Lecture 3 - Variables and variability

Alon B. Muhame

September 12, 2024

For today, variables and variability. The end goal:

- Have a framework for discussing data science questions or problems
- A way to link to an underlying statistical/data science model

This will provide structure for us later

Variables

Let's begin by considering an example: Does sleep deprivation affect level of hunger and attraction to dessert foods? How would you go about answering this question?

Definition 1
A variable is any characteristic of a person or thing that can be assigned a number or category. Variables can be classified as: [1]

1. *categorical (a group designation)* [2] *or*
2. *quantitative (meaningful numerical value).*

Some variables are **quantitative**, taking numerical values on which ordinary arithmetic operations make sense. Other variables are **categorical**, taking category designations. (If there are only two designations, the variable is also called **binary**.)

Definition 2
An ***observational unit*** *s the person or thing being measured. An example include: people, cities, universities*

Definition 3
Data *are the values measured or categories recorded on individual entities of interest. These individual entities on which data are recorded are called* ***observational units****. The recorded characteristics of the observational units are the* ***variables of interest****.*

Suppose I think university or Institute students text too much. If I ask you how many text messages you have sent today, do you think everyone will have the same answer? Can you suggest any explanations?

So, it turns out that one of the overarching goals in Statistics, data science, machine learning and other data related disciplines aim to explain variability. If everybody was the same, we wouldn't need Statistics, data science and the world wouldn't be very interesting!

[1] A variable is a name that refers to a value. The following statement assigns the value 5 to a variable named X, **X** = 5

[2] qualitative variables may also include: strings such as letters and words, Timestamp objects to represent dates and times, and tuples to represent latitude, longitude pairs. It also introduces Geopandas, a library for working with location data)

Examples

1. Suppose I collect the following data on you:
   - Number of text messages you have sent today
   - Whether your most recent text message was incoming or outgoing
   - Time of your first text message today
   - Year in school
   - Number of course units in diploma course
   - Mobile telephone number for your most recent text message
2. Explain why the following are *not legitimate* variable definitions:
   - How many of you have not sent at least one text message today [3]
   - Whether women tend to send more text messages than men This is a research question [4]

Definition 4
Research question: Why you are gathering the data/often looks for pattern in a variable or compares variable behavior across groups or looks at the relationships among variables.:

The type of the variable available in the dataset determine to the big extent the type of analysis to conduct, the questions to pose to data and the model to select for training. Consequentially, an important question for us as data scientists and practitioners is which approach to use when conducting exploratory data analysis , univariate or multivariate? There are scenarios for select univariate analysis and multivariate at different points in the analysis [5].

Numbers

One of the variable we work with in data science and machine learning are **Numbers**. Tools like Python provide for working with many kinds of data, including numbers, words, dates, times, and locations (latitude and longitude). Examples include : integers, floats etc

Arithmetic

On variables such as numbers, using tools like Python, we can perform computions. Uisng Python operators that perform arithmetic such as addition and subtraction are + and -, variables (especially) quantitative can be manipulated to make real life sense.

[3] The variable (with you all as the observational units) would be whether or not you have sent at least one text message
"How many. . . " is a summary of that variable

[4] *The variables (questions to each observational unit) are gender identification and number of text messages*

[5] However, the data science question has a big say on the determination of the analysis to conduct and this can sometimes create an overlap)

Math Functions

In addition, to aritthmetic operators such as addition and substraction, tools like Python provides functions that compute mathematical functions like sin and cos, exp and log. [6]

Calculating with Variables

In the Lab session using Python, we'll use variables to solve a problem involving compound interest. It might not be the most exciting example, but it will leverage what we have studied until upto this point so far, and it reviews exponentiation and logarithms, which we are going to need.

If we invest an amount of money, P, in an account that earns compounded interest, the total accumulated value, V, after an interval of time, t, is [7]:

$$V = P\left(1 + \frac{r}{n}\right)^{nt}$$

where r is the annual interest rate and n is the compounding frequency. [8]

Example 3: work with a partner next to you in class

For each of the following research questions, identify the observational units and variable(s), and specify whether the variable(s) are quantitative or categorical. **Be sure to state any assumptions you have to make.**

1. How much did an average Kampala mid-class consumer spend in China Town Supermarket store in the two weeks since its openning?

2. Do WITI Institute students who attend all- courses (religiously) tend to have lower grade point averages than those who do not attend regularly? (*Hint: Identify and classify two variables*)

3. Is the price of a house related to its size? (*Hint: identify and classify two variables*)

4. Does asking students about the number of texts they have sent first to their friend impact whether they assess themselves as texting too often?

Summary

These notes introduces variables, which are names that refer to values, and two kinds of values, quantitative (e.g., integers and floating-point numbers) and qualitative (such as categorical including binary categories of 0 or 1).

[6] However, they are not part of Python itself, but they are available from a library, which is a collection of values and functions. The one we'll use is called NumPy, which stands for "Numerical Python", and is pronounced "num pie". Before you can use a library, you have to import it. We will see how to compute functions such as compound interest, logarithms of numbers in Python in the Lab sessions

[7] Example adapted from Elements of Data Science by Allen. B. Downey, accessible `https://allendowney.github.io/ElementsOfDataScience` here

[8] For example, if you deposit $2,100 in a bank paying an annual interest rate of 3.4% compounded four times a year, we can compute the balance after 7 years by defining these variables:

These variables will be crucial to assess the data science concept of variability in real life or business cases in later stages of this course - expecially exploratory data analysis (EDA).