OVERVIEW OF ENGINEERING STATISTICS [DEVORE 1.1]

- **DATASET (DEFINITION):** **Data** (or a **dataset**) is a set of observations or measurements (called **data points**).
- **THE NEED FOR STATISTICS: VARIATION IN DATA**

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

  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!!**
- **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)
- **RANDOM PROCESSES (DEF'N):** A **random process** is one whose outcome cannot be predicted a priori.
- **THE NEED FOR PROBABILITY: UNCERTAINTY IN PROCESSES**

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

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!
- **PROBABILITY (DEFINITION):** **Probability** is the quantitative study of uncertainty.

(CONTINUES ON NEXT PAGE)

OVERVIEW OF ENGINEERING STATISTICS [DEVORE 1.1]

- **ESTABLISHING MEANINGFUL DATA: POPULATIONS**

  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.

- **VARIABLES (DEFINITION):** A **variable** is a characteristic that may change among objects in a population.

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

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

- **THE NEED FOR SAMPLES: ENORMOUS POPULATIONS**

  Unfortunately, most populations are vastly huge, which causes various issues:

  – There are hundreds of millions of people in the US – it takes too much time & money to poll every person!

  – There are billions of cans of soda – if taste-testers tested every can of soda, there would be no soda to sell!

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

- **SAMPLE (DEFINITION):** A **sample** is a subset of a population.

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

- **EXAMPLES OF POPULATIONS, VARIABLES, AND SAMPLES:**

POPULATION	POP. TYPE	VARIABLE	TYPICAL SAMPLE
All students	finite concrete	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	finite hypothetical	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:**

  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**.

  But inference involves describing the sample visually/numerically, which is a branch called **Descriptive Statistics**.

  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.

- **STATISTICAL INFERENCE (DEFINITION):** **Statistical Inference** (or just **inference**) is the quantitative study of samples to draw conclusions of populations.

- **DESCRIPTIVE STATISTICS (DEFINITION):** **Descriptive Statistics** is the organization, summary, visualization and presentation of data that conveys useful information about the data.

OVERVIEW OF ENGINEERING STATISTICS [DEVORE 1.1]

- **UNIVARIATE & MULTIVARIATE DATA:**

  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)

- **PROBABILITY & STATISTICS COVERED IN THIS FIRST COURSE:**

  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.

- **PROBABILITY & STATISTICS NOT COVERED IN THIS FIRST COURSE:**

  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)