Summary (Unit 2)

In this unit, we discussed the first step in the big picture of statistics — **production of data.**

Production of data happens in two stages: **sampling** and **study design**.

Our goal in sampling is to get a **sample that represents the population of interest well**, so that when we get to the inference stage, making conclusions based on this sample about the entire population will make sense.

We discussed several biased sampling plans, but also introduced the "family" of probability sampling plans, the simplest of which is the **simple random sample**, that (at least in theory) are supposed to provide a sample that is not subject to any biases.

In the section on study design, we introduced 3 types of design: observational study, controlled experiment, and sample survey.

We distinguished among different types of studies and learned the details of each type of study design. By doing so, we also expanded our understanding of the issue of establishing causation that was first discussed in the previous unit of the course. In the Exploratory Data Analysis unit, we learned that in general, association does not imply causation, due to the fact that lurking variables might be responsible for the association we observe, which means we cannot establish that there is a causal relationship between our "explanatory" variable and our response variable.

In this unit, we completed the causation puzzle by learning under what circumstances an observed association between variables CAN be interpreted as causation.

We saw that in observational studies, the best we can do is to control for what we think might be potential lurking variables, but we can never be sure that there aren't any others that we didn't anticipate. Therefore, we can come closer to establishing causation, but never really establish it.

The only way we can, at least in theory, eliminate the effect of (or control for) ALL lurking variables is by conducting a randomized controlled experiment, in which subjects are randomly assigned to one of the treatment groups. Only in this case can we interpret an observed association as causation.

Obviously, due to ethical or other practical reasons, not every study can be conducted as a randomized experiment. Where possible, however, a double-blind randomized controlled experiment is about the best study design we can use.

Another very common study design is the survey. While a survey is a special kind of observational study, it really is treated as a separate design, since it is so common and is the type of study that the general public is most often exposed to (polls). It is important that we be aware of the fact that the wording, ordering, or type of questions asked in a poll could have an impact on the response. In order for a survey's results to be reliable, these issues should be carefully considered when the survey is designed.

We saw that with **observational studies** it is **difficult to establish** convincing evidence of a **causal relationship**, because of lack of control over outside variables (called lurking variables). Other pitfalls that may arise are that individuals' behaviors may be affected if they know they are participating in an observational study, and that individuals' memories may be faulty if they are asked to recall information from the past.

Experiments allow researchers to take control of lurking variables by **randomized assignment to treatments**, which helps provide more convincing evidence of causation. The design may be enhanced by making sure that subjects and/or researchers are **blind** to who receives what treatment. Depending on what relationship is being researched, it may be difficult to design an experiment whose setting is realistic enough that we can safely generalize the conclusions to real life.

Another reason that observational studies are utilized rather than experiments is that certain explanatory variables — such as income or alcohol intake — either cannot or should not be controlled by researchers.

Sample surveys are occasionally used to examine relationships, but often they assess values of many separate variables, such as respondents' **opinions** on various matters. Survey questions should be designed carefully, in order to ensure unbiased assessment of the variables' values.

Throughout this unit, we established guidelines for the ideal production of data, which should be held as standards to strive for. Realistically, however, it is rarely possible to carry out a study which is completely free of flaws. Therefore, common sense must frequently be applied in order to decide which imperfections we can live with, and which ones could completely undermine a study's results.

> **(Optional) Outside Reading:** Little Handbook – Design & Sampling (one long & one short)
