Overview of Engineering Statistics

Engineering Statistics
Section 1.1

Josh Engwer

TTU

22 January 2016

The Need for Statistics: Variation in Data

Throughout history, people have collected **data** about certain characteristics of objects, phenomena and processes.

Definition

Data (or a **dataset**) is a set of observations or measurements (**data points**).

Alas, **no real-world dataset has all data points of the exact same value:**

- Houses in a subdivision do not have the exact same price.
- US blockbuster films do not have an exact running time of 90 minutes.
- A collection of steel rods do not have the exact same tensile strength.
- Not all people answer a survey with the exact same set of responses.
- Not all gas stations in a city have the exact same price for unleaded fuel.
- Not all trees in a forest have the exact same branching.
- Not all cookies in a box produced in a factory are exactly the same size.

The data itself is overwhelming & provides little-to-no insight/info/conclusions!
So how to use the data to reliably draw useful conclusions? **STATISTICS!!**

Definition

Statistics is the quantitative handling of data to draw useful conclusions.

Why does Data inherently have Variation??

Because the world is immensely complex:

- Humans are not 100% perfect.
- Instruments never measure to 100% accuracy.
- Materials/substances are never 100% pure.
- Behaviours and processes never act in 100% isolation.
- Future events can never be 100% predicted in advance. (see next slide)

The Need for Probability: Uncertainty in Processes

Life is full of processes whose outcome cannot be predicted ahead of time:

Definition

A **random process** is a process whose outcome cannot be predicted a priori.

Examples of random processes:

Gambling:	Flipping a Coin, Games of Chance (Blackjack, Roulette, ...)
Meteorology:	Weather Systems, Path of a Tropical Cyclone
Economics:	Stock Prices, Demand for Oil
Social Sciences:	Behaviour in People (e.g. fads)
Biology:	Behaviour of Infectious Disease
Engineering:	Instrumentation Errors, Noise in Signals
Physics:	Entropy, Heisenberg's Uncertainty Principle

If we can't predict the outcome, what's the next best thing?
Use **Probability** to determine the **likelihood** of a particular outcome!

Definition

Probability is the quantitative study of uncertainty.

Establishing Meaningful Data: Populations

How to ensure that collected data is lean & meaningful? Define a **population**!

Definition

A **population** is a well-defined set of all objects with desired characteristic(s).
A **finite population** has a finite number of objects.
An **infinite population** has an infinite number of objects or is uncountable.
A **concrete population** is a population that actually exists.
A **hypothetical population** is a population that cannot exist but is still useful.
A **variable** is a characteristic that may change among objects in a population.

POPULATION	POP. TYPE	VARIABLES (Numerical/Categorical)
All students	finite concrete	height (N/C), weight (N), eye color (C)
All topsoil	infinite concrete	pH (N/C), N_2 conc. (N), fertilizer (C)
All possible topsoil pH levels	infinite hypothetical	pH Group (N/C), region (C)
All possible chess games	finite hypothetical	total moves (N), stalemate? (N/C)

The Need for Sampling: Enormous Populations

Unfortunately, most populations are vastly huge:

- There are hundreds of millions of people in the US.
- There are billions of cans of soda.
- There are trillions of cells in the human body.
- There are too many birds (no one knows an accurate count!)

The enormity of most populations of interest causes various issues:

- It takes too much time & money to poll every single person in the US!
- If taste-testers tested every can of soda, there would be no soda to sell!
- If every cell was drawn from a person, the person would die!
- It's too hard for scientists to capture & tag every bird!

The fix to this intractable problem is to take a **sample** of the population:

Definition

A **sample** is a subset of a population.

As it happens, most methods of statistics involve **samples**.

Numerical & Categorical Variables

It's clear that some characteristics are numerical & others are not:

Definition

A **numerical variable**'s possible values are numbers.
A **categorical variable**'s possible values are labels.

WARNING: Numbers can sometimes function as labels!

POPULATION	VARIABLE	TYPICAL SAMPLE
All students	height (N)	$6.1', \ 3.9', \ 5.6', \ 4.0'$
	height (C)	Tall, Short, Tall, Short
	eye color (C)	Blue, Blue, Hazel, Brown
All possible chess games	total moves (N)	$23, 20, 57, 89, 89, 9, 121$
	stalemate? (C)	No, No, Yes, No, No, Yes
	stalemate? (C)	$0, \ 0, \ 1, \ 0, \ 0, \ 1$
	stalemate? (N)	$0, \ 0, \ 1, \ 0, \ 0, \ 1$

As shown above, some variables can either be numerical or categorical. The choice in such a situation is usually determined by context.

Two Branches of Statistics with Probability as a Link

Unfortunately, people want to learn about characteristics of entire populations, but a sample is a subset of a population, and by comparison is quite small!

The solution is a branch of Statistics, called **Statistical Inference**:

Definition

Statistical Inference (or just **inference**) is the quantitative study of samples to draw conclusions of populations.

But it turns out inference involves describing the sample visually and/or numerically, which is another branch called **Descriptive Statistics**:

Definition

Descriptive Statistics is the organization, summary, visualization and presentation of data that conveys useful information about the data.

Since using samples to infer information about an entire population by its nature involves uncertainty, **Probability** also plays a role in inference.

Finally, **Probability** can draw conclusions about a sample from a population.

Univariate & Multivariate Data

How many characteristics are measured at the same time for each object leads to different methods:

Definition

Univariate data involves observations/measurements w.r.t. one variable.
Bivariate data involves simultaneous measurements w.r.t. two variables.
Multivariate data involves simultaneous measurements w.r.t. many variables.

EXAMPLE UNIVARIATE SAMPLES:

- Student Heights (in ft) – measured 01/01/2016: 6.1, 3.9, 5.6, 4.0
- Student Weights (in lb) – measured 11/11/2015: 205, 135, 183, 141
- Student Eye Colors – measured 10/10/2015: Hazel, Blue, Brown, Hazel

EXAMPLE BIVARIATE SAMPLE:

- Student Heights & Weights (in ft & lb) – measured 12/12/2015:
  (6.1, 197), (3.9, 136), (5.6, 187), (4.0, 141)

EXAMPLE MULTIVARIATE SAMPLE:

- Student Heights, Weights & Eye Colors – measured 12/12/2015:
  (6.1, 197, Hazel), (3.9, 136, Blue), (5.6, 187, Brown), (4.0, 141, Hazel)

Statistics & Probability Covered in this Course

So what does this **first** course in **Engineering Statistics** cover?

Definition

(1^{st} Course in Engineering Statistics)

A 1^{st} **course in Engineering Statistics** (MATH 3342) covers:

Descriptive Statistics	(Chapter 1)
Probability	(Chapter 2)
Random Variables	(Chapters 3-4)
Central Limit Theorem	(Sections 5.3,5.4)
Point Estimation	(Chapter 6)
1-Sample Inference	(Chapters 7-8)
2-Sample Inference	(Chapter 9)

Moreover, only **univariate** data will be used.
Bivariate & **multivariate** data will never be considered in this course.

Math Requirements for this Course

The Bad News:

- Algebra: Powers, Roots, Polynomials, Rational Functions
- Algebra: Logarithms, Exponentials
- Algebra: Basic Factoring & Equation Solving
- Calculus I: Computation of Limits, Derivatives & Integrals : /
- Calculus II: Integration by Parts, Partial Fraction Decomposition :(
- Calculus II: Improper Integrals: $\int_0^\infty f(x)\,dx, \quad \int_{-\infty}^\infty f(x)\,dx$:(
- Calculus II: Sums of Finite, Geometric & Telescoping Series :(
- Basic Set Theory: Unions, Intersections, Complements, Empty Set

The Good News:

- Some proofs will be shown, but HW & exams involve no proofs at all! :)
- No Calculus III (except for an occasional partial derivative in proofs) :)
- No trig functions (except from rare integrals: e.g. $\int \frac{1}{1+x^2}\,dx = \arctan x$) :)
- No vectors or matrices!!! Woohoo!!!

Statistics & Probability **NOT** Covered in this Course

So what does a **second** course in **Engineering Statistics** cover?

Definition

(2^{nd} Course in Engineering Statistics)

A 2^{nd} **course in Engineering Statistics** (in your dept??) would cover:

Bivariate Probability	(Sections 5.1,5.2)
Many-Sample Inference	(Chapters 10-11)
Fitting Models to Data	(Chapters 12-13)
Goodness-of-Fit Inference	(Chapter 14)
Nonparametric Inference	(Chapter 15)
Quality Control Charts	(Chapter 16)

Beyond Engineering Statistics

- Bayesian Inference
- Robust Inference
- Decision Theory
- Survival Analysis
- Design of Experiments
- Sampling Theory
- Stochastic Processes
- Time Series Forecasting
- Statistical Physics
- Econometrics
- Game Theory
- Machine Learning
- Business Analytics
- Uncertainty Quantification
- Risk Management
- Monte Carlo Simulation
- Stochastic Differential Equations (SDE's)

Textbook Logistics for Section 1.1

- Ignore the stem-and-leaf display in Figure 1.1 (pg 5)
  - Stem-and-leaf displays will never be used (as explained in the 1.2 Slides)
- Skip ”Enumerative Versus Analytic Studies” section (pg 9-10)
  - Both types of studies will be encountered throughout the course.
  - However, identifying/distinguishing their type is not crucial here.
- Skip ”Collecting Data” section (pg 10)
  - There are several ways to obtain a sample from a population.
  - Some sampling methods may be more appropriate than others.
  - This is superfluous for Descriptive Statistics & Probability (Ch 1-4)
  - On the other hand, this is crucial for Statistical Inference (Ch 6-9)
  - Hence, this section will be covered in Section 5.3

Fin

Fin.