Refining Your Research Question

You probably already have a general research question in mind, such as

- Why are some baseball players paid more than others?
- What factors influence housing prices?
- What makes students attend or skip lectures?

Before you can start to analyse your data to find answers, you need to refine your general question into a specific question phrased in terms of the data you have collected. In particular:

1. **You can only answer a question about the population from which your experimental units were drawn**.

- *What is a population and a sample?* A population is any entire collection of people, animals, plants or things that we are interested in studying. It is the entire group we are interested in, which we wish to describe or draw conclusions about. The idea of statistics is instead of collecting data from the entire population, which could be costly, lengthy or impossible; we instead collect data from a sample, which in most cases can be thought of as some collection of members of the population. For each population there are many possible samples. See ‘sample’ in Glossary section for information regarding samples. It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included. For example, a researcher may be interested in the trunk length of living African elephants. It is almost impossible to measure the trunk length of all the African elephants, so they may have a sample of 30 elephant trunks. So here the sample is the 30 elephants but the population is all living African elephants.

- *What are your experimental units?* Suppose you are studying what factors influence housing prices and you are using the dataset ‘house prices’ discussed in the second example on the page on understanding your dataset. The experimental units in this case are towns, not individual houses. So you will only be able to answer a question about what aspects of a *town* tend to make its average house prices higher or lower, and not any questions about what makes a particular *house* more or less expensive.

- *Who or what might your data collection method have missed?* Suppose you are interested in how to get more students to attend lectures. If you have collected data from students through a survey distributed in a lecture, you are likely only to have gotten information from students who already usually attend, and not from the students who do not. Thus the questions you could answer from that dataset are about why some students *do* attend lecture.

2. **You can only answer a question that relates to the data you have collected about these experimental units.**

- *What data have you collected?* You cannot answer a question about how the age of a baseball player influences his salary from the baseball dataset, because the age of the players is not recorded there.

- *How does the data you have relate to the data you really want?* Suppose that on your student survey you ask, “How many times this semester have you skipped lecture?” Students may not remember exactly, or they may not answer honestly. Thus you can only answer a question about how often students *report* skipping lectures, and not about how often they actually skip them.

3. **A specific research question should take the form of a question about the relationships among the variables you have measured on your sample.**

Often, this will take the form of a question about how one or more *predictor variables* relate to a *response variable,* and what aspect of the *response variable* are you interested in. For example are you interested in the average of your response variable, i.e. the mean or the median. Or are you interested in the spread of your response variable such as the standard deviation, interquartile range or variance.

Examples of questions that involve a single predictor variable and a single response variable:

- Do towns that border on the Charles River have higher or lower average house prices than towns that do not?

  Here the predictor variable is “borders on the Charles River” (yes or no) and the response variable is “house price”, and we are interested in the average of the response variable.

- How does a baseball hitter’s number of homeruns in the season relate to his salary?

  Here the predictor variable is “number of homeruns in the season” and the response variable is “salary”.

Examples of questions that involve more than one predictor variable and a single response variable:

- How do the distance to Boston employment areas and accessibility of radial highways influence average housing prices in a town?

- How do the socioeconomic factors crime rate, pupil-teacher ratio and percent lower status in the population relate to house prices in a town?

4. **You must decide how many variables you can consider in one question.**

- *If you are interested in how more than one predictor variable relates to a given response, you get more information by studying the predictor variables all together than by studying each one independently.*

One reason for this is that there may be *interactions* between the predictors that you can only see by considering them all together. For instance, it may be that housing prices in a town are higher if the town is on the river *unless* the town is heavily industrialised, in which case being by the river lowers prices. That is, the effect of the first predictor, being by the river, might depend on the second predictor about amount of industry.

- *The more variables you want to study together, the more complicated the analysis will be.*

  There are three levels of statistical analysis. From simplest to most complicated, they are:

  1. Summary statistics and graphical methods,
  2. Testing for significance (getting p-values),
  3. Model building.

If you have more than three variables related to your research question, the only technique that will work is model building. So you may consider how to frame your research questions with this in mind.

- *The more variables included in your research question, the larger the sample size you need to get reliable results.*

See the sheet “sample size and models” for more on this.

5. **You cannot generally answer questions about causes unless you have done a controlled experiment.**

A controlled experiment is a study in which experimental units have been put at random into either an experimental or control group and those in the experimental group are given some sort of intervention, whereas those in the control group are given no intervention or some placebo.

None of the examples discussed above are experiments (baseball data, housing prices and student survey) because no intervention was given.

An example of an experiment would be if students were randomly assigned to two groups, and one group was given lecture notes and the other group was not, then you observed lecture attendance for the members of both groups. Then you could ask if giving students lecture notes caused them to skip lectures.