STAT 305: Lecture 4

Chapter 2: Data Collection

and

Chapter 3: Elementary Descriptive Statistics

Course page: imouzon.github.io/stat305

Recap

Section 2.3: Principles of Effective Experimentation

Recap

Terminology

We described an ideal simple experiment and defined a few associated terms (2.3.1, 2.3.2)

- **Managed variable**: variables where we choose the value
  - **Controlled variable**: a managed variable that only takes one value throughout our experiment.
  - **Experimental variable**: a managed variable taking different values for different runs.
- **Response variable**: the output value; the variables whose values we wish to effect using our experimental variables
- **Concomitant variable**: Variables that we record but are not of interest.
- **Extraneous variable**: All other changes in our system.

Comparitive Studies

We discussed the importance of **Comparative studies** (2.3.3)

- def: a study in which the goal is to compare two or more approaches/methods/ideas/etc.
- As always, there are many things that we can not control when we make measurements and collect data
- When we want to compare a new method to a old method, we need to be aware that the uncontrolled circumstances that existed when we first studied the old method do not exist anymore
- Take away: in order to *know* that difference in results are due to the difference in the method used (and not the difference in the uncontrolled circumstances) we must collect new data on both methods - this way, both methods be studied under the same uncontrolled circumstances.

Techniques for Dealing with Extraneous Variables

Some aspects of the environment can not be controlled, even though they may effect the results we see. We call those aspects **extraneous variables**.

Though we can not *control* extraneous variables, we can plan our experiment to minimize their impact (2.3.2, 2.3.4)

Note: These techniques are part of how the experiment is *designed* - we decide this before any data is collected.

Technique 1: Blocking

- Designing the experiment with homogeneous mini-environments.
- In this way, regardless of the random/uncontrolled events that occur in the mini-environment, we know every observation experienced the same event.
- When collecting the data, we record the identity of the block used. Since the block used will have different values depending on the observational unit, we call the block identifier a "blocking variable"

Techniques for Dealing with Extraneous Variables

Technique 2: Randomization

- Using random assignment at all possible chances to "average out" the systematic changes that occur as we perform each run of our experiment.
- Used to choose the assignment of order, location, worker, partners, etc.

Techniques for Dealing with Extraneous Variables

Technique 2: Randomization, continued

Example: A chemist performs 20 runs of a synthesis, 10 using substrate A and 10 using substrate B. It is believed that the chemist could become more adept with each run.

Attempt 1: No randomization on order:

- **Plan**: the chemist performs all 10 syntheses using A and then performs all 10 syntheses using B.
- **Impact**: we have a changing extraneous variable ("adeptness") that will benefit the last runs
- **Result**: we can't tell the difference between whether our results change based adeptness or substrate choice

Attempt 2: Order chosen using randomization

- **Plan**: the chemist using a system to randomize the order of the 20 runs
- **Impact**: the chemist's change in adeptness will not benefit only one substrate
- **Result**: everything's ok!

Dealing With Extraneous Variables, cont.

Technique 3: Replication

- def: the process of repeating a run of the experiment more than once for each combination of experimental variables.
- results on the first run should be similar to the results on the second run if no experimental variables are changed - changes are the result of run-specific extraneous variables.
- after repeating multiple times, impact of run-specific effects average out
- neat: replication is strongly connected to the concept of reproducibility - the results I get should be similar to the results you get

In Summary

The main point is this: an effective experiment is **designed** to account for the environmental conditions that could influence our response. Doing this takes a lot of planning.

Section 2.4

Some Common Experimental Plans

Common Experimental Plans

Designing Experiments

Experiments don't just happen, they have to be planned out ahead of time.

With **replication**, **randomization**, **blocking** we have introduced some common techniques used in planning (or **designing**) an experiment.

Common combinations of these techniques can be used as frameworks around which we can build our real-life experiments.

Some of these combinations have become so useful that we named them to make explaining the experimental design easier.

Design 1: Completely Randomized Experiments

Description

- All experimental variables are of primary interest (no controls, no blocking)
- Randomization used at every possible point where we choose how to treat the experimental units

Examples

- Example 12 (page 51): Golf ball drive distances study by G.Gronberg
  - 10 balls each at 3 different compressions (80, 90, and 100)
  - How would you design an experiment to determine which type can be driven the furthest?
- Example 13 (page 51-52): Pelletizing experiment

Design 2: Randomized Complete Block Experiments

Description

- At least on factor is a blocking variable
- "Blocks" are created by combinations of the levels of the blocking variables
- Within each block, every combination of the primary experimental variables are used at least once
- Randomization is used where possible

Examples

- Example 12 (page 51): Golf ball drive distances study by G.Gronberg
  - Gronberg didn't drive the golf balls all at the same time
  - instead, he personally drove 10 golf balls from each compression each night, 6 nights a week, for 3 weeks.

Design 3: Randomized Incomplete Block Experiments

Description

- At least one factor is a blocking variable
- "Blocks" are created by combinations of the levels of the blocking variables
- Within each block, some combinations (but not all) of the primary experimental variables are used at least once while others are not used at all
- Randomization is used where possible

How does this work??

- There are some clever ways to assign the combinations of the primary experimental variable levels to the blocks that allow us to get "more" out of the data we collect.
- How these block assignments come about gets explained in Chapter 8.

Section 2.5

Preparing to Collect Engineering Data

Read Independently

Chapter 3

Elementary Descriptive Statistics

Section 3.1

Elementary Graphical and Tabular Treatment of Quantitative Data

Summarizing Univariate Data

Introduction: Creative Writing Workshops

Two methods of teaching a creative writing workshop are being studied for their effectiveness of improving writing skills. First, two groups of creative writing students who were randomly assigned to one of two different 3-hour workshops. At the end of the workshop, the students were given a standard creative writing test and their score on the test was recorded.

Exam Scores for Two Groups of Students Following Different Courses

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

Exam Scores for Two Groups of Students Following Different Courses

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

We may have several questions we are interested in answering using this data. For instance,

- Which group did better on average?
- Which group has the most consistent scores?
- Were there any unusually low or high scores in either group?
- If we ignore unusual scores, which group is better?
- Which group had the most scores over 80?
- ...

However, none of these are immediately clear looking at the raw recorded data.

The Purpose of Summaries

Certain questions can and should be asked across many types of experiments.

But seeing data in this kind of *flat* format presents lots of problems for a person trying to understand the relationship between the two groups.

Summaries are tools (mainly mathematical or graphical) which help researchers quickly understand the data they have collected.

The purpose of a summary is to faithfully present aspects of the data in such a way that

- we are capable of answering the types of core questions about the data asked on the previous page,
- we are able to identify more complicated aspects of the data that we may want to investigate further.

Key Idea: Good summaries should be quickly interpreted, provide clear insight into the data, and be widely applicable.

Simple Graphical Summaries

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

Simple graphical summaries aim to provide a better view of the entire set of data. The best graphs are able to make important points clearly and give valuable insights with closer study. Producing good graphs is an art.

Two common graphical summaries

- Stem and Leaf Diagrams
- Dot Diagrams

Frequency Tables

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

- **Class**: A grouping of the observations
- **Frequency**: The number of observations in a class
- **Relative Frequency**: The proportion of the observations in the class
- **Cumulative Relative Frequency**: The proportion of observations in the current class or a previous class.

Histograms

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

A **histogram** is essentially a graphical representation of a frequency table.

Tips for useful frequency tables

1. Use equal class intervals
2. When the goal is to compare multiple groups, use uniform scales on each graph (i.e., keep lengths consistent)
3. Show the entire vertical axis (especially for relative frequency histograms)