STAT 305: Lecture 2

Amin Shirazi

2019-08-19

Why Engineers Study Statistics

Chapter 1: Introduction, Continued

Chapter 2: Data Collection

Course page:
ashirazist.github.io/stat305.github.io

Section 1.2

Basic Terminology, Continued

Types of Data Structures

The most basic way to think about data is to imagine how the the raw observations could be organized once collected.

Collected data can be referred to as a **data set**. If the data set is simple enough, we can store it in a **data table** or **flat file**. Traditional data tables store values relating to a single observation/unit/individual as a row of the table. Each column in the table represents a value for some observed characterstic observed.

Example: Failure time of lightbulbs

A single brand and model of lightbulb is being examined for average failure time. Five bulbs were run until they burned out and their lifetime was recorded in hours. The first bult lasted 521.4 hours, the second bulb lasted 501.2 hours, the third bulb lasted 541.8 hours, the fourth bulb lasted 498.1 hours, and the fifth bulb lasted 528.2 hours.

What and Why

Terms

Data Structures

Types of Data Structures

Example: Failure time of lightbulbs, continued

Assembling the results in a data table could look like this:

Bulb Number	Failure Time (hours)
1	521.4
2	501.2
3	541.8
4	498.1
5	528.2

Each bulb tested gets its own row - which row is attached to which bulb is identified by the first column. The only feature being observed is failure time - so only one column of observations are recorded for each bulb.

Notice:

- Failure Time is a **quantitative continuous** variable.
- This is a **univariate data set**.

What and Why

Terms

Data Structures

Types of Data Structures

Example: Type of bill, date of payment, and payment amount for Mediacom

```
Customer      Type      Date          Amount
John Doe      Internet  01-05-2015    110.00
John Doe      Phone     01-15-2015     10.00
John Doe      Internet  02-05-2015    110.00
John Doe      Phone     02-15-2015     10.00
John Doe      Internet  03-05-2015    110.00
John Doe      Phone     03-15-2015     10.00
...           ...       ...           ...
John Doe      Internet  01-05-2016    110.00
John Doe      Phone     01-15-2016     10.00
Jane Doe      Internet  04-12-2015     90.00
Jane Doe      Internet  05-12-2015     90.00
...           ...       ...           ...
Jane Doe      Internet  01-12-2016     90.00
```

Notice:

- Type of bill is is a **Qualitative** variable.
- Amount paid is **quantitative discrete**.

What and Why

Terms

Data Structures

Types of Data Structures

Example: Machine Parts

> Suppose we get a shipment of 5000 machine parts and would like to verify that the shipment meets the standards the machinist agreed to. We take out 100 parts and examine them carefully. To verify that the parts are as strong as we anticipated, we measure the "Rockwell hardness" with a machine that is accurate to the first decimal place. We also examine each part for scratches and record it weight. Further, we run the part in a test machine to determine if it works correctly.

In this case, we are gathering **4** values on each part. So for instance, the first of the 100 parts we examine could have a measured Rockwell hardness of 3.2, no scratches, a weight of 1.7562 g, and it works correctly. The second of the 100 parts we examine could have a measured Rockwell hardness of 3.1, no scratches, a weight of 1.7901 g, and does not work correctly.

What and Why

Terms

Data Structures

Types of Data Structures

The data as recorded by the researcher might look like this

```
Part identifier: 1/100
  Rockwell Hardness: 3.2
  scratches: no
  weight (g): 1.7562
  functioning: yes

Part identifier: 2/100
  Rockwell Hardness: 3.1
  scratches: no
  weight (g): 1.7901
  functioning: no

...

Part identifier: 100/100
  Rockwell Hardness: 3.4
  scratches: no
  weight (g): 1.7651
  functioning: yes
```

What and Why

Terms

Data Structures

Types of Data Structures

Which we could turn into structured data table like this:
The data as recorded by the researcher might look like this

```
part rockwell_hardness   weight scratches functioning
1                  3.2   1.7562        no         yes
2                  3.1   1.7901        no          no
.                    .        .         .           .
.                    .        .         .           .
.                    .        .         .           .
100                3.4   1.7651        no         yes
```

When data is arranged like this, with each sampling unit on its own row, the data is said to be in **wide format**.

What and Why

Terms

Data Structures

Types of Data Structures

However, we could also structure a data table like this:

```
part    measurement    value
1       Rockwell         3.2
1       weight        1.7562
1       scratches         no
1       functioning      yes
2       Rockwell         3.1
2       weight        1.7901
2       scratches         no
2       functioning       no
.       .                  .
.       .                  .
.       .                  .
100     functioning      yes
```

When data is arranged like this, with each sampling unit on its own row, the data is said to be in **long format**. Long format matches each recorded value to a unique set of identifiers called **keys** - in this case, for example, the first row matches the recorded value 3.2 uniquely to the

Types of Data Structures

The complexity of our data we gather changes based on our objective. Consider the following scenarios:

Scenario 1: Simple Data Structure We have designed a less expensive method for cleaning the byproduct of our production process. We wish to get an estimate of how well it works by using it to clean multiple samples of the byproduct.

- Our data will consist of a identifier to distinguish one sample from another and a measure of cleanliness after treatment with the new method.

Scenario 2: Complex Data Structure Synthesis of a certain chemical can be done in a number of ways. We are considering two sets of substrates, three environments where production can occur, and three chemists to perform the synthesis. Our goal is to get the purest end product.

- We must gather data on substrate, environment, the chemist's identity, and the resulting purity.

What and Why

Terms

Data Structures

Factorial Studies

Factorial Studies involve scenarios in which several process variables are indentified as being of interest and data are collected under different settings of these process variables.

We call the process variables **factors** and the possible settings for a process variable its **levels**

Complete Factorial Studies are factorial studies where data is collected from each possible combination of the levels of the factors.

Partial Factorial Studies are factorial studies where data is collected from some (but not all) possible combinations of the levels of the factors.

What and Why

Terms

Data Structures

Factorial Studies Example

> A pair of chemists, Walter and Jessie, are attempting to synthesize a chemical product and consider purity to be the most important quality. There are three environments available to them (a winnebago, a basement, and a laboratory) and two precursors (pseudoephedrine/methylamine). They are both willing to take the role of "lead cook" and will try all their options in order to get the best results.

- What parts of this synthesis are being treated as variables which can be controlled at the start of the experiment?
- What are the possible values for each of these variables?
- How many ways can the variables be combined?

What and Why

Terms

Data Structures

Factorial Studies Example, cont

Here are all the possible combinations of the factors:

$$(\text{\# of Cooks}) \cdot (\text{\# of Environments}) \cdot (\text{\# of Precursors}) = 2 \cdot 3 \cdot 2 = 12$$

cook	environment	precursor
walter	winnebago	psuedoephedrine
walter	winnebago	methylamine
walter	basement	psuedoephedrine
walter	basement	methylamine
walter	lab	psuedoephedrine
walter	lab	methylamine
jessie	winnebago	psuedoephedrine
jessie	winnebago	methylamine
jessie	basement	psuedoephedrine
jessie	basement	methylamine
jessie	lab	psuedoephedrine

Factorial Studies Example, cont

After testing each scenario, Walter and Jessie decide that the best combination to use is Walt as cook in the lab with methylamine. However, a new "chemist" Victor has joined the group and is going to try to be the cook and "follow the recipe" in the lab. Jessie also tries a new environment, South America, where only methylamine is available.

- If we consider the all the past combinations to be part of this new study, how many combinations of factor levels are now possible?
- Victor never works in the Winnebago, the basement, or South America. Walter never works in South America.

What and Why

Terms

Data Structures

Factorial Studies Example, cont

	cook	env	precursor
1.	walt	winne	pseudo
2.	walt	winne	methylamine
3.	walt	basement	pseudo
4.	walt	basement	methylamine
5.	walt	lab	pseudo
6.	walt	lab	methylamine
7.	jessie	winne	pseudo
8.	jessie	winne	methylamine
9.	jessie	basement	pseudo
10.	jessie	basement	methylamine
11.	jessie	lab	pseudo
12.	jessie	lab	methylamine
13.	jessie	so. am.	methylamine
14.	victor	lab	methylamine

Section 1.3

Measurement: It's Importance and Difficulty

What and Why

Terms

Measure

Key Words

If You Can't Measure, You Can't Do Statistics

Or Engineering For That Matter

- **Validity**: faithfully representing the aspect of interest
- **Precision**: the amount of variation in repeated measures
- **Accuracy**: aka "unbiasedness"; how close a measurement is to the true value "on average"

We **calibrate** to improve accuracy

Section 1.4

Mathematical Models

What and Why

Terms

Measure

Math Models

Mathematical Models and Data Analysis

> **Mathematical Model**: A description of a physical system using mathematical concepts and language.

Identifying mathematical relationships between parts of a system allows us to describe complexity in simple terms.

Example: Height of an Object in Projectile Motion

We can describe the relationship between height of a projectile y and time t as

$$y = h_0 + v_h \cdot t - \frac{1}{2}gt^2,\ t \geq 0,$$

where

- h_0 is the initial height,
- v_h is the initial vertical velocity, and
- g is the (constant) acceleration due to gravity

What and Why

Terms

Measure

Math Models

Example: Height of an Object in Projectile Motion, cont.

$$y = h_0 + v_h \cdot t - \frac{1}{2}gt^2,\ t \geq 0,$$

However, this is not what we see in real life for a variety of reasons. This model assumes

1. g is constant as the ball falls, while g actually depends on the distance between the object and earth,
2. g is a known to infinite accuracy, while we would be using a value that is estimated,
3. Gravity is the only force acting on the object, ignoring drag force, electrical attractions, etc.
4. There are no other changes in the system (for instance, changes in air pressure)

We can fix these by writing a better relationship *or* we can accept that some things won't be known and use a **stochastic model** - a mathematical model that specifically allows for variation (or "randomness"). Understanding how these **stochastic models** work is a major focus of this course.