Conceptual Foundations of Inference

Statistical Sleuth Chapter 1

Introduction and Goals

The goals for this lab are:

1. to remind you of some definitions for standard terms related to statistical inference (you will hopefully have seen most of these in your first statistics course, but we may give more careful definitions here);
2. to introduce you to a randomization test for inference about the difference between two means (I don't expect that you will have seen this before);
3. to introduce you to the use of R in the RStudio server (most of you have not seen this before, but a few have).

Example Study

- Example from T. Amabile, "Motivation and Creativity: Effects of Motivational Orientation on Creative Writers," *Journal of Personality and Social Psychology* 48(2).
- **Study design**:

47 subjects randomly assigned to one of two groups:

1. "Intrinsic" group; 24 subjects asked to fill out this questionarre, then write a haiku about laughter:

> *INSTRUCTIONS:* Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).
>
> __ You get a lot of pleasure out of reading something good that you have written.
> __ You enjoy the opportunity for self-expression.
> __ You achieve new insights through your writing.
> __ You derive satisfaction from expressing yourself clearly and eloquently.
> __ You feel relaxed when writing.
> __ You like to play with words.
> __ You enjoy becoming involved with ideas, characters, events, and images in your writing.

2. "Extrinsic" group; 23 subjects asked to fill out this questionairre, then write a haiku about laughter:

> *INSTRUCTIONS:* Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).
>
> __ You realize that, with the introduction of dozens of magazines every year, the market for free-lance writing is constantly expanding.
> __ You want your writing teachers to be favorably impressed with your writing talent.
> __ You have heard of cases where one best-selling novel or collection of poems has made the author financially secure.
> __ You enjoy public recognition of your work.
> __ You know that many of the best jobs available require good writing skills.
> __ You know that writing ability is one of the major criteria for acceptance into graduate school.
> __ Your teachers and parents have encouraged you to go into writing.

The haiku were then rated by 12 poets on a 40 point scale for creativity, and the average score across the 12 judges calculated.

- **Research Question**: Is a subject's creativity score affected by the type of motivation (intrinsic or extrinsic) induced by the questionairre?

A first look at the data

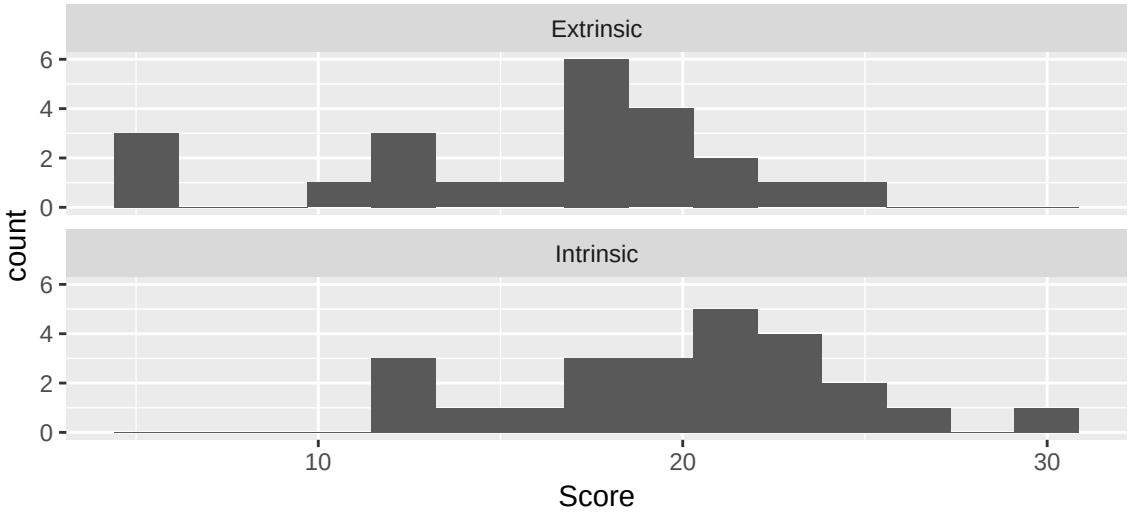
```
# load packages we will need: these provide extra functionality to R
library(ggplot2) # for making plots
library(dplyr) # for data manipulation

# read the data set into R, store it in a data frame called "creativity"
creativity <- read.csv("http://www.evanlray.com/data/sleuth3/case0101_creativity.csv")

# how many observational units (rows) and variables (columns) in the data set?
dim(creativity)
```

```
## [1] 47  3
```

```
# display the data for the first 5 observational units
head(creativity, n = 5)
```

```
##   X Score Treatment
## 1 1   5.0 Extrinsic
## 2 2   5.4 Extrinsic
## 3 3   6.1 Extrinsic
## 4 4  10.9 Extrinsic
## 5 5  11.8 Extrinsic
```

```
# make a histogram to summarize scores within each group
# Score goes on the x axis, with a separate facet (panel) for each Treatment
ggplot(data = creativity, mapping = aes(x = Score)) +
  geom_histogram(bins = 15) +
  facet_wrap( ~ Treatment, ncol = 1)
```




```
# group the data by which treatment was applied,
# then summarize it by finding the mean and standard deviation separately for each group
creativity %>%
  group_by(Treatment) %>%
  summarize(
    mean_score = mean(Score),
    sd_score = sd(Score)
  )
```

```
## # A tibble: 2 x 3
##   Treatment mean_score sd_score
##   <fct>          <dbl>    <dbl>
## 1 Extrinsic       15.7     5.25
## 2 Intrinsic       19.9     4.44
```

Populations and Samples

- **Population**: A defined group of people or things we want to learn about.

In this example, the population is...

- **Sample**: A smaller group of people or things that are selected for a study. These are the people or things we record data about.

In this example, the sample is...

- **Observational Unit**: An individual person or thing in a sample.

In this example, the observational unit is...

- **Types of Samples**:
  - **Randomized**: People selected from the population randomly
    - Simplest type is a **simple random sample**: All groups of n observational units are equally likely to be selected as the sample.
    - (See book for others)
  - **Convenience Sample**: We picked a few people because they were easy to pick.

In this example, the sample is...

Types of Variables, Experiments vs. Observational Studies

- In statistics, a **variable** is a quantity that is measured about each observational unit in the study.

In this example, the variables are...

- **Explanatory** and **Response** (or **Outcome**) variable: Does changing the value of the explanatory variable cause the value of the response to change?

In this example, the explanatory variable is...

In this example, the response (or outcome) variable is...

- **Confounding variable**: Related to both the explanatory variable and the response. Its presence makes it hard to establish the outcome as being a direct consequence of explanatory.

In this example, some examples of potential confounding variables are...

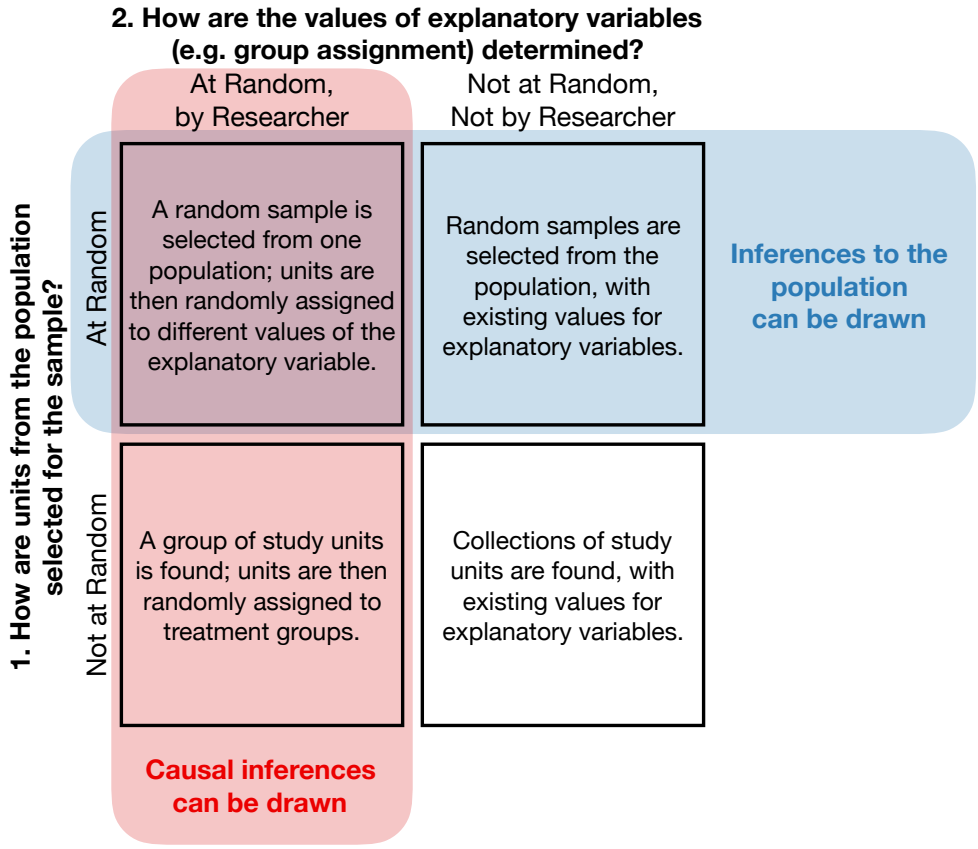
- Types of studies:
  - **Randomized experiment**: The investigator controls the assignment of experimental units to groups and uses a chance mechanism (like the flip of a coin) to make the assignment.
  - **Observational study**: Assignment of units to groups is not in the control of the researcher.

In this example, the study is (randomized experiment or observational study?)...

Scope of Conclusions

Two *separate* questions:

1. Can we make inferences about the population (i.e., can we say something about what's going on in the population based on what we see in our sample)?
    - If the sample was selected randomly from the population, then **yes**
    - If the sample was not selected randomly from the population, then **no**
2. Can we make claims about a causal relationship between the explanatory and response variables?
    - If the values of the explanatory variable were assigned randomly by the researcher, then **yes**
    - If the values of the explanatory variable were not assigned randomly, then **no**



In this example, what is the scope of conclusions we might be able to draw?

Parameters and Statistics

- A **parameter** is an unknown numeric value that describes a feature of a probability model.
  - In intro stats: "a number summarizing values in the population we are studying."

In this example, what are some relevant parameters?

- A **statistic** is a number that can be calculated from the data in a sample.

In this example, what statistics might we use to estimate the unknown parameters? What are the values of these statistics for our sample?

Hypothesis Test Set-Up

A hypothesis test is one way to answer the question "do the data provide evidence that there is a difference in creativity scores for people who see the intrinsic or extrinsic motivation questions?"

- The **null hypothesis** typically states that nothing interesting is going on.

In this example, what is the null hypothesis? State it in a written sentence (in context) and using an equation involving the parameter(s).

- The **alternative hypothesis** states that something interesting is going on.

In this example, what is the alternative hypothesis? State it in a written sentence (in context) and using an equation involving the parameter(s).

Finding a p-value for the test

- The **p-value** is the probability of obtaining a test statistic as extreme or more extreme than the value of the statistic we obtained for our sample, if the null hypothesis is correct.
  - In this sample, the difference in means was 19.883 - 15.739 = 4.144. The p-value is the probability of obtaining a sample difference in means at least this large, if the overall average difference in means is 0.
  - A small p-value means that it would be unlikely to obtain a sample statistic as extreme as the one we got if the null hypothesis is true, so is evidence against the null hypothesis.
  - A calculation of the p-value will be based on the sampling distribution of the statistic, working under the assumption that the null hypothesis is true.
- The **sampling distribution** of a statistic is the distribution of values for that statistic that we could obtain from all possible samples of size n.

A simulation-based calculation (by hand)

To find the sampling distribution, we will simulate (artificially re-create) the assignment of creativity scores to the intrinsic and extrinsic groups over and over again, assuming that each person in the study was equally likely to be in either of those groups. For each of these artificially generated samples, we will calculate the difference in means for the two groups. This will let us see how the results we actually observed in our sample compare with the kinds of results we might have observed if the intrinsic/extrinsic group assignment had no relationship to creativity scores.

Throughout this procedure, let's assume that in every replication of the experiment, the researchers assign 24 participants to the "intrinsic" group and 23 cases to the "extrinsic" group.

One Simulation

Each group has 47 slips of paper with the creativity scores for each person in our sample.

1. Shuffle your slips of paper well.
2. Deal them into two stacks: one with 24 papers (the intrinsic group) and one with the other 23 (the extrinsic group)
3. Calculate the mean score for your intrinsic group:

4. Calculate the mean score for your extrinsic group:

5. Calculate the difference in scores: (intrinsic group mean) - (extrinsic group mean)

A second simulation

Repeat steps 1 through 5 again to generate a second random assignment of our subjects to the two groups, find the means for each group, and the difference in means.

Combining our results on the board

Please add your group's results to the board. Does it appear to be likely to obtain a sample difference in group means as large as 4.144 if in fact group assignment does not affect average creativity scores?

Simulation in R

Doing that by hand was too much trouble, and resulted in too few samples to get a really good sense of the sampling distribution. Let's do this again using R to make things easier.

Your goal for today is just to sign in and run a few commands in RStudio. I won't ask you to write any R code of your own, and you don't necessarily need to understand how these commands work for now. We will talk about them more in future classes.

1. Open a web browser and go to rstudio.mtholyoke.edu
2. Your user name is your email address, **without the `@mtholyoke.edu`** part, and your password is your MHC password.
3. Click on `File`, `New Project`; `Version Control`; `Git`
4. Type in the repository URL: https://github.com/mhc-stat242-s2019/Lab0.git
5. Press the Tab key to fill in the project directory name
6. Change the "Create project as subdirectory of" value from "~" (home folder) to "~/stat242" by clicking "Browse" then creating a "stat242" folder and clicking "Choose". (I recommend this as a place to put all of your github repos for this class)
7. Click on `Create Project`
8. The first time you connect to github from the RStudio server, it may prompt you with a forbidding "authenticity can't be established" message. You should type "yes" into the box to trust this host and click "Okay".
9. Click on the file named `lab00_intro.Rmd` in the File browser pane on the lower right corner of the screen.

That will guide you through the document and the process of running code to do the simulations above.