Lab 1: Data Description and Visualization

Javier E. Flores

January 28, 2019

Introduction

In today's lab, we are going to learn how to numerically describe data features via **summary statistics**. While these numeric measures are (independently) useful in describing data, their descriptive efficacy is compounded when paired with **visual summaries** of data. This considered, we will also learn a few data visualization techniques.

To facilitate our learning and practice of these skills, we will use the Student Alcohol Consumption dataset and data generated by your responses to the survey you all completed over the weekend. On the course website, there should be a de-identified version of the survey data for you to download and use to proceed with the lab.

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission.

Data

Consider the following questions:

Q1) Suppose we want to use these data to learn about the studying habits of all students at Grinnell College. Are these data a population or a sample? Are there potential sources of bias? Explain.

Q2) What if we wanted to use these data to learn about the studying habits of the students enrolled in STA209-04 (this class)? Are the survey data a population or a sample? Is there bias? Explain.

Q3) Think back to the types of bias we previously discussed in class. Does the class survey exhibit any of these biases? If so, which? Explain.

Categorical Variables

Recall from our first lecture that variables can be categorized into one of two general types: **categorical** or **quantitative**. As I mentioned at the time, determining the variable type is important in deciding how data should be analyzed.

In this section, we will learn summary statistics and visualization methods for data characterized by either a single categorical variable or two categorical variables. The Student Alcohol Consumption data will be used for each example.

Single Categorical Variable

Regardless of the exact type - nominal, ordinal, or binary - categorical variables all consist of categories (what a profound statement ☺). As such, the most natural way of summarizing these data would be to describe the total number of cases for each category. This is referred to as computing the **frequency** for each category. **Frequency tables** are often used to display the frequencies for each of a variable's categories. Shown below is an example table generated by the Student Alcohol consumption data. The frequencies of each category of "Mjob" are displayed under the column labeled "Count". In addition to the frequencies for each category, the total number of cases, "N", is also provided.

Tally

Mjob	Count
at_home	59
health	34
other	141
services	103
teacher	58
N=	395

Oftentimes (honest) statisticians aren't satisfied with reporting only frequencies. Instead, statisticians prefer to supplement frequencies with their corresponding **proportion** (and vice-versa).[1] The proportion is found by dividing the category's frequency by the total number of cases in the sample, i.e.

$$\hat{p}_A = \frac{n_A}{n}.$$

The $\hat{p}_A$ in the notation above may be read as "the sample proportion in group A". Similarly, n_A may be read as "the sample frequency of group A", and n refers to the total number of cases in the sample. As an example, if we wanted to compute the proportion of cases whose mothers worked in healthcare, we would compute

$$\hat{p}_{health} = \frac{n_{health}}{n} = \frac{34}{395}.$$

As seen in the example above, there are a few symbols statisticians use for the sake of "efficiency" (actually, we're just lazy and don't want to write words). The table below describes a few of these symbolic conventions.

Table 1: Commonly used statistical symbols

Statistic	**Example**	**Description**
Proportion (population)	p_A	The proportion of category A cases in the *population*.
Proportion (sample)	$\hat{p}_A$	The proportion of category A cases in the *sample*.
Frequency (population)	N_A	The frequency of category A cases in the *population*.
Frequency (sample)	n_A	The frequency of category A cases in the *sample*.

Fortunately, you **do not** need to compute either frequencies or proportions by hand. Thanks to our friend Minitab, you can obtain both of these by following these steps:

1) Go to the "Stat" menu and select "Tables" -> "Tally Individual Variables".

2) Select the categorical variable you are interested in. Click "Counts" to obtain the frequencies and "Percents" for the proportions.

Q4) Create a frequency table for the survey question "Which genre of music do you listen to most often?". Include the result as a table in your write-up.

Q5) Using the frequency table in Q4), compute the proportions for each category. Show your work and compare it to the appropriate Minitab output. Include the results as a table in your write-up. In order to reduce the length of your report, you may combine the tables for Q4) and Q5).

As I mentioned earlier, the effectiveness of numerical summaries is increased when supplemented by a visual summary. For single categorical variables, arguably the most well known visual summaries are **bar charts** and **pie charts**. To create these in Minitab,

[1] To understand why, consider this example: Who is the better hitter - Player A who hit 8 pitches or player B who hit 4? What if I told you Player A received 100 pitches and Player B received 4?

1) Go to the "Graph" menu and select "Bar Chart" or "Pie Chart".

2) Select "Simple" and click "Ok".

3) Select the variable of interest and click "Ok" to create the chart.

Q6) Using Minitab, create both bar and pie charts for the survey question from Q4).

Q7) Which of these is more effective in communicating information? Or are they both equally effective? Explain your choice and rationale.

Two Categorical Variables

With a mastery of single categorical variables under your belt, we can extend the concepts of **frequency** and **proportion** to two categorical variables. Thankfully, this extension isn't too difficult. Instead of a frequency table that looks like this:

	Category	**Frequency**
Variable 1	A	n_A
	B	n_B
	C	n_C
	D	n_D

we now have one that looks like this:

		Variable 2		
	Category	E	F	G
Variable 1	A	n_{AE}	n_{AF}	n_{AG}
	B	n_{BE}	n_{BF}	n_{BG}
	C	n_{CE}	n_{CF}	n_{CG}
	D	n_{DE}	n_{DF}	n_{DG}

Each cell entry, n_{AE} for example, is the frequency of cases that are in *both* the corresponding row-variable category (A in this example) and column-variable category (E in this example). This table is often referred to as a **two-way frequency table**. While this table is a bit busier than a single frequency table, it provides a lot of additional information. As an example, let's consider the two-way table of "Mjob" and "Address" from the School Alcohol Consumption data.

Rows: Mjob Columns: address

	R	U	All
at_home	22	37	59
health	3	31	34
other	34	107	141
services	18	85	103
teacher	11	47	58
All	88	307	395

Cell Contents
Count

Recall that "R" designates rural addresses and "U" urban addresses. In the table above, a row and column labeled "All" are provided. The entries for this row and column are referred to as the **row-marginal** and **column-marginal** frequencies, respectively. Notice how they are simply the sums of all entries within the row (for row-marginal) or column (for column-marginal). For example, looking at the first row, we see that $22 + 37 = 59$. Note also that the marginal frequencies are *exactly* the same as the

frequencies for either the row or column variable. To see this, compare the "All" column above to the frequency table for "Mjob" shown previously.

Using two-way frequency tables, we can compute the proportion for a specific category combination, or **cell-specific proportion**. If, for example, we wanted to know which proportion of students lived in a rural setting and had mothers who worked in health care, we would compute

$$\hat{p}_{health,R} = \frac{n_{health,R}}{n} = \frac{3}{395}.$$

We can also compute a **conditional proportion**. An example would be the proportion of students who have mothers working in healthcare *given*, or *conditional* on, that they live in a rural setting

$$\hat{p}_{health|R} = \frac{n_{health,R}}{n_R} = \frac{3}{88}.$$

Make note of the difference between this and the previously computed proportion. For the conditional proportion I am considering only the rural cases (hence the phrasing, "given that [a case] lives in a rural setting"). In contrast, the cell-specific proportion is computed using all cases. Consider again the two-way table generated from Variable 1 and Variable 2.

		Variable 2		
	Category	E	F	G
Variable 1	A	n_{AE}	n_{AF}	n_{AG}
	B	n_{BE}	n_{BF}	n_{BG}
	C	n_{CE}	n_{CF}	n_{CG}
	D	n_{DE}	n_{DF}	n_{DG}

Using the table above, we may compute the

- cell-specific proportions, e.g. $\hat{p}_{AE} = \frac{n_{AE}}{n}$, where n is the total number of cases;
- row-marginal proportions, e.g. $\hat{p}_A = \frac{n_A}{n}$, where $n_A = n_{AE} + n_{AF} + n_{AG}$;
- column-marginal proportions, e.g. $\hat{p}_E = \frac{n_E}{n}$, where $n_E = n_{AE} + n_{BE} + n_{CE} + n_{DE}$;
- row-conditional proportions, e.g. $\hat{p}_{E|A} = \frac{n_{AE}}{n_A}$;
- and column-conditional proportions, e.g. $\hat{p}_{A|E} = \frac{n_{AE}}{n_E}$.

To construct a two-way frequency table in Minitab,

1) Go to the "Stat" menu and select "Tables" -> "Cross-tabulations and Chi-Square"
2) Select the variables you want to use for the rows and columns of the table and click "Ok".

Q8) With Minitab, create a two-way frequency table using the question "Are you an introvert or extrovert?" as the row variable and the question "Are you spontaneous or methodical?" for the columns.

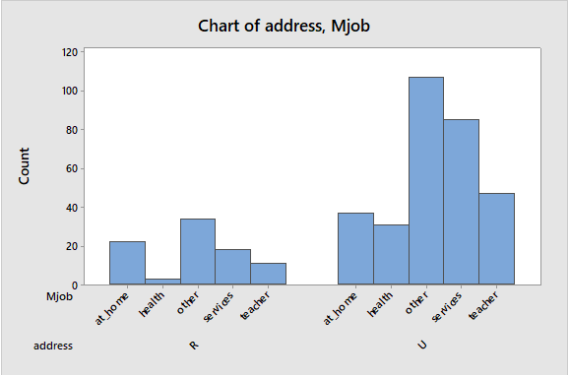
Q9) Compute all row-conditional and column-conditional probabilities. Show your work.

Q10) Which characteristic - spontaneous or methodical - is more likely to describe the extroverts in our class? Explain. (Think carefully before you answer!)

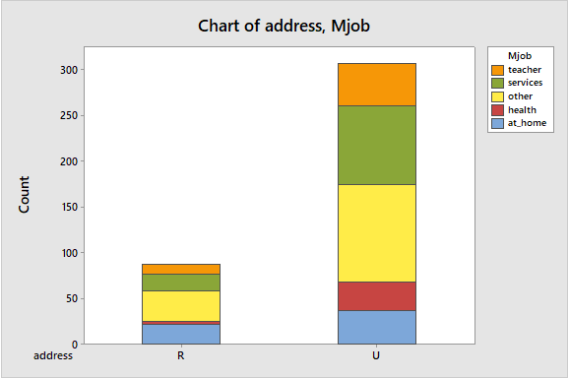
Q11) Are spontaneous individuals in our class more likely to be extroverts? Explain. (Think carefully before you answer!)

Just as with a single categorical variable, there are a few visualization options available for two categorical variables. These include **clustered bar charts** (aka "side-by-side" bar charts) and **stacked bar charts** (aka "segmented" bar charts). Examples of each are provided below. Note that for each of these graphs, there is an **outer**, or cluster-determining, variable ("address" below) and **inner** variable ("Mjob" below).

Clustered Bar Chart



Stacked Bar Chart



To create either of these charts in Minitab,

1) Go to the "Graph" menu and select "Bar Chart".

2) Select either "Cluster" or "Stacked" and hit "Ok".

3) Select the two categorical variables of interest. The outer variable should be selected first and the inner variable second. Hit "Ok"to create the chart.

Stacked (segmented) bar charts are particularly useful for displaying conditional proportions. Minitab has added functionality which allows the output of conditional proportions on any stacked bar chart produced:

1) Go to the "Graph" menu and select "Bar Chart" -> "Stacked".

2) Select the two categorical variables of interest. The outer variable should be selected first and the inner variable second. Note that the graph will condition on the outer variable.

3) Click "Chart Options" and select both "Show Y as Percent" and "Within categories at level 1 (outermost)". Hit "Ok" to create the chart.

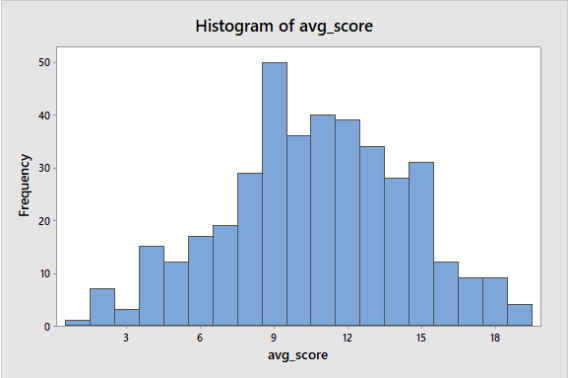
Q12) Create a stacked bar chart showing responses to "Are you an introvert or extrovert?" conditional upon being spontaneous or methodical.

Quantitative Variables

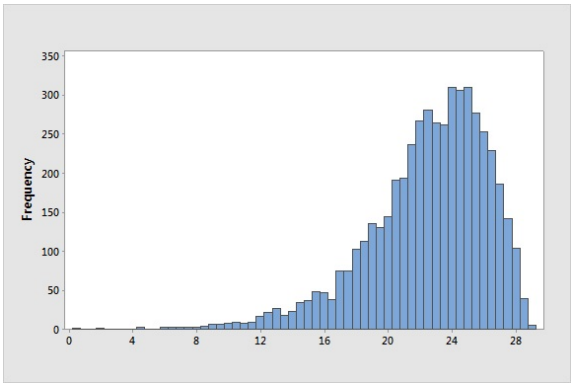
Having surveyed a few of the summarization and visualization tools used for categorical data, we will next consider similar tools for quantitative data. As before, we will use the Student Alcohol Consumption dataset to demonstrate each concept.

Single Quantitative Variable

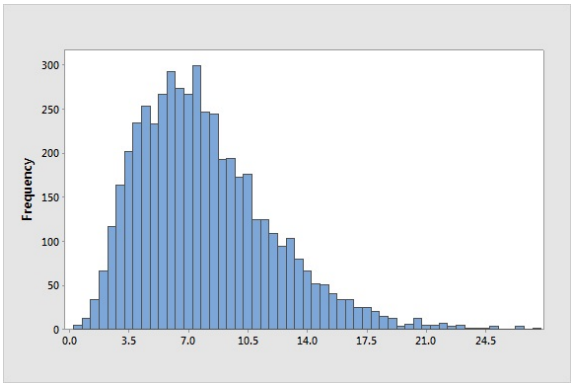
In contrast to categorical variables, quantitative variables aren't as straightforward to describe. There aren't categories we can use to obtain frequencies, and without frequencies, we can't talk about proportions. However, quantitative variables *can* be described by their **distribution**. The distribution of a quantitative variable is the values the variable assumes and how often they are observed. You might be thinking that this sounds an awful lot like frequencies, and you would be right! To emphasize the similarity between the distribution of a quantitative variable and the frequencies of a categorical variable, consider the figure below.



This figure is commonly referred to as a **histogram**, and is often use to visualize distributions. One thing you should immediately notice is how much this histogram looks like a bar chart. The difference in the histogram, though, is that each bar, or bin, represents data grouped in small intervals over the range of the quantitative values. In the example above, we see that the variable takes on values between 0 and about 20. Each bin is then defined by creating small, equally-sized intervals that cover this range (e.g. 0-1, 1-2, etc.). Using histograms, we are able to visually describe the *shape*, *center*, and *spread* of a variable's distribution. In the histogram above, for example, the distribution's shape appears to be relatively **symmetric**[2] about its center, which is somewhere between 9 and 12. In contrast, the histograms below show variables whose distributions are **skewed**.



Left-Skewed



Right-Skewed

[2]The distribution looks the same on either side of an imaginary vertical line drawn through its center

To be more specific, the distribution on the left is **left-skewed** since there is a tail of data pulling out to the left, and the distribution on the right is **right-skewed** since the tail of data is pulling to the right.

In order to create histograms in Minitab,

1) Go to the "Graph" menu and select "Histogram".

2) Select "Simple" and click "Ok".

3) Choose the quantitative variable of interest and click "Ok".

Q13) Create histograms of the responses to the questions "How much time (in hours) do you spend on social media each week?" and "How much time (in hours) do you spend studying each week?". Include both histograms as images in your write-up.

Q14) Describe the shape of each histogram created in Q13). What conclusions can you draw from these figures?

Up until now, we've discussed only the shape of a distribution. However, as I mentioned earlier, a distribution is also characterized by its *center* and *spread*. We refer to (numeric) summaries of a distribution's center as measures of**central tendency**, and summaries of a distribution's spread as measures of **dispersion**.

Some common measures of central tendency include the

- **mean**, or average, of the data;
- **median**, or the middle value when data are ordered from smallest to largest;
- and **mode**, or most frequently occurring data value.

Some common measures of dispersion are the

- **standard deviation**, or the degree of *variability* (aka spread) in the data, and
- **range**, the absolute difference between the variable's **maximum** and **minimum** values.

Other important numerical summaries of quantitative data are **percentiles**. The P^{th} percentile of your data is the value that is *greater than* P percent of your data. Some frequently reported percentiles are the **first quartile**, Q_1, and **third quartile**, Q_3, which are the 25^{th} and 75^{th} percentiles, respectively. Note that the second quartile is the same as the median.

Using the first and third quartiles we can obtain yet another measure of dispersion, the **Interquartile Range** (IQR). A common way to numerically summarize an entire distribution is to report the **five number summary**: the minimum, Q_1, median, Q_3, and maximum.

To compute these statistics in Minitab,

1) Go to the "Stat" menu and select "Display Descriptive Statistics".

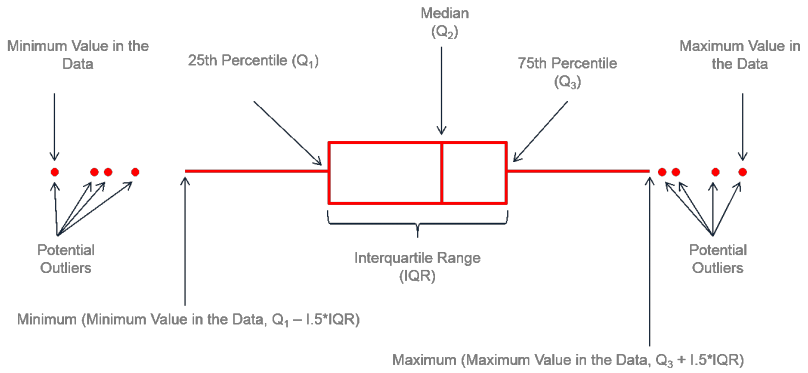
2) Choose the variables of interest and click "Ok".

Q15) Report the mean, median, standard deviation, and IQR of the responses to "How much time (in hours) do you spend studying each week?".

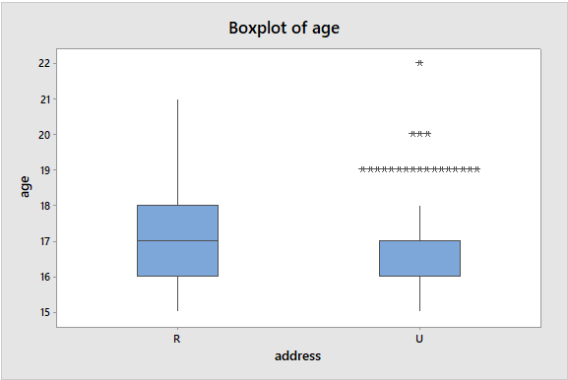
Q16) Suppose there was a student who (badly) lied about their study time in hopes of making a good impression, and reported studying 170 hours each week. Which of the statistics you reported in Q15) do you expect to change (if any)? If you think some statistics will change, explain how they would change and why.

(Visually) Bridging the Quantitative-Categorical Gap

More often than not, statisticians aren't interested in *only* categorical variables or *only* quantitative variables but in how sets of variables might relate, regardless of their general type. In instances where we'd like to visualize the relationship between a categorical and quantitative variable, **boxplots** tend to be pretty useful tools. The diagram below describes each aspect of a boxplot.



The utility of a boxplot in illustrating the categorical-quantitative relationship lies in how compact a representation the boxplot is of the quantitative variable's distribution. Since it is compact, boxplots of quantitative distributions *conditioned* on each level of a categorical variable can be created. Shown below are boxplots of age for each type of address.



The leftmost boxplot corresponds to the age distribution of all respondents who live in rural locations (i.e. age conditioned on living in a rural location), and the rightmost corresponds to the age distribution of urban respondents.

To create boxplots in Minitab,

1) Go to the "Graph" menu and select "Boxplot".
2) Select "Simple" under "One Y" if you want a single boxplot of the overall distribution of your quantitative variable, or select "With Groups" under "One Y" if you want to see the distributions after conditioning on a categorical variable.

Q17) Use boxplots to answer the question: "Do introverts spend more time studying than extroverts?" Include your plots along with a few sentences to explain your reasoning.

Q18) Use one or more of the techniques discussed in this lab to answer the question: "Which genre of music do Social Studies students listen to most often?" Include any relevant figures, and provide rationale for the chosen technique(s) and response. Do not exceed 5 sentences in your explanation.

Q19) Use one or more of the techniques discussed in this lab to answer the question: "Do students spend more time on social media or studying?" Include any relevant figures, and provide rationale for the chosen technique(s) and response. Do not exceed 5 sentences in your explanation.

Challenge (Optional)

In class, I've mentioned that there is statistical software that can be just as (or even more) powerful than Minitab. Personally, my favorite software is R. R is free, open source, and used widely by statisticians. If (after these few days in my class) you already aspire to become a statistician, I highly recommend that you pick up R.

Towards this end, the challenge for this lab is to generate all of the statistics and graphics we discussed (one- and two-way frequency tables, bar/pie charts, histograms, boxplots, and numerical summaries for quantitative variables) using R.

If you need to download R, visit this link. I also recommend downloading RStudio (the open source version), which makes working in R much easier.

Unlike Minitab, R does not have a point-and-click interface. Rather, users write lines of code to be executed by the computer. This makes for a steeper learning curve, but it also means that you have a lot more flexibility in terms of what you can create (*much* prettier graphs!). You can find a lot of basic tutorials on youtube and google to get oriented.

Incentive for doing this (beyond the value of the skill itself) will come in the form of **extra credit** on an exam. You may work on this challenge as a group, but I expect individuals to turn in separate copies of their own code and generated statistics/figures.

Also worth noting is that turning in this challenge may prompt a brief assessment in which you will be asked to state the code used to generate any one of the results produced.

This challenge may be turned in and redeemed for credit at any point prior to the first exam, which will tentatively be scheduled one week after finishing chapter 3.