ENGG 319

Probability & Statistics for Engineers

Section #01

Probability, Statistics

&

Data Analysis

L01 Dr. Sameh Nassar F16

Probability and Statistics

- We are continually confronted with the collection of information (or ***facts***) (or ***data***).

Statistics:

- Is the branch of scientific inquiry that provides methods for organizing & summarizing data and for using information in the data to draw various ***conclusions***.

Probability:

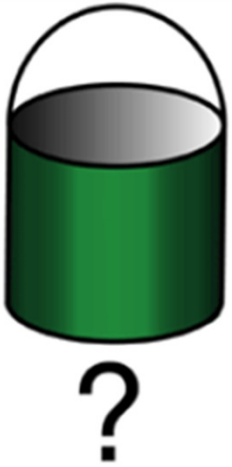
- Refers to the study of ***randomness*** and ***uncertainty***. In any situation in which one of a number of possible outcomes may occur, the ***theory*** of probability provides methods for quantifying the *chances* or likelihoods associated with the various outcomes.

" How do you want it – the crystal mumbo-jumbo or statistical probability? "

Probability and Statistics

- Information is gathered in the form of ***samples*** or collections of ***observations***.
- ***Samples*** are collected from ***populations*** that are collections of all individuals (or individual items) of a particular type.
- The true interpretation of data analysis through modern statistical methods cannot be achieved without some formalism in probability.
- Thus, it is common to study probability before statistics since the elements of probability allow to quantify the strength or confidence in the obtained conclusions.

Probability and Statistics



Statistics: Given the information in your hand, what is in the box?

Probability: Given the information in the box, what is in your hand?

Based on: *Statistics*, Norma Gilbert, W.B. Saunders Co., 1976.

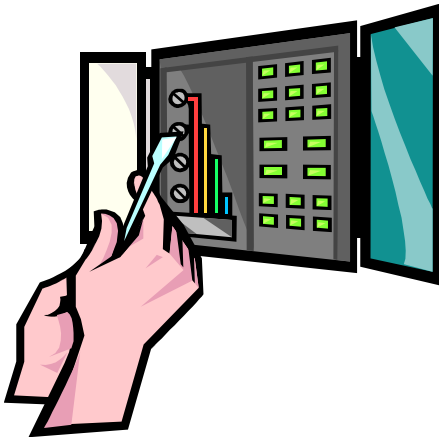
Probability and Statistics

- "Problems in **Probability** allow us to draw conclusions about characteristics of hypothetical data taken from the population based on **known features** of the population".

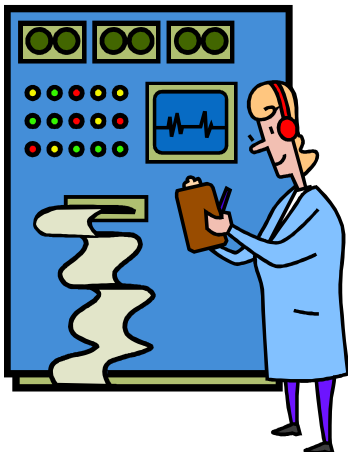
- "Problems in **Statistics**: the sample along with inferential statistics allow us to draw conclusions about the population, with making clear use of elements of Probability".

Why Probability and Statistics?

- By quantifying and explaining the variability in engineered products and systems, the corresponding design can be improved.



- By analyzing data from a process, more sense is gained of where in the process changes may be made to improve the quality of the process.



Why Statistics?

Market Analysis:

- Are people more likely to stop at a Starbucks if they've seen a recent TV advertisement for their coffee?

Health Study:

- Does a low-carbohydrate diet result in significant weight loss?

Political Analysis: How can we predict election results using exit polls?

Engineering study:

- What is the rate at which left-turn accidents occur at intersections without left-turn-only lanes?
- What is the percentage of all items manufactured on the assembly line that are defective?

Statistics: Basic Ideas

Statistics is "the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data". (*The Random House College Dictionary*)

The science of Statistics is commonly applied to two types of problems:

1. Summarizing, describing, and exploring data.
   "Descriptive Statistics"

2. Using sample data to infer the nature of the data set from which the sample was selected.
   "Inferential Statistics" → "*getting information out of a data so we can draw a conclusion about a situation or perhaps make a prediction*"

Fundamental Elements of Statistics

- **Population:** All data of interest. Usually large, sometimes conceptual. "Totality of data".
- **Sample:** Subset of the population for whom we have data.
- **Variable:** a characteristic or property of the data of interest.

- A ***Parameter*** is a numerical summary of the population.
- A ***Statistic*** is a numerical summary of a random sample taken from the population.
  - **Randomness** is **crucial** to experimentation

Example #1

In California in 2003, a special election was held to consider whether Governor Gray Davis should be recalled from office.

An exit poll sampled 3160 of the 8 million people who voted.

- **What is the sample and the population for this exit poll?**

Population:	**8,000,000**
Sample:	**3,160**

From Agresti/Franklin: Statistics 2006

Types of Variables

- **Categorical:**
  - Each observation belongs to one of a set of categories.
  - Key feature is the percentage in each of the categories.

- **Quantitative:** 
  - Observations take numerical values.
  - Key features are center and spread.
  - Values can be discrete (set of separate numbers) or continuous (values form an interval)

How to Describe Quantitative Data?

Measures of location (center) in the sample:

- Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n*.

Measures of the central tendency of the sample are:

(1) Sample Mean.

(2) Sample Median.

(1) Sample Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

For a Population:

$$\text{mean} = \mu$$

How to Describe Quantitative Data?

- Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n*.
- Arrange $x_1, x_2, \ldots, x_n$ in an increasing order of magnitude.

(2) Sample Median:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{if } n \text{ is even} \end{cases}$$

- The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is not influenced by extreme values (or outliers).

How to Describe Quantitative Data?

Measures of the variability (spread) of the sample:

- Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n*.

Measures of the variability of the sample are:

(1) Sample Range.

(2) Sample Variance.

(2) Sample Standard Deviation.

(1) Sample Range:

$$\text{Sample range} = x_{\text{maximum}} - x_{\text{minimum}}$$

How to Describe Quantitative Data?

(2) Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

or:

$$s^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]$$

"used with data of many decimal places"

(3) Sample standard deviation:

$$s = \pm\sqrt{s^2}$$

For a Population:

$$\text{variance} = \sigma^2 \quad , \quad \text{standard deviation} = \sigma$$

Example #2

3-way catalytic converters have been installed in new vehicles to reduce pollutants from car exhaust emissions. However these converters unintentionally increase the level of ammonia in the air. *Environmental Science & Technology* (Sept. 2000) published a study on the ammonia levels near the exit ramp of a San Francisco highway tunnel. The data represent daily ammonia concentrations (ppm) on 8 randomly selected days during afternoon drive-time in the summer of 1999.

1.53	1.50	1.37	1.51	1.55	1.42	1.41	1.48

a) Find the mean daily ammonia level in air in the tunnel

b) Find the median daily ammonia level in air in the tunnel

c) Find the variance of the ammonia level

Example #2 (Sol.) "1/2"

(a) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}\sum_{i=1}^{8} x_i$$

$$= \frac{1}{8}[1.53+1.50+1.37+1.51+1.55+1.42+1.41+1.48]$$

$$\bar{x} = \frac{11.77}{8} = 1.471 \quad \text{ppm}$$

(b) $$x_{\text{arranged}} = \{1.37,\ 1.41,\ 1.42,\ 1.48,\ 1.50,\ 1.51,\ 1.53,\ 1.55\}$$

$$\tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}(x_4 + x_5) = \frac{1}{2}(1.48+1.50)$$

$$\tilde{x} = 1.49 \quad \text{ppm}$$

Example #2 (Sol.) "2/2"

(c)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - \bar{x})^2$$

$$= \frac{1}{7}[(1.53-1.471)^2 + (1.50-1.471)^2 + (1.37-1.471)^2 + (1.51-1.471)^2$$

$$+ (1.55-1.471)^2 + (1.42-1.471)^2 + (1.41-1.471)^2 + (1.48-1.471)^2]$$

$$s^2 = 0.004 \quad \text{ppm}^2$$

Textbook Readings

- 1.1
- 1.2
- 1.3
- 1.4
- 1.5
- 1.6
- 1.7