What is research?

"In the broadest sense of the word, the definition of research includes any gathering of data, information, and facts for the advancement of knowledge."
- Martyn Shuttleworth

"Research is a process of steps used to collect and analyze information to increase our understanding of a topic or issue."
- J.W. Creswell

Source: http://en.wikipedia.org/wiki/Research

The research process can be conceptualized as follows.

FORMULATING THE RESEARCH HYPOTHESIS

Most studies are conducted because researchers want to learn about the relationship between two or more variables in a specific population. When a research question (or hypothesis) is developed, the following should be evident:

- The variables under consideration should be clearly defined
- The population being studied should be clearly identified

In quantitative studies, the research question typically takes on one of the following formats:
<table>
<thead>
<tr>
<th>Type</th>
<th>Example Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlational</td>
<td>(1) Is there a relationship between adoption and suicide risk in young adults?</td>
</tr>
<tr>
<td></td>
<td>(2) Is eating breakfast cereal in the morning associated with healthier weights in children?</td>
</tr>
<tr>
<td>Experimental</td>
<td>(3) What is the difference in tumor incidence rates between mice exposed to tobacco smoke vs. exposed to clean air?</td>
</tr>
<tr>
<td></td>
<td>(4) Does acupuncture improve symptoms of fibromyalgia? More specifically, are the reduction in Fibromyalgia Impact Questionnaire (FIQ) scores greater for those patients undergoing actual acupuncture vs. simulated acupuncture?</td>
</tr>
</tbody>
</table>

**VARIABLES DEFINED IN THE RESEARCH HYPOTHESIS**

The above examples each involve both an *explanatory* and a *response* variable.

**Variable Roles**

The *explanatory* variable (also called the *independent* variable) is one that may explain or cause changes in a *response* variable (also called the *dependent* variable). The response variable measures the outcome or result of a study.

**Exercises:**

1. For each of the example research questions listed above, identify both the explanatory and response variables of interest.

<table>
<thead>
<tr>
<th>Question</th>
<th>Explanatory Variable</th>
<th>Response Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Consider the following abstract.

Identify both the explanatory and response variables of interest in this study.

Explanatory variable: _____________________________________

Response variable: _____________________________________

3. Consider the following abstract.

Identify both the explanatory and response variables of interest in this study.

Explanatory variable: _____________________________________

Response variable: _____________________________________

Note that variables are not inherently dependent or independent – this depends on what role they play in the study. Also, there is no restriction on the number of variables that can be considered in a study, though a research question should never be unnecessarily complex.
INTRODUCTION TO STUDY DESIGNS

In a typical statistics course, the majority of the course outline focuses on data analysis techniques. It is extremely important, however, that a statistician (or a data scientist) also have a sound understanding of study design.

“There is no point to analyzing data from a study that was not properly designed to answer the research question under investigation. In fact, there’s a real point in refusing to analyze such data lest faulty results be responsible for implementing a program or policy contrary to what’s really needed.”

-- Gerard E. Dallal

The remainder of this handout will provide an overview of various study designs and will also discuss their advantages/disadvantages.

First, note that there are two basic approaches for designing a study to investigate whether there is a relationship between an explanatory variable and a response variable.

### Broad Classifications of Study Designs

<table>
<thead>
<tr>
<th>Classifications</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>The researcher purposefully manipulates the values of the explanatory variable(s) and measures the effect of the manipulation on the outcome of interest (i.e., the response variable).</td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td>The researcher simply observes information on the outcome of interest (the response) and makes comparisons across the values of the explanatory variable. There is no specific intervention in an observational study.</td>
</tr>
</tbody>
</table>

Exercise: Classify each of the following examples as either an experimental or observational study design.

<table>
<thead>
<tr>
<th>Example</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>A study was conducted to see if those with high blood pressure (above 164/89 mmHg) did worse on tests of memory than those with normal blood pressure.</td>
<td></td>
</tr>
<tr>
<td>An examination of the medical records of Swedish men was conducted to investigate whether those who were overweight had a higher risk of kidney cancer.</td>
<td></td>
</tr>
<tr>
<td>A study was conducted to see if patients with bipolar disorder who were given a high dose of omega-3 fats from fish oil saw improved symptoms more so than those given a placebo.</td>
<td></td>
</tr>
</tbody>
</table>
As mentioned, experiments involve the researchers manipulating something and then measuring the effect of that manipulation on some outcome of interest. This manipulation involves what are called treatments. A treatment is a specific experimental condition applied to the subjects. Treatments are defined by the different values (or combinations of values) of the explanatory variable(s). When subjects are randomly assigned to one treatment or another, the study is referred to as a randomized controlled experiment.

Confounding Variables and the Importance of Randomized Controlled Experiments

To understand why this randomization of treatments is important, consider the following definitions.

<table>
<thead>
<tr>
<th>Extraneous Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>A confounding variable both affects the response variable and is also related to the explanatory variable. If a variable is a confounding variable, then its effect cannot be separated from the effect of the explanatory variable.</td>
</tr>
<tr>
<td>A lurking variable is a potential confounding variable that is not measured and is not at all considered in the study.</td>
</tr>
</tbody>
</table>

Confounding might make it appear as though a relationship exists between the explanatory and response variable when this is in reality not the case. For example, consider the following case studies.

<table>
<thead>
<tr>
<th>Case Study: Breakfast Cereal and Healthy Weight in Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 2012 study was conducted to investigate the relationship between eating breakfast cereal and weights of children. One summary of this study was headlined, “Breakfast Cereals Prevent Overweight in Children.” The article claims that based on this study, we can conclude that regularly eating cereal for breakfast every day (even super-sugary cereals!) is tied to healthy weights for children. Source: worldhealthme.blogspot.com.</td>
</tr>
</tbody>
</table>
Case Study: Breakfast Cereal and Healthy Weight in Children

Questions:

1. Can you think of other potential confounding variables related to eating breakfast cereal that might also affect whether or not a child maintains a healthy weight?

![Diagram showing the relationship between eating breakfast cereal and having a healthy weight, with confounding variables depicted as ellipses between the two major variables.]

2. Note that the appearance of an association between eating breakfast cereal and whether or not a healthy weight is maintained could potentially be due to these confounding variables. Given this information, do you think it is fair for the aforementioned headline to use language such as “breakfast cereals prevent...”? Why or why not?
Case Study: Adoption and Suicide Risk

In September of 2013, researchers from the University of Minnesota published a study in the journal *Pediatrics*. This study was described in a *Fox News* article titled “Adopted teens may be at higher risk of suicide.” This article described the results as follows: [The researchers] examined data from an existing University of Minnesota study of 692 adopted children and 540 non-adopted siblings in Minnesota… All of the adopted kids, who were between 11 and 21 years old during the study period, had been taken in by their families before age two. Almost three quarters of the adopted children were born abroad, most of the foreign-born children were from South Korea and 60 percent of those were girls. At the beginning of the study, and again about three years later, the researchers asked participating families if the children had made a suicide attempt. Over the three years of the study, 56 children attempted suicide at least once, according to the family members’ reports. Of those kids, 47 were adopted and nine were not adopted.

Questions:

1. Can you think of any other potential confounding variables related to adoption that might also affect suicide attempts?

2. The following quote was taken from the *Fox News* article ([foxnews.com](http://foxnews.com)).

   *When previous self-harm behavior was taken into account, researchers calculated that adopted teens were 3.7 times more likely to attempt suicide than the other teens… When the researchers adjusted for other factors often linked with suicidal thinking or behavior, including drug use, depression, academic struggles and personality traits like alienation and impulsivity, the increased risk for adopted kids remained…*

   *Other mediating factors, not considered in our study, may include: heritable risk, prenatal factors, factors unique to relinquishment by a biological parent, early trauma, weak attachment to adoptive families and loss of cultural identity and ethnic discrimination,” Keyes told Reuters Health by email.*

   What confounding variables were considered by the researchers in their analysis? What other lurking variables did the researchers identify?
Case Study: Adoption and Suicide Risk

3. Below are actual headlines used to publicize the results of this study. Do you think these headlines are fair, or are they misleading? Explain your reasoning.

- “Adopted teens may be at higher risk of suicide” (foxnews.com)
- “Adopted teens face high suicide risk” (medpagetoday.com)
- “Adoptees four times more likely to attempt suicide” (medscape.com)

4. Write a headline to publicize the results of this study that is purposefully misleading.
Back to Why Randomization of Treatments is Important

When an observational study indicates that there is an association between the two variables under study, it is quite possible that lurking variables may be affecting the association. For example, consider the cereal study. Recall that those who ate cereal for breakfast tended to be at healthier weights. Does the cereal cause this effect? Not necessarily! It also seems reasonable to conclude that students with high physical activity levels will be hungry in the morning and will eat more cereal; moreover, students with a high physical activity level will be at a healthier weight. This set of relationships could be the real reason for the study’s results!

Now, suppose that the study had been conducted differently. Instead of simply observing the children’s habits at breakfast, suppose that the researchers randomly assigned half of the children to eat breakfast cereal every day and the other half to not eat breakfast cereal. Now, what would we conclude if the study results were similar? That is, what if we still saw that those who tended to eat cereal for breakfast tended to be at healthier weights? Now, with a randomized controlled experiment, we wouldn’t be able to argue that the results were attributable to physical activity level. The random assignment of subjects to experimental conditions should have taken care of this – theoretically, half of the students with high physical activity levels would have been eating breakfast cereal, and the other half would not.

Because of the random assignment, the groups of subjects receiving the different experimental conditions are theoretically balanced on all other variables. The only way the two groups should differ is in whether or not they’re eating cereal every day (the treatment intervention). So, even though lurking variables are present in all studies, randomized controlled experiments account for lurking variables by balancing out their effects between the different treatment groups. If it had been a designed experiment that had indicated that kids who ate breakfast cereal every day tended to have healthier weights, we would have been much more certain that we had uncovered a causal relationship.

The Big Idea: Observational studies allow us to conclude only that an association exists. The best method for determining causal relationships is to conduct a randomized controlled experiment!

Note that it is unlikely that any explanatory variable provides the sole explanation for changes in the response variable. Confounding variables are almost always present, and only sometimes are they measured and accounted for in the study. You should always think about the possible effects of confounding variables (especially with observational studies).
Case Study: ECT and Gain in Functional Status

A recent study investigated whether the number of electroconvulsive therapy (ECT) treatments was associated with a gain in functional status for patients hospitalized because of severe depression. The study methodology was described as follows: “The primary study data were collected using a retrospective chart review completed at a large Midwestern inpatient and outpatient psychiatric hospital setting in an urban area with a population of more than 100,000. Records included 278 psychiatric inpatient and outpatient admissions, with some receiving ECT treatments, after April 1, 2005, and where discharged before April 30, 2009. The use of these records was obtained from individuals authorized to provide consent.”

The researchers initially posed the following research question: “Does the implementation of a larger number of ECT treatments lead to a larger gain in functional status for these patients?”

A summary of the data is shown below.

They carried out a linear regression analysis and arrived at the following conclusion: “The results of this study indicate that a larger number of ECT treatments leads to a larger gain in functional status for patients hospitalized with severe depression (p = .0044).”

Questions:

1. Is the study design observational or experimental? Explain your reasoning.

2. Is the research question framed in a correlational or experimental format? Is this acceptable? Explain why or why not.

3. What do you think of the researchers’ conclusion? *Hint: think of potential confounding variables!* Do you agree or disagree with their statement, and why?
Given the previous discussion, you might be wondering why we ever do anything but experimental research. Clearly, a randomized experiment provides stronger evidence of a cause-and-effect relationship than does an observational study. When is an observational study preferable (or required)? Consider the following.

Questions:

1. Suppose you want to conduct a study to determine whether smoking during pregnancy is associated with lower birth weight of the baby. Which type of study design should you choose? Explain your reasoning.

2. Suppose you’re interested in investigating whether women are more likely to suffer depression than are men. What type of study design would you use? Explain your reasoning.

Finally, consider studies that have investigated whether smoking causes lung cancer.

Questions:

1. Are the studies that investigate this question observational studies or designed experiments? Explain.

2. Given your answer to the previous question, one could argue that we can’t definitively establish causation. Why is it, then, that we as a society overwhelmingly believe that smoking does in fact cause lung cancer?