STAT 305: Chapter 1

Part II

Amin Shirazi

ashirazist.github.io/stat305_s2020.github.io

Why Engineers Study Statistics

Chapter 1: Introduction, Continued

Chapter 2: Data Collection

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 bulb 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 types of data?

(1) Hardness : quantitative / Continuous

(2) weight : " "

(3) scratch : qualitative

(4) Functioning : "

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**.

What and Why

Terms

Data Structures

Factorial Studies

Recall:

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 (also known as **Full Factorial Studies**).

Partial(Fractional) 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 (RV, a basement, and a laboratory) and two precursors (Chemical compound) (pseudoephedrine/methylamine). They are both willing to try all their options in order to get the best results.

response variable

- 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?

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	RV	psuedoephedrine
walter	RV	methylamine
walter	basement	psuedoephedrine
walter	basement	methylamine
walter	lab	psuedoephedrine
walter	lab	methylamine
jessie	RV	psuedoephedrine
jessie	RV	methylamine
jessie	basement	psuedoephedrine
jessie	basement	methylamine
jessie	lab	psuedoephedrine
jessie	lab	methylamine

If we collect data from each of these combinations, we have performed a **A Complete Factorial Study**

What and Why

Terms

Data Structures

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.

- If we consider the all the past combinations to be part of this new study, how many combinations of factor levels are now possible?

cooks env. C.C

$3 \times 3 \times 2 = 18 + (1 \times 1 \times 2) = \underline{20}$

What and Why

Terms

Data Structures

- Victor never works in the RV, 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	RV	pseudo
2.	walt	RV	methylamine
3.	walt	basement	pseudo
4.	walt	basement	methylamine
5.	walt	lab	pseudo
6.	walt	lab	methylamine
7.	jessie	RV	pseudo
8.	jessie	RV	methylamine
9.	jessie	basement	pseudo
10.	jessie	basement	methylamine
11.	jessie	lab	pseudo
12.	jessie	lab	methylamine
13.	jessie	so. am.	methylamine
14.	jessie	so. am.	pseudo
15.	victor	lab	methylamine
16.	victor	lab	pseudo

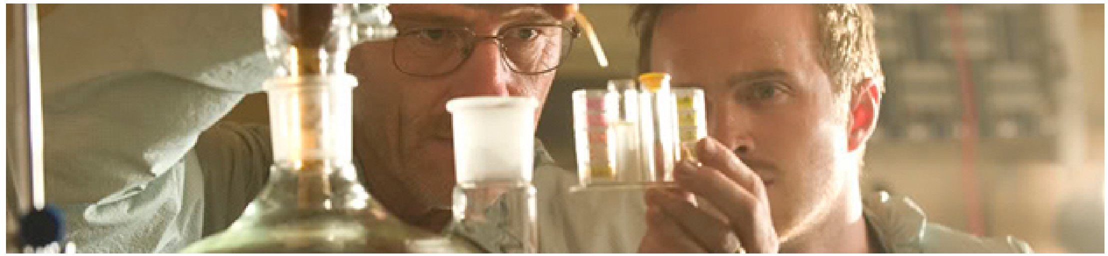
2

What and Why

Terms

Data Structures

Factorial Studies Example, cont



In this case, we would have a **Fractional Factorial Study** - a factorial study in which no data is collected for some possible combinations.

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

- Success in statistical engineering studies requires the ability to measure
- Methods of measurements are available for some physical properties (length, mass, temperature, ...)
  - Often, the behavior of anengineering system can be adequately characterized in terms of such properties
- If it cannot, engineers must carefully define what is about the system that needs observing and then create a suitable method of measurement.

What and Why

Terms

Measure

Key Words

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

Example:

Two students wanted to conduct a factorial study comparing $\boxed{\text{joint strengths}}$ for combinations of three different woods and three glues.

- Didn't know how to have access to strength-testing equipment, so invented their own.
- Suspend a large container from one of the pieces of wood involved and poured water into it until the weight was sufficient to break the joint.
- Knowing the volume of the water poured into the container and the density of the water, they could determine the force required to break the joint.

$(\#\text{ levels})^{\#\text{ Factors}}$ (if levels of different factors are the same size)

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

What and Why

Terms

Measure

Key Words

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

Measurement and its importance and difficulty

- **Validity**: appropriately represent the feature of interest.

  Variation is always present in collecting data.

  - Some come from the the objects under study as they are never alike(that might be of interest to see if the variation is due to the object)
  - Some of it is due to the fact that the measurement processes have their own inherent variablity.

What and Why

Terms

Measure

Key Words

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

Measurement and its importance and difficulty

- **Precision**: the amount of variation in repeated measurement of the same object

A measurement system is called **precise** if it produces small variation in repreated measurement of the same object

- Precision is the internal consistency of a measurement system: typically, it can be improved only with basic changes in the configuration of the system.

What and Why

Terms

Measure

Key Words

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

Measurement and its importance and difficulty

Precision of a measurement is important, but for many purposes it alone is not adequate.

- **Accuracy**: or **Unbiasedness**; how close a measurement is to the true value "on average".

Accuracy is the agreement of a measuring system with some external standard. It can be changed without extensive physical change in a measurement method. So, we **calibrate** to improve accuracy.

What and Why

Terms

Measure

Key Words

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

Measurement and its importance and difficulty

- **Calibration** of a system against the standard (bringing the measurement system in line with the standard) can be
  - As simple as comparing the measurement system to a standard
  - Developing an appropriate conversion scheme and then using converted values in place of recording observed measurements.

What and Why

Terms

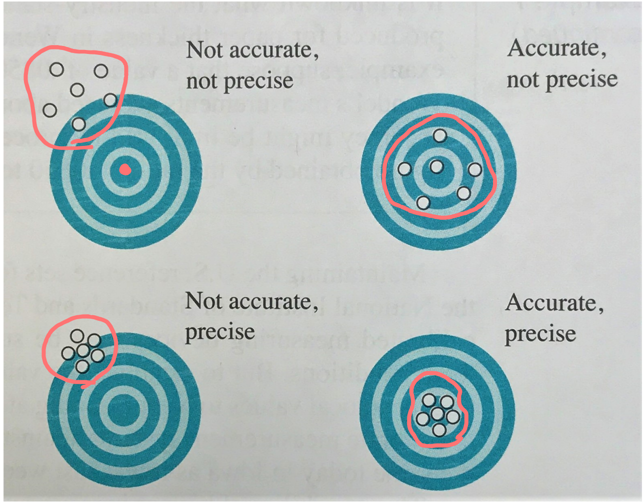
Measure

Key Words

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

Accuracy VS. Precision

- **Accuracy**: how close a measurement is to the true value "on average"
- **Precision**: the amount of variation in repeated measurement of the same object
- Comparing measurement to target shooting



What and Why

Terms

Measure

Key Words

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

Validity, Accuracy and Precision

Section 1.4

Mathematical Models

What and Why

Terms

Measure

Math Models

Mathematical Models and Data Analysis

A discussion on the relationships of mathematics to the physical words and to engineering statistics.

> **Mathematical Model**: A description of a physical system using mathematical concepts and language (in terms of symbols, equations, numbers, and the like)

- Identifying mathematical relationships between parts of a system allows us to describe complexity in simple terms.
- An effective mathematical model is the one which is **simple** and has **predictive ability**.

What and Why

Terms

Measure

Math Models

Mathematical Models and Data Analysis

Example: Height of an Object in Projectile Motion

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

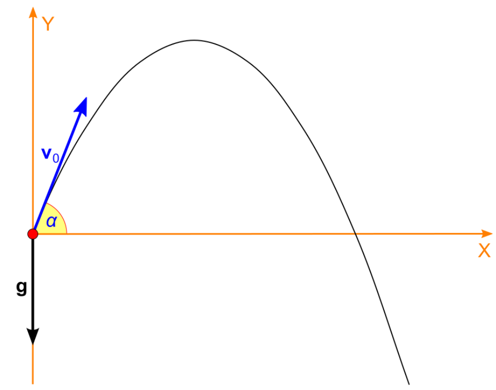
$$h = h_0 + v_h \cdot t - \frac{1}{2}gt^2,\ t \geq 0,$$

where

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

Mathematical Models and Data Analysis

Example: Height of an Object in Projectile Motion



What and Why

Terms

Measure

Math Models

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

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.

What and Why

Terms

Measure

Math Models

What's my point

- We cannot say there is no variation in the measurement or the relation under study is just affected by the components defined in the mathematical model.
- We can control some parts of the variation by planning the data collection process
- There is always some error out of control which are stochastic (random)
- Statistical methods help to deal with this randomness