STAT 305: Chapter 1

Amin Shirazi

Course page:
ashirazist.github.io/stat305_s2020.github.io

Why Engineers Study Statistics

Chapter 1: Introduction

Section 1.1

Engineering Statistics: What and Why

What and Why

Engineers In General

What Do Engineers Do?

- Design/build/operate/improve some system
- Use both **quantitative theories** (i.e., mathematical) and **scientific principles** (i.e., physics, chemistry, psychology(?)) as a guide
- Obvious issue - math/science vast, no one knows everything.
- Additionally, engineers must work outside of "lab conditions" - there is no control over the environment, the users, the timing, ...

So, here's the situation:

The "system" you want to improve is essentially unique

There are competing theories that all seem equal

Experts disagree bitterly about what to do

What's an engineer to do???

Option 1: Just Give Up

A few drawbacks to this one

Option 2: Gather Some Data

Figure Out What Really Matters in the System

Figure Out How a New System Is Gonna Work

What and Why

Engineers In General

But ...

- Without specific training in data collection and analysis, engineers' attemps can be haphazhard and poorly conceived!
- Valuable time and resources are wasted!
- Ambiguous conclusions are reached!

So, engineers need a good toolkit for best possible data collection and interpretation.

What and Why

Engineers In General

Engineering Statistics can help!

What is statistics?

Statistics is the science of collecting, presenting, analyzing, and making decisions from data. Often, as an engineer, it is necessary to **collect and interpret data** that will help in understanding how a new system or product works.

Statistics has applications to engineering through quality control, process control, reliability, risk management, system identification, design of experiments, etc.

What and Why

Engineers In General

Engineering statistics} is the study of how best to

> collect engineering data,
>
> summarize or describe engineering data, and
>
> draw formal inference and practical conclusions on the basis of engineering data,

all while recognizing the reality of variation.

We can break down this study into three main tasks:

> **Summary:** Decribe, summarise and display data
>
> **Inference:** Draw conclusion from data
>
> **Interpretation:** Explain those conclusion in layman's terms (i.e. to people outside statistics)

What and Why

Engineers In General

Data?

What Do I Mean Data?

Data is **essentially just information** we can record.

Examples are incredibly easy to come up with:

- Students with majors and courses they enrolled in

```
student   major         course
John      Philosophy    EASY 101
Kate      Engineering   SMRT 500
Mike      Mathematics   MATH 000
```

- Amount I can bench over time

```
Date     Weight
08/01    55 lbs.
08/02    56 lbs.
08/03    57 lbs.
08/04    59 lbs.
...      ...
12/21    345 lbs.
```

Really trivial to come up with examples

Getting Data is Easy

Getting Good Data is Hard

What and Why

Engineers In General

Data?

Example 1

Two Competing Theories

Goal: Load gears into a continuous carburizing furnace to minimize distortion during heat treating

- **Theory 1**: Load the gears laid in a stack
- **Theory 2**: Hang the gears from a bar
- **Complications**: No two gears are exactly the same - if we test it how do we decide what was the effect of the gear and what was the effect of the loading method?
- **Good engineer**: tried it both ways and collected the distortion measures from multiple attempts

```
Method    Distortion measure (.0001 in.)
Hung      7, 8, 8, 10, 10, 10, 10, 11, 11, ..., 31, 36
Laid      5, 8, 8,  9,  9,  9,  9, 10, 10, ..., 19, 27
```

What and Why

Engineers In General

Data?

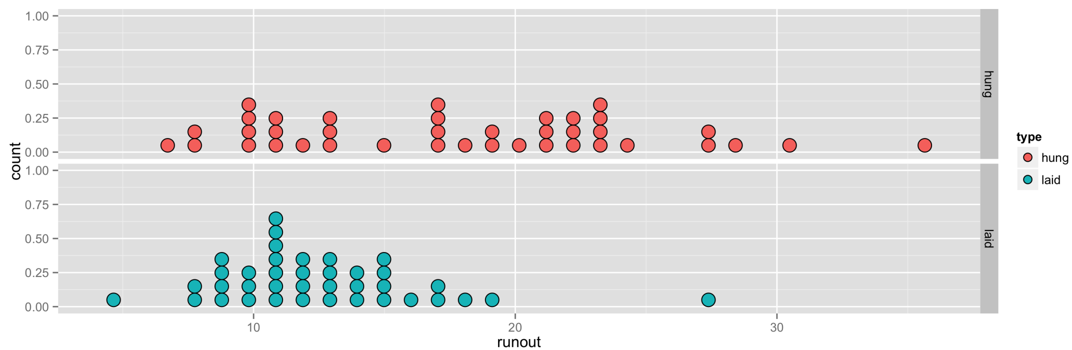
Example 1

Two Competing Theories

Goal: minimize distortion

```
Method    Distortion measure (.0001 in.)
Hung      7, 8, 8, 10, 10, 10, 10, 11, 11, ..., 31, 36
Laid      5, 8, 8,  9,  9,  9,  9, 10, 10, ..., 19, 27
```

Plots and summary values help us see what's going on:




```
Mean hung runout:  17.9
Mean laid runout:  12.6
```

Looks like laying gears in a stack is best

What and Why

Engineers In General

Data?

Example 1

Engineering Statistics

Engineering Statistics

Collecting good data is part of what engineering statistics is concerned with, but of course that's only the first part. **Recall:**

> **Engineering Statistics** is the study of how *best* to
>
> 1. Collect engineering data
> 2. summarize or describe engineering data, and
> 3. draw formal inferences and practical conclusions on the basis of engineering data
>
> all while recognizing the reality of variation

Recognizing the "reality of variation" in the distortion example led the engineer to run multiple tests. This helped rules out the possibility that hanging gears is better: if hanging the gears were actually the right approach, it would be astronomically unlikely that it would have led to that many comparatively large distortions.

Example 1

- If the engineer concludes that laying the gears in a stack is better in minimizing the distortion, but more expensive!
- Now, the question is if it is really worth it to use the results based on that experiment.
- How sure are we to use laying method?

Statistics can help to reach reliable conclusions.

All good engineers use statistical tools

The only question is whether they will use good ones

Section 1.2

Basic Terminology

What and Why

Terms

Basic terminology

It's important we speak the same language. This section introduces common terminology related to statistical studies, types of data, and types of data structures.

Population vs. sample

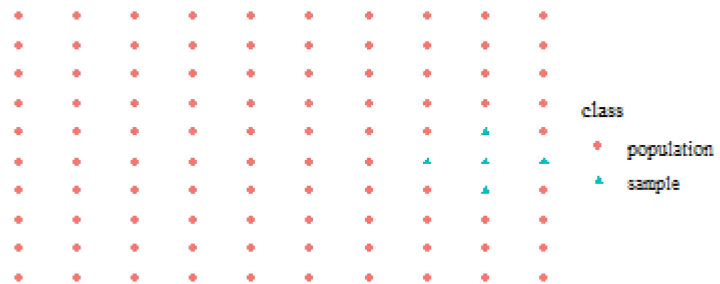
> A **population** is the entire group of objects about which one wishes to gather information in a statistical study.
>
> A **sample** is the group of objects on which one actually gathers data.

What and Why

Terms

Basic terminology

Population vs. sample



> The relationship between a population and a sample. In this example, we have 100 parts and 5 are examined in order to verify acceptibility. Notice we say *one sample*, not *five samples*.

What and Why

Terms

Basic terminology

Population vs. sample

Example: Heat treating gears, cont'd

> Population: All the gears with same make and model as those included in the experiment
>
> Sample: The 77 gears arranged, tested, and measured for distortion

What and Why

Terms

Basic terminology

Population vs. sample

Example: [Fiscal cliff]

On Dec. 1-2, 2012,Gallup conducted a study to find out what proportion of Americans prefer a compromise on the Fiscal Cliff issue. 1000 adults were randomly selected for telephone interviews. The adults were aged 18 and older and living in any of the 50 U.S. states or the District of Columbia.

> Population:
>
> Sample:

What and Why

Terms

Basic terminology

Population vs. sample

Example:

Esbit manufactures fuel pellets out of compressed hexamine powder. Suppose a new shipment of 100 pelletizing machines arrives, and the goal of a new study is to determine the quality of this particular new shipment.

5 machines out of the 100 are randomly selected for comprehensive testing in which each produces 200 pellets, and each pellet's mass, volume, flash point, and rate of combustion are measured.

> Population:
>
> Sample:

Types of Studies

What and Why

Terms

Types of Studies

Types of studies

Experimental study

When an engineer **collects data**, (s)he must decide how active to be. Should the engineer manipulate the process or let things happen and record the results?

> An **experimental study** (or, more simply, an experiment) is one in which the investigator's role is active. Process variables are manipulated, and the study environment is regulated.

Randomized study

> In a **randomized experiment**, investigators control the assignment of treatments to experimental units using a chance mechanism (like the flip of a coin or a computer's random number generator).

Types of studies

Example:[Smoking - experimental]

To answer the question of if smoking affects lung capacity in young men, an investigator does the following.

Finds 100 men age 20 who do not currently smoke. Randomly assigns 50 of the 100 men to the smoking treatment and the other 50 to the non-smoking treatment. Those in the smoking group smoke a pack a day for 10 years while those in the control group remain smoke free for 10 years. Measures lung capacity for each of the 100 men. Analyze, interpret, and draw conclusions from data.

Types of studies

Observational study

An **observational study** is one in which the investigator's role is basically passive. A process or phenomenon is watched and data are recorded, but there is no intervention on the part of the person conducting the study.

Example:[Smoking - observational]

To answer the question of if smoking affects lung capacity in young men, an investigator does the following.

Finds 100 men age 30 of which 50 have been smoking a pack a day for 10 years while the other 50 have been smoke free for 10 years. Measures lung capacity for each of the 100 men. Analyze, interpret, and draw conclusions from data.

Types of Data

Data

Engineers encounter many types of data. It's important to have a way to distinguish all the different types of data you will see.

Qualitative data

> **Qualitative** or **categorical** data are the values of basically nonnumerical characteristics associated with items in a sample. There can be an order to qualitative data, but aggregation and counting are required to produce meaningful numeric values from such data.

Example:

> - male/female
> - green/blue
> - condition A/ condition B

What and Why

Terms

Types of Studies

Data Types

Data

Quantitative data

> **Quantitative** or **numeric** data are the values of numerical characteristics associated with items in a sample. These are typically counts of the number of occurrences of a phenomenon of interest or measurements of some physical property of the items.

Example:

> - mass of fuel pellets
> - temperature of the engine
> - furnance of grear roundout

Data

Quantitative data

Types of numeric data:

- **Discrete:**

  > measurements are separated points (e.g. pages of a book)

- **Continuous:**

  > measurements lie in a continuum (e.g. mpg of a car)

What and Why

Terms

Types of Studies

Data Types

Data

Univariate data

> **Univariate data** arise when only a single characteristic of each sampled item is observed.

Multivariate data

> **Multivariate data** arise when observations are made on more than one characteristic of each sampled item.
>
> A special case is when there are two characteristics - **bivariate**

What and Why

Terms

Types of Studies

Data Types

Data

Example:

hung	laid
7, 8, 8, 10, 10, 10, 10, 11, 11, 11, 12, 13, 13, 13, 15, 17, 17, 17, 17, 18, 19, 19, 20, 21, 21, 21, 22, 22, 22, 23, 23, 23, 23, 24, 27, 27, 28, 31, 36	5, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 16, 17, 17, 18, 19, 27

Univariate or **bivariate**?

Example:

Arrange the data in a table, where:

> Each row is a sample unit, or thing that you measure (gear, in this case).
>
> Each column is a variable, or characteristic that you control or measure.

Arrangement	Runout
hung	18
hung	8
hung	23
laid	13
laid	11
laid	11
...	...

What and Why

Terms

Types of Studies

Data Types

Repeated measures/ paired data

When multivariate data consist of several determinations of basically the same characteristic (e.g., made with different instruments or at different times), the data are called **repeated measures data**.

In the special case of bivariate responses, the term **paired data** is used.

Example:[Paired distortion]

For the gears heat treating example, the measurements were actually made on the 77 gears both *before* and after *heat* treating.

Data Structure

What and Why

Terms

Types of Studies

Data Types

Data structure

Data structures

It is common for several sets of conditions to be compared with each other, in which **several samples** are involved. Here are two structures for multisample studies.

Response variable

> A **response variable** (or dependent variable) is the outcome of a study.

Factor

> A **factor** is any numerical or categorical variable with a finite set of possible values. A **level** is the value of the factor.

(complete) factorial study

A **(complete) factorial study** is one in which several process variables (and settings of each) are identified as being of interest and data are collected under each possible combination of settings of the process variables. The process variables are usually factors.

Example:[Pelletizing machine, pg. 6, 12]

Experimentation with a pelletizing machine using a $2 \times 2 \times 2$ or 2^3 factorial structure. The researchers are measuring the percentage of acceptable fuel pellets for various situations. The factors and respective levels are:

- **Die volume** - low volume vs. high volume
- **Material flow** - current method vs. manual filling
- **Mixture type** - no binding agent vs. with binder

There are then 8 sets of conditions under which data are collected.

What and Why

Terms

Types of Studies

Data Types

Data structure

Example:[Pelletizing machine, pg. 6, 12]

Volume	Flow	Mixture
low	current	no bunder
high	current	no bunder
low	manual	no bunder
high	manual	no bunder
low	current	binder
high	current	binder
low	manual	binder
high	manual	binder

When there are many factors or levels are involved, the number of sampling units in a complete fractional study can quickly reach an impractical size.

What and Why

Terms

Types of Studies

Data Types

Data structure

Fractional factorial study

> A **fractional factorial study** is one in which data are collected for only some of the combinations that would make up a complete factorial study. \end{df}`

Example:[Pelletizing machine, cont'd]

	Volume	Flow	Mixture
2	high	current	no bunder
3	low	manual	no bunder
5	low	current	binder
8	high	manual	binder