1.2 – Modeling with Descriptive Statistics

1 Descriptive Statistics

Descriptive analytics is the analysis of data to create informative descriptions of what has happened so far [1]. Descriptive statistics is an essential component of descriptive analytics that are used to summarize data. In most cases they help us measure center and spread of data [2]. We will get into the specific statistics we use to measure center and spread in the next two readings, but first, how can we model with descriptive statistics?

2 Types of Questions

As we look at how we apply modeling to descriptive statistics, we first need to understand what type of questions we can answer using it. This is because, when we remember the mathematical modeling triangle, understanding what we need to find is one of the key inputs to selecting a model. Some questions you can answer using descriptive statistics are:

- What is the story of the data set? What can the data collectively tell me about the group that the data represents?
- How are these two data sets different, and what might those differences mean in context?
- How does this data point compare with the rest of those in a given dataset?
- How can I compare data points from two different data sets?

3 Applying the Modeling Triangle

The mathematical modeling triangle we learned in the first reading applies here, but we can tailor our general *transform*, *solve*, and *interpret* steps to be more specific to building a model for descriptive statistics. The general triangle with the customizations for descriptive statistics are found in Figure 1.

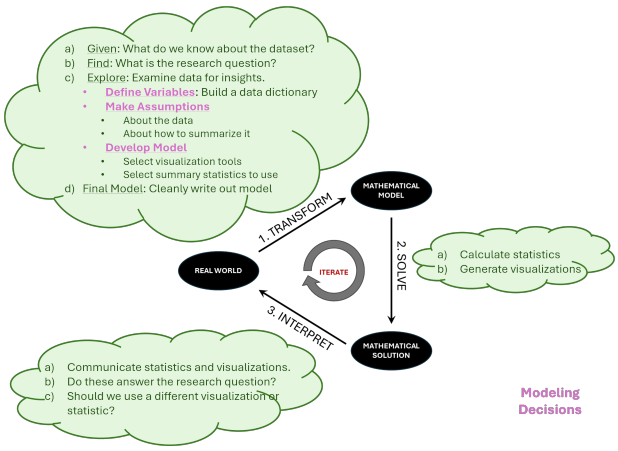

Figure 1: The mathematical modeling triangle as it is applied to descriptive statistics. The pink writing identifies where the modeling decisions are made.

3.1 Transform

As shown in Figure 1, the flow through the transform step is the same. We start with what we are *given*, the data set. We also need to understand the question we are trying to answer; this is what we need to *find*. Once the inputs are understood, we *explore* the data to develop our model. Even the components of exploration are the same: define variables, make assumptions, and develop model. However, what is unique, is what we do within those components.

3.1.1 Define Variables

Within descriptive statistics, understanding the dataset is instrumental to being able to conduct analysis. Within a dataset there are normally always rows and columns, with each row representing a record for which data has been collected [1]. The columns represent a measured characteristic or outcome [1]. The columns of the dataset are our variables. For example, in Table 1, there are four variables and five records of data. Just by looking at the table can you tell what each piece of information is? What is the units for run time? How long is the run? Is the deadlift column the weight or the score?

Table 1: Sample Cadet Fitness Data

ID	Run Time	Deadlift	Age
A001	14.8	300	20
A002	13.5	275	19
A003	15.2	320	21
A004	13.9	290	20
A005	14.1	310	19

As a part of defining our variables, we need to think through and define what each of these columns represent so that we don't get confused later in the analysis. We can define our variables neatly by creating a data dictionary. This data dictionary is an essential **modeling decision** as it annotates how you are interpreting the meaning behind the data. If a definition changes, then your model may need to change as well.

Definition 1.2.1 (**Data Dictionary**)
A table that describes the meaning of each variable in a dataset. [1]

The data dictionary should include a brief definition of the variable to include the units, expected range of values, and type of variable. A list of the types of variables is found in Figure 2. A data dictionary for the data in Table 1 is given in Table 2.

Types of Variables

Numerical: a variable that takes on numerical values [1]

- Continuous: a variable that can take on infinitely many values in an interval [4]
- Discrete: a variable that can only take on specific, separate values, often whole numbers, such as the number of students in class or the number of goals scored in a game [4]

Categorical: a variable that distinguishes among subjects by sorting them into a limited number of categories [3]

- Nominal: labels categories without an inherent order or ranking [3]
- Ordinal: a variable that is measured using a scale that places an order on it, such as t-shirt sizes small, medium, large, and extra large [3]
- Binary: a variable with only two possible values, usually coded as 0 and 1 [3]

Figure 2: The types of variables and their definitions.

Table 2: Data Dictionary Sample Cadet Fitness Data

Variable	Description
`ID`	Unique identifier for each cadet (nominal categorical)
`Run Time`	Time (in minutes) to complete the Army 2-mile run portion of the AFT Range: 13.5-15.2 (continuous numerical)
`Deadlift`	Maximum deadlift weight lifted (in pounds) during the AFT. Range: 275-320 (discrete numerical)
`Age`	Age of the cadet in years. Range: 19-21 (discrete numerical)

3.1.2 Make Assumptions

There are two different types of assumptions we need to consider throughout descriptive statistics, assumptions about the data and assumptions on how to summarize it.

One way we make assumptions about the data is through **data cleaning**. Data cleaning is the process of correcting and reformatting relevant data [1]. Often you will be doing analysis on data you did not collect and therefore unable to control the quality of the data to ensure everything is entered appropriately and is complete. If we were to delete or remove every record that is incomplete or contains a possible error, we lose information that may be valuable to identifying trends and patterns. Depending on the error we may be able to recognize a typo, or gain understanding from what was meant. With data cleaning we can make the decision to edit the record to make it make more sense, as long as we describe what action was taken and why it was *reasonable* and *necessary* to do.

Let's look at an example. Table 3 contains additional records of the same dataset found in Table 1. In looking at this new set of data, we can see records that have issues. For example, record `A010` has "lbs" written within the cell with the deadlift weight. When we start working with this data, Microsoft Excel, or any other database framework, is not going to recognize "295 lbs" as a number because it has text. However, we can assume that it was meant to be entered as "295". We can edit this cell to remove the text, but we need to ensure we make a note as to what we did and why it was reasonable and necessary.

For example:
We edited the deadlift entry for record `A010` to remove the text. This is reasonable because the text identified the units, which have been defined in Table 2. This is necessary to be able to use that data as a number within Excel.

What other potential errors can you identify?

Table 3: Sample Cadet Fitness Data, continued.

ID	Run Time	Deadlift	Age
A010	14.0	295 lbs	20
A011	10.7	280	
A012	15.5	4000	22

Record `A011` doesn't have an age specified. Do we need to discard the full record? Not necessarily; maybe the research question doesn't address age, and the other data look fine. Similarly, the run time for `A011` is over two minutes faster than the others we've seen. Does this mean it's an error? Not necessarily, maybe that cadet is a track athlete and remarkably fast. Regardless of how you choose to interpret the data, you do need to identify and flag these potential issues so that as you proceed with your analysis you can return to these items if they cause issues. Table 4 provides an example of how to clearly annotate the changes you make during data cleaning. Keep in mind that these are **modeling decisions** and there isn't one correct answer; the way you clean the data may be different than this example.

NOTE: the only WRONG answer in data cleaning is to make up or create data. For example, it is not valid to say record `A011` is missing an age, so we're going to input 20 as the cadet's age. It is possible to *impute* data, which is the process of estimating missing data, but we will not discuss imputation in this course.

Another way we make assumptions in descriptive analytics is in how we choose to summarize the data. Depending on the research question, maybe we choose to summarize the data using the mean value instead of the median or we identify the range of the data as more important than the standard deviation. Don't worry, we go into more de-

Table 4: Data Cleaning Justifications

ID	Variable	Edit	Justification
A010	Deadlift	Remove "lbs"	This is reasonable because the text identified the units, which have been defined in Table 2. This is necessary to be able to use that data as a number within Excel.
A011	Run Time	None	This is reasonable because it may be a very fast cadet. It is necessary to keep it to establish trends and identify outliers.
A011	Age	None	This is reasonable because excluding the entire record removes other valuable data. This is necessary to establish trends and identify outliers.
A012	Deadlift	Change value from 4000 to 400 by removing an extra 0	This is reasonable because 400lbs could be lifted, while 4000lbs is unlikely. This is necessary to establish relavant conclusions.

tail into these topics in the next reading. These choices are **modeling decisions** that affect how you answer the research question.

3.1.3 Develop Model

What is a model in the context of descriptive statistics? In this case, the model we are working toward includes one or more visualizations and the descriptive statistics that help answer the research question. Both of these items help translate the data into something simple to help us make sense of it. A complete generalized model should include:

- The type of visualization(s) to be used
- The specific variable(s) being assessed
- The assumptions associated with the data and descriptive statistic selection
- The specific statistics to be calculated

3.2 Solve

The solve step of the mathematical modeling triangle includes calculating the statistics you chose and generating the appropriate visualizations.

3.3 Interpret

During the interpret step you will communicate the statistics and visualizations. Do these sufficiently answer the research question? If they don't fully answer the research question, do we need to iterate to try a different model?

4 Ethical Checklist

Remember that the ethical checklist we use requires three things: data validity, model validity, and clear communication. Below are some questions to think through as you consider whether your model is ethical.

- **Data validity**. Where does your data come from? How and why was it collected? Normally, when you are using data collected by someone else, there should be some report or description of how the data was collected. How much did you need to clean the data? If you had to clean a lot of it, what are the possible implications about this data set?
- **Model validity**. Examine your modeling decisions.
  - While cleaning, did you choose to remove records? Why did you remove them? Does removing them add bias into your results? How did you handle outliers (e.g. the cadet who ran a 10.8 minute two mile)? Did you remove the outlier, modify it, or include it? If you change how you treated outliers, does it significantly change your results? Which option provides a more thorough analysis?
  - Are your visualizations appropriate for what you are trying to convey?
- **Communication**. Are your visualizations clear? Are they misleading, or do they convey an honest representation of the data?

5 Conclusion

Descriptive statistics are a foundational tool for modeling because they help us organize, summarize, and interpret data in a meaningful way. By applying the modeling triangle to descriptive analytics, we recognize that our modeling decisions, from how we define variables to how we handle missing data, shape the story we tell. These decisions carry implications for validity, clarity, and ethical responsibility. As we continue learning about measures of center and spread, keep in mind that modeling is not just about doing math; it's about making sense of information in context and communicating insights clearly and responsibly.

References

[1] Frederick Hillier et al. *MA103 Mathematical Modeling: Introduction to Management, Science, & Business Analytics with Connect*. McGraw-Hill, 2024.

[2] Purdue University. *Descriptive Statistics*. 2018. URL: https://owl.purdue.edu/owl/research_and_citation/using_research/writing_with_statistics/descriptive_statistics.html?utm_source=chatgpt.com.

[3] W. Paul Vogt. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*. Sage Publications, 2005.

[4] Dennis Wackerly, William III Mendenhall, and Richard Scheaffer. *Mathematical Statistics with Applications*. Brooks/Cole, 2008.