here are a few details about the study.

- In 1995, the researchers collected data on over 400,000 men and women age 50 to 71
- Patients were followed until 2008
- By the end of the study, over 52,000 patients had died
- Research question: "Did the amount of coffee these people drink affect how much longer they lived"

This study is an example of an observational study. Subjects weren't told how much coffee to drink, they just did what they would do otherwise. In all observational studies we have to be careful about interpreting our observed associations. In this study, our explanatory variable is the amount of coffee drank and the outcome is how much longer the subjects lived. Suppose the coffee drinkers tended to live longer so there's a positive association. It may be tempting to say coffee causes long life but there are other possible reasons why an association between coffee and longevity may have been observed.

Mechanisms that can result in an observed association between the outcome and explanatory variable in an observational study are:

1. **Causation**: changes in the explanatory variable cause the outcome to change.
2. **Reverse causation**: changes in the outcome cause the explanatory variable to change.
3. **Association**: the relationship between the explanatory variable and the outcome is a coincidence.
4. **Common cause**: causes both the explanatory variable and the outcome to change.
5. **Confounding variable**: is associated with the explanatory variable, making it impossible to know whether to attribute changes in the outcome to the explanatory variable or to the confounding variable.

In the coffee study, whether a subject had diabetes is a possible common cause. Perhaps because of dietary restrictions, diabetics drink less coffee and diabetics also tend to have shorter lives. One of the many possible confounders could be smoking. Smokers tend to drink more coffee so smoking status is associated with our explanatory variable, and smoking also has a causal affect on our outcome length of life.

In Simpson's paradox, we saw that there are also variables that can be lurking or hidden variables. A **lurking variable** is a variable that is not accounted for in the analysis but may affect the nature of the relationship between the explanatory variable and the outcome. Lurking variables can be confounding variables, or the source of a common response, or another variable that, when considered, changes the nature of the association. In the coffee study, method of preparation is a possible lurking variable. We can't say if differences in the composition of espresso versus filtered coffee affect the protective effects of coffee, thus influencing the relationship between coffee drinking and length of life.

Controlling for confounding variables is a particularly difficult part of the analysis of any statistical study. In the next section, we will talk about the use of experiments to mitigate the possibility of confounders.