Chapter 1

Introduction to Data

This ActiveStats document contains a set of activities for Introduction to Statistics, MA 207 at Carroll College. This is a non-calculus based statistics class which serves many majors on campus. This document is intended for the classroom teacher to support students in active engagement with statistics on a daily basis. This document is not designed to be given to students as is. Rather, it is a teacher resource.

The activity set is designed to work alongside the OpenIntro *Introductory Statistics with Randomization and Simulation* textbook by Diez, Barr, and Cetinkaya-Rundel. The chapters in ActiveStats are numbered to align with OpenIntro, though the subsections may differ. OpenIntro is an open source curriculum with accompanying data sets. OpenIntro is the textbook resource to direct students to for out-of-class reading assignments and review. We also use the Cartoon Guide to Statistics as a supplement for assigned reading.

Data sets for ActiveStats can be found at mathquest.carroll.edu/activestats/data/ or on the class Moodle page.

1.1 Controlled Experiments

Reading Assignment 1.1. Read the syllabus and become familiar with the course page on Moodle. Also sign into WeBWorK, which you can find via the class Moodle page. ▲

Preview Activity 1.2. Before most classes there is a preview activity to complete. This may relate to the reading or to the upcoming lesson. There is no preview prior to the first day of class, but check online for a preview activity to complete *before* the next class.

Goals: (1) Introduce controlled vs. observational experiments, (2) gather data from an in-class randomized controlled experiment

Teacher Note: We start by gathering data on the first day. Then we have something to plot and something to investigate with basic descriptive stats. The primary attributes we are interested in for the controlled study are cup size and count of M&Ms. The teacher should randomly assign the cups. (We used paper dixie cups in 3 oz and 5 oz sizes.) Gender and breakfast (and the other categories) are not controlled variables, allowing for a discussion about the difference between a controlled variable and an uncontrolled variable.

Activity 1.3. M&Ms: controlled experiment activity

1. Your instructor will give you a cup.
2. Using your cup, take as many M&Ms as you would like to eat from the bowl at the front of the room.
3. When you return to your seat please record the following on the Google Sheet linked from our class Moodle page
   (a) What was the size of your cup?
   (b) How many M&Ms did you take?
   (c) What is your gender? (M/F)
   (d) What is your class standing? (Freshman, Sophomore, Junior, Senior)
   (e) What state are you from?
   (f) Did you eat breakfast? (yes/no)
   (g) Do you like M&Ms? (yes/no)

   Now we have data. This data can be combined with data from similar classes to create a larger data set for our investigations.

▲

Follow-up: What do you notice? What do you wonder?
Teacher Note: The primary research question focuses on the impact of cup size on how many M&Ms are consumed. What other research questions might we investigate with this data? We could consider the impact of gender, whether they ate breakfast, or any of the other variables. However, we focus on the cup size because it was randomly assigned, emphasizing the role of controlled experiments in statistical inquiry.

Our primary research question for the M&M task is whether the size of the cup affects the average number of M&Ms taken. This relates to research on plate size and diet (link to NPR article on plate size and calories). How does the size of your dinner plate affect the number of calories you consume?

It might be tempting to address this research question with anecdotes. Anecdotes are stories about individual data points. For example:

- Anecdote 1: The student at the end of the table took way more M&Ms. Maybe its because he skipped breakfast or maybe it was because he had a bigger cup.
- Anecdote 2: The student in the corner didn't take much. Maybe its because she doesn't like chocolate, or maybe it was because she had a smaller cup.

The problem with anecdotes is that you can find one to support almost any theory or opinion. Statisticians rely on randomized controlled experiments to test theories, rather than anecdotal evidence.

In order to investigate our research question about the impact of cup size on candy quantity, there are a number of questions to consider.

Key Questions: **Teacher Note:** - These could be turned into clicker questions with multiple choice responses. This would be helpful for building class norms from day 1, if clicker questions are to be routine.

1. What population are we interested in studying?
   (M&Ms, cups, students in our class, students at Carroll, adults in general, people in general, candy in general)

2. Why was randomization important? (Open response)

3. How could we best describe the number of M&Ms chosen?
   (List everyone's counts and cup size, create 2 graphs of the results for the two cup sizes, provide the minimum and maximum for each cup size, provide the mean for each cup size, provide the median for each cup size, other)

4. How does the size of the cup play a role? (Open response)

5. What graphical display might help us better understand the data?

6. Is there a difference between the number of M&Ms chosen with the two different sized cups?

7. Make a prediction about the experiment regarding the effect of the size of the cup.

8. Make a prediction about the experiment regarding the effect of gender.

Activity 1.4. Balance and vision controlled experiment

Teacher Note: As an alternative to the previous controlled experiment, the balance and vision controlled experiment provides a different first experience (without the need to buy M&Ms.) The difference in means for the balance task tends to be so strong that students aren't left to wonder whether it will be significant. For the cup size experiment, the average difference may be less severe. For time constraint reasonings, we recommend choosing just one "first day experiment."

How important is vision for balance? Are people better at balancing in the dark (where they aren't distracted by things) or in the light (where they can orient themselves visually)? How can we test this?

Each person is different and there are likely many contributing factors that influence their ability to balance on one leg. Name some factors. **Teacher Note:** Shoe type, athletic experience, proper night's sleep, level of intoxication

- Anecdote 1: My friend Bob has trouble balancing in the dark. He seems to use his vision to help him balance.

- Anecdote 2: My friend Ursula can balance on one foot on a stand up paddleboard, even with her eyes closed.

The antidote to anecdotes: Let's do a randomized controlled study and gather data from many participants.

1. Find a partner and a time keeping device

2. When it is your turn to balance, you must raise your foot to at least knee level during the balancing (like Captain Morgan.) No hopping or leaning on other supports.

3. Have one partner go first while the other one keeps time. Flip a coin to determine whether the first participant will balance with eyes open or closed. Heads = eyes open, tails = eyes closed.

4. Record the number of seconds that the participant remained balanced. If you get to 60 seconds, stop and record 60 seconds as the time.

5. Repeat with the other partner. Flip the coin again to determine treatment vs. control group.

6. Record the following data in the class spreadsheet for each participant.

   (a) Eyes open (yes/no)
   (b) Length of time for balancing (in seconds, maximum of 60 sec)
   (c) What is your gender? (M/F)
   (d) What is your class standing? (Freshman, Sophomore, Junior, Senior)
   (e) Do you consider yourself to be a student athlete? (yes/no)

▲

Characteristics of a Controlled Experiment

- Replication: Same treatments assigned to different participants/ subjects/ experimental units
- Control: Grouping similar experimental units
  - Control Group: Mimics normal as much as possible
  - Experimental Group: Some sort of experimental treatment is given to this group
  - Otherwise, the groups are treated exactly the same
- Randomization: All treatments and groups (to the extent possible) must be assigned randomly. Participants are not allowed to choose their own groups. Why?
  - Randomization smooths biases and other confounding variables among the groups.
  - In truly random selection, every unit has equal probability of getting each treatment.
- Ethical: If human subjects are involved in the study, usually an external review is performed before the research begins. This review process focuses on the safety and ethical considerations of the research process. This is done through your local Institutional Review Board (IRB).
- Conclusion: A cause/effect relationship can only be determined by a randomized controlled experiment.

Characteristics of an Observational Study

- Researchers gather data without attempting to influence a variable
- Studies can be prospective (collecting data going forward) or retrospective (using existing data)
- Random sampling (because we generally can't gather data about a full population)
  - Simple random sampling
  - Stratified random sampling (take a random sample in each subsection of the population)
  - Cluster sampling (data is clumped into clusters and some clusters are randomly selected)
  - See section 1.4 in the OpenIntro text for more details.
- Ethical concerns may center more around privacy / confidentiality of data
- Conclusions may be made about associations, but not cause/effect relationships.

Characteristics of a Survey

- Surveys are conducted through questionnaires or interviews, rather than relying on a researcher to observe or measure the data.
- Data is gathered without attempting to influence a variable.
- Surveys may used similar sampling technqiues to those listed above.
- Surveys, such as phone and mail based surveys, tend to have high rates of non-response. Non-response can result in bias if certain subsets of the population are more likely to respond than others. For example, are older or younger voters more likely to respond to a telephone survey about their voting intentions?
- Conclusions may be made about associations, but not cause/effect relationships.

Activity 1.5. Is this a controlled experiment?

1. A group of students wants to determine whether talking on a cell phone while driving causes people to get into accidents. They request records from the Helena Police Department for all motor vehicle accidents over the past year and calculate the percentage of accidents in which cell phones were a factor.
   (a) Observational Study
   (b) Controlled Experiment
   (c) Survey
   (d) None of these

If this is a controlled experiment: What is the treatment? Can you conclude a cause/effect relationship?

2. A marketing research group wants to know how influential the judges on "America's Got Talent" are in determining who the audience votes for. They gather a sample of people and randomly assign them to two groups. The first group watches an entire show (including performances of contestants and judges commentary), while the second group watches only the performances. At the end the researchers compare the votes from the two groups.

   (a) Observational Study
   (b) Controlled Experiment
   (c) Survey
   (d) None of these

   If this is a controlled experiment: What is the treatment? Can you conclude a cause/effect relationship?

3. A psychologist wants to investigate the impact of dish size on the amount of snack food a college student will choose. She randomly assigns some students to a larger cup and some students to a smaller cup and asks them to choose the amount of M&Ms they would like to eat. She then tracks the count of candies per student.

   (a) Observational Study
   (b) Controlled Experiment
   (c) Survey
   (d) None of these

   If this is a controlled experiment: What is the treatment? Can you conclude a cause/effect relationship?

▲

Example 1.6. If you were interested in the effect of hand sanitizer on the spread of illnesses on campus, a study could be designed with a treatment group which is assigned to use hand sanitizer 4 times a day for 2 weeks and a control group which does not use hand sanitizer. Track the illness rates and types among both groups.

The controlled variable in this case is the use of hand sanitizer. The outcome variable is the illness rates and types.

Ethics in research: In order to improve the study, the researcher decides to expose all of the participants to the flu virus, to better detect how well the hand sanitizer works. Is that okay ethically? Why or why not?

Activity 1.7. Describe a controlled experiment that could be used to establish a cause / effect relationship for each of the following scenarios.

1. A supplement manufacturer wants to demonstrate that their products promote weight loss.

2. A shoe manufacturer wants to show that their shoes improve running speed.

3. A textbook author wants to show that her textbook leads to better comprehension than her competitor's textbook.

4. An online store wants to see which of 3 new store fronts (web pages) results in higher sales.

▲

1.2 Descriptive Statistics

Reading Assignment 1.8. Go to the course website and click on the link to our free open source textbook: OpenIntro Stats by Diez, Barr, and Cetinkaya-Rundel. Read Sect 1.3 (Overview of data collection principles) and Sect 1.5 (Experiments). Then answer the preview questions. If you would prefer a paper copy of the textbook, you may purchase one online. ▲

Preview Activity 1.9. One of the goals of this preview is to prove that you have found the class textbook and read the intended sections. We will return to this book many times. While the book may exist only digitially, you are expected to read it. Claims that there is no textbook for Stats class will be met with dismay.

Previews for other days may involve reading as well as gathering data, creating simulations, and conducting statistical tests.

1. In the section on anecdotal evidence, what three anecdotes are provided? [Check all that apply]

   - Mercury poisoning from swordfish
   - Death by heart attack
   - Time until graduation
   - Death by shark attack
   - Charter school success
   - Online dating failure

   Solution: It's the first 3 options.

2. In order to determine whether a medication is helpful in preventing a heart attack, which type of study should a medical researcher perform? [Select one]

   - An observational study
   - A randomized experiment
   - A survey
   - A placebo study

3. In a double-blind study

   - Both the researcher and the participant take a placebo pill.
   - The participant does not know whether they are in the treatment or control group, but the researcher does.
   - The participant does not know whether they are in the treatment or control group, and also does not know that they are being studied.
   - Neither the researcher nor the participant know whether the participant is in the treatment or control group.

- None of the above.

Task #2
If you have a survey for students (to gather demographic data), this would be a good time for that. Categories such as home state, gender, height, length of commute to class (in minutes and in miles), major, year in school, etc. make for interesting data for descriptive statistics.

Goals: (1) Basic descriptive statistics and (2) graphing tools in both Excel and TinkerPlots.

Teacher Note: A primary goal of this section is to become comfortable with basic descriptive statistics and graphing tools. This prepares students to work with data in the first lab (day 3). We use both Excel and TinkerPlots. Select a data set with multiple columns of data. Be sure to have qualitative and quantitative data so that students can explore multiple ways of organizing data. We selected a video game data set adapted from Kaggle.com. If your students come from a common major/discipline, you may want to choose a data set that engages that discipline. For this activity, the teacher can demonstrate the techniques or point students to videos of the techniques.

Activity 1.10. Descriptive statistics and basic graphs in Excel and TinkerPlots

Download the *videoGamesXBoxPS3* data set from the class website. The columns in the video game correspond with the following attributes.

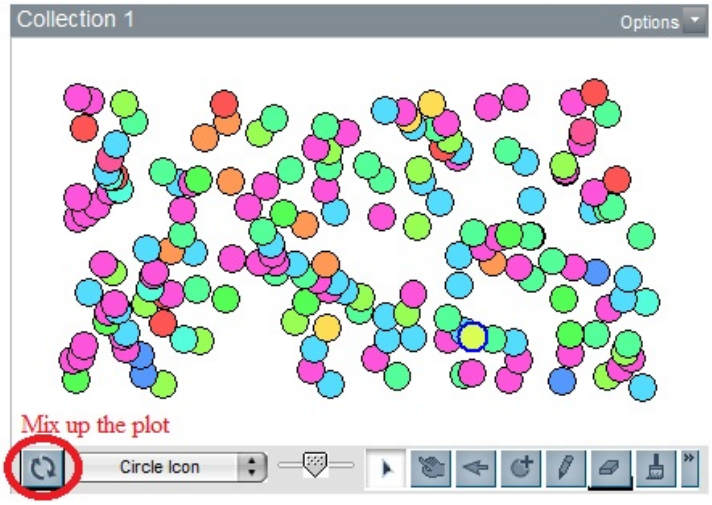
column name	description and units
name	game title
platform	type of gaming system
year of release	when the game first appeared for this platform
genre	game type: e.g. sports, shooter, etc.
NA sales	sales in North America, in millions of dollars
global sales	total sales, in ?millions? of dollars
critic score	average critic score, maximum of 100
critic count	number of critics
user score	average user score, maximum of 10
user count	number of users who provided scores
rating	E = everyone E10+ = everyone ten and older T = teen M= mature

The above table is sometimes referred to as a code book or code library. When you generate a data set, it is polite to include a code book so that others can identify what you measured and how.

Now let's tinker with the data

- Open the data set in Excel. Look through the rows and columns to see what sort of data is available.
- Use *ctrl-A* to select all of the data, and use *ctrl-C* to copy the data to the clipboard in Excel.
- Open TinkerPlots and drag an empty table into the workspace. Use *ctrl-V* to paste the data into the table in TinkerPlots.
- Within TinkerPlots, drag a plot into the workspace. Spend 5 minutes exploring your data. Click on different columns in your table. Click and drag on points in your plot. Check in with a neighbor and see how they are organizing their data. In short: Tinker with your data.

Next we will explore with purpose. As you sort the data for each of the following, keep in mind that you can always reset your plot to an unsorted cloud of data using the *Mix up plot* command in the lower left corner of the plot. If your cloud of dots is all grey, this means you have not yet selected an attribute to explore. Select a column of your choice in your table to see the data colored by that category.



In order to copy an image from TinkerPlots, go to Edit, then Copy As Picture. Then copy the image into your Word document. Note: The copy-paste feature may not work as smoothly with Google Docs.

1. Sort the data by platform.
2. Sort the data by platform and by year. What do you notice? What do you wonder?
3. Sort the data by platform and by rating. What differences do you notice in ratings based on platform?
4. Sort the data by platform and user review. What do you notice? What do you wonder?

5. Sort the data by year (and not platform). Create a boxplot in TinkerPlots.

   When you sorted the data by year, the color of dots changed subtly because the data is quantitative and can take on any numerical value in a range of values. When you sorted by ratings, the dots were is contrasting colors. If red means mature and yellow means teen, we do not anticipate mixing red and yellow to create orange halfway between them.

6. Find another example of a quantiative (numeric) variable and another example of a categorical (qualitative) variable in the video games data set.

▲

Activity 1.11. Descriptive statistics in Excel

Open the *videoGamesXBoxPS3* data set in Excel. Now we will use features of Excel to find common descriptive statistics.

1. Find the mean, median, and mode for the user ratings and for the critic ratings for all data in the sample. Among the three measures of center (mean, median, and mode) which would be the most useful for summarizing this data and why?

2. Find the standard deviation and variance for both the user ratings and critic ratings.

3. The five number summary of a single variable data set consists of the minimum, lower quartile, median, upper quartile, and maximum. Find the five number summary for both the user ratings and the critic ratings.

4. Explore the basic graphing options in Excel, under the Insert tab. Are there any graphs you can be make easily to summarize the data? **Teacher Note:** There are limits on what Excel can do easily, so remind the class that TinkerPlots may be an easier tool for graphs, particularly for histograms and boxplots.

5. Use the descriptive statistics tool in Excel. What types of information are provided by this tool?

   Choose the Data Analysis option in the Data tab. Select Descriptive Statistics from the list of options. Then highlight the range of data you would like summarized, check the boxes "Labels in first row" and "Summary statistics," and click OK.

   Note: If you do not see Data Analysis in your Data tab, do a quick online search for how to add the Data Analysis toolpack for Excel.

▲

Activity 1.12. Practicing Descriptive Statistics
Teacher Note: Time permitting, create a worksheet focused on descriptive statistics and basic graphs in Excel and TinkerPlots. Focus on measures of center and measures of spread. If you have a local data set, relevant to your students/community, use that.

Note: Students may be at rather different places in their technology skills (and technology anxiety). Consider a worksheet for students to work on independently or in small groups so that the teacher is free to circulate the lab to support those most in need

▲

1.3 Lab 1 - Descriptive Statistics with TinkerPlots and Excel

Reading Assignment 1.13. 1. Read pages 7-23 in the Cartoon Guide to Statistics. Take note of the definitions of the mean, median, the 5-number summary, the IQR, and the standard deviation.

2. On the Moodle page you will find the lab that we will be doing in class. In the lab instructions you will find several videos intended to teach you some of the basics of MS Excel and Tinkerplots. Watch these videos and practice the skills on your own before you get to class. You will be working on the actual lab tasks with a partner during class.

3. Answer the preview questions on Moodle

▲

Preview Activity 1.14. 1. Presume that we have a data set that contains the salary of every member of Apple Inc. Which measure of center would be most appropriate to describe this data if we want to know about the typical Apple employee? [median, mean, min, max, Q1, Q3, standard deviation, IQR]

2. Presume that we have a data set that contains a test score for every student in a statistics class. Which measure of center would be most appropriate to describe how the class did as a whole? [median, mean, min, max, Q1, Q3, standard deviation, IQR]

3. Which measure shows the spread of the middle 50% of the data? [median, mean, min, max, Q1, Q3, standard deviation, IQR]

4. What is the middle value in a box plot? [median, mean, min, max, Q1, Q3, standard deviation, IQR]

5. Let's presume that we have a data set of 6 numbers, the first 5 of which are: 1, 3, 7, 9, and 9. Which of the following additional data points would make the standard deviation in the data set the largest? [100, 0, -1, 9, 10]

Activity 1.15. Lab 1 (in separate file)

- Work with the M&M cups data, finding summary statistics and creating a plot.
- Initial focus on identifying the population, sample, and research question.
- Compare/contrast summary statistics for two groups.
- Compute five number summary and mean and standard deviation in Excel
- Create basic dot plots and histograms in TinkerPlots
- Interval estimates for predicting variable
- Basic interpretations, in context, for all of the above plots and descriptive statistics.

▲

1.4 Sampling methods - Optional

Reading Assignment 1.16. Your reading assignement is:

- Task #1 Cartoon Guide to Statistics pages 89 to 97
- Task #2 OpenIntro book, section 1.4.2: Four sampling methods (approximately 4 pages)

▲

Preview Activity 1.17.

1.5 Summary

Summary 1.18. Student learning outcomes from Chapter 1

1. Students are able to distinguish between controlled experiments and observational studies.
2. Students are able to compute descriptive statistics in Excel and TinkerPlots. These descriptive statistics include mean, median, mode, standard deviation, variance, range, and 5 number summaries.
3. Students are able to create graphical representations of data in Excel and TinkerPlots. Graphs include histograms, line plots, boxplots, and scatterplots.
4. Students are able to compare 2 or more data sets or subections of data using both graphical representations and summary statistics.
5. (Optional) Students are able to identify a variety of sampling methods.
