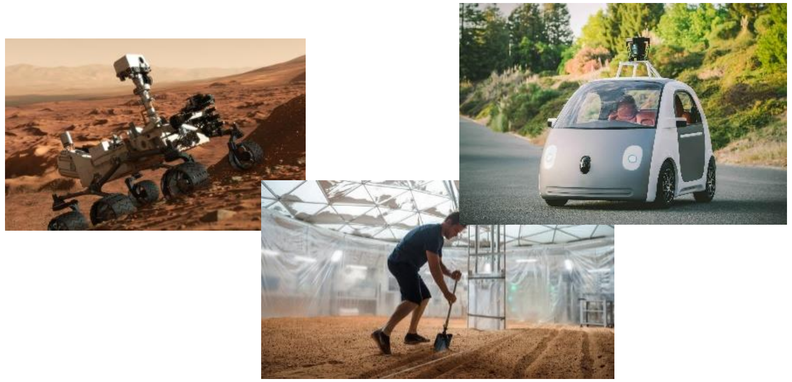
Introduction of IE241

Motivation

- Machine Learning
- Artificial Intelligence
- Optimization
- Optimum Control
- Planning
- Markov Decision Process
- Influential Diagram
- Decision Tree
- Dynamic Control
- Game Theory
- Search
- Stochastic Programming
- Reinforcement Learning
- Bandit problem

⋮

Engineering is all about decision makings



What are the differences in these decision-making strategies?

What are the common aspects in these decision-making strategies?

What type of decision making framework will be used?

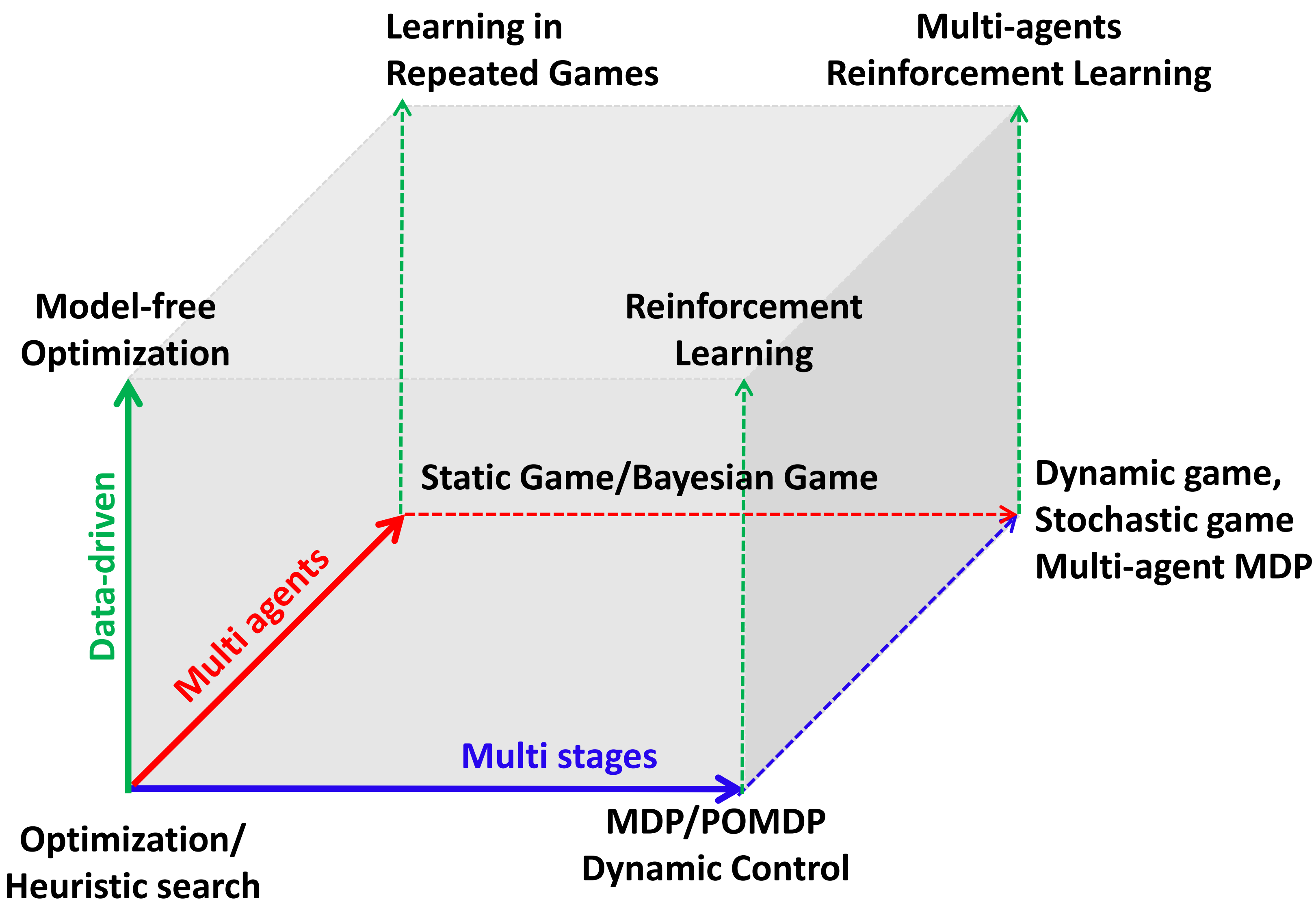
- Single stage or multi stages
- Single decision maker or many decision makers
- Model based or model-free

"Decision makings under uncertainties"

How to model uncertainties?

- **Epistemic uncertainty** (systemic uncertainty):
  Uncertainty arising through lack of knowledge
  - Model uncertainty
  - State uncertainty
- **Aleatoric uncertainty** (statistical uncertainty):
  Uncertainty arising through an underlying stochastic system

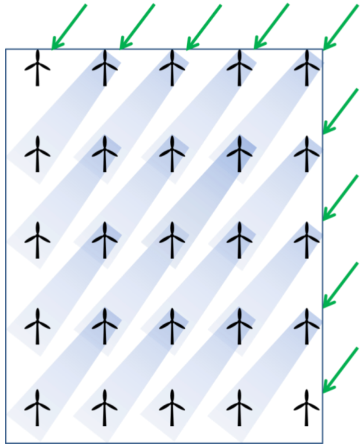
Classified decision making methods



Real-world task

Modeling

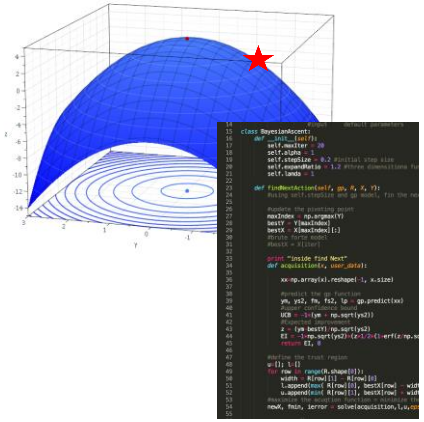
Formal task (model)



$$\underset{\boldsymbol{x}}{\text{maximize}} \sum_{i=1}^{N} P_i(\boldsymbol{x}; \theta, U)$$

Solving

Algorithm (program)



Validation

Are we building the right model?

Verification

Does the algorithm capture all the essential aspects of the model?

Is this solution good for the target system?

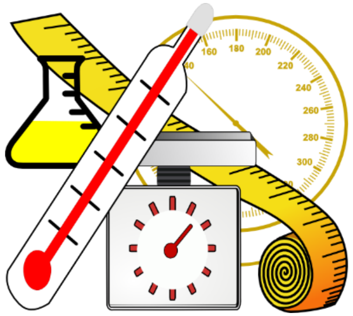
Data can help model more realistic and derive more accurate solution!

How to solve a problem using data

Problem

Tools

Measurement tools

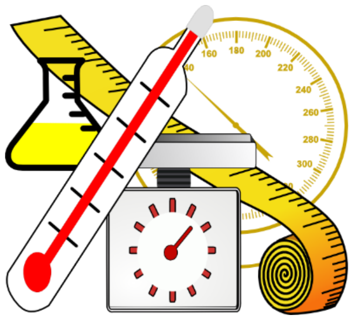


How to solve a problem using data

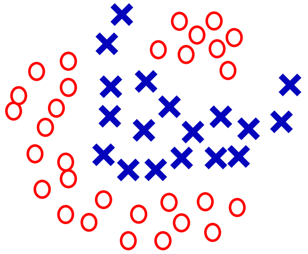
Problem

Tools

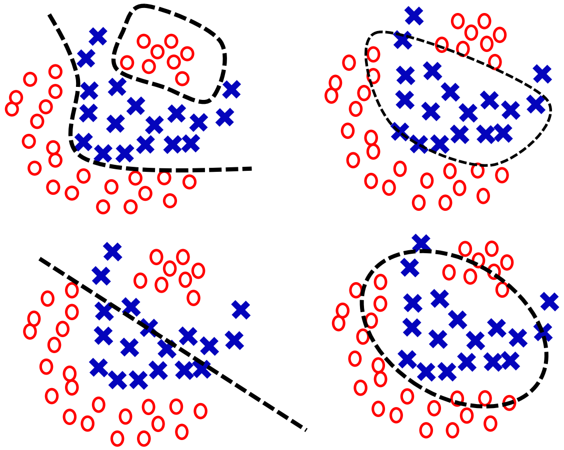
Measurement tools



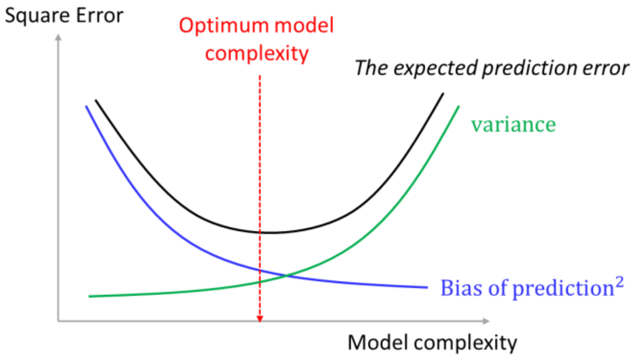
Data



Algorithm or model



Evaluation measures



$$\text{data} \quad + \quad \text{model} \quad = \quad \text{prediction}$$

- **Data** : observations, experience,…
- **Model**: a form of prior knowledge, assumptions, belief
  - ✓ Functional model
  - ✓ Probabilistic model
- **Prediction** : the new knowledge obtained by combining the data and model
  - ✓ Regression
  - ✓ Classification
  - ✓ Clustering

Data

Training data set D

$D = \{(x_i, y_i); i = 1, \ldots, m\}$

$x_i = (x_{i1}, \cdots, x_{in})$

$y_i \in \mathbb{R}$ or $y_i \in \{1, \ldots, N\}$

m input Feature vectors

x_1	x_2	$\cdots$	x_n	y
x_{11}	x_{12}	$\cdots$	x_{1n}	y_1
x_{21}	x_{22}	$\cdots$	x_{2n}	y_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_{m1}	x_{m2}	$\cdots$	x_{mn}	y_m

m outputs

model

Functional form $f(x; \theta)$
Is usually given

Learning Algorithm

Using training data set, a learning algorithm finds the best hypothesis function $h(x)$ that **is believed to accurately predict** the output y for a given query input x

θ^*

prediction

Query input
x_*

Hypothesis
$f(x; \theta^*)$

Predicted output
y_*

Input feature vector
$x_* = (x_{*1}, x_{*2}, \ldots x_{*n})$

If $y_* \in \mathbb{R}$: **Regression**

If $y_* \in \{1, \ldots, N\}$: **Classification**

CHAPTER 1
Introduction

Contents

Motivations

- It is interesting to note that billions of dollars are spent each year by U.S. industry and government for data from experimentation, sample surveys, and other data collection procedures.
- This money is expended solely to obtain information about phenomena susceptible to measurement in areas of business, science, or the arts.
- The implications of this statement provide keys to the nature of the very valuable contribution that the discipline of statistics makes to research and development in all areas of society.

What is Statistics

- Webster's New Collegiate Dictionary defines statistics as "*a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.*"
- Rice (1995), commenting on experimentation and statistical applications, states that statistics is "*essentially concerned with procedures for analyzing data, especially data that in some vague sense have a random character.*"
- Freund and Walpole (1987), among others, view statistics as encompassing "*the science of basing inferences on observed data and the entire problem of making decisions in the face of uncertainty.*"
- Mood, Graybill, and Boes (1974) define statistics as "*the technology of the scientific method*" and add that statistics is concerned with "*(1) the design of experiments and investigations, (2) statistical inference.*"

What is Statistics

uncertainty

random character

data

the design of experiments

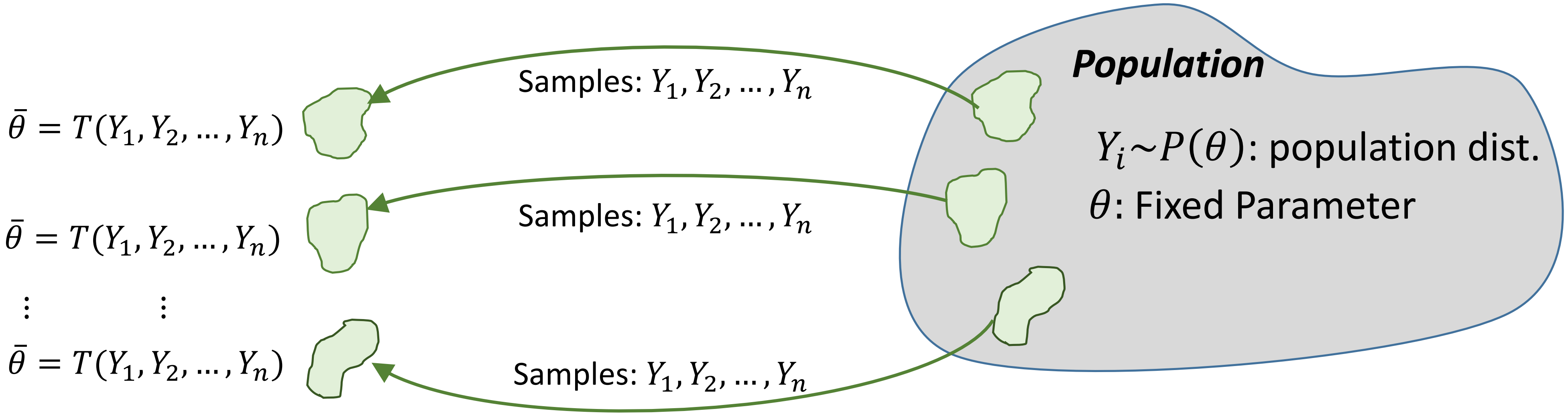
inference

making decisions

Statistics is a theory of information, with inference making as its object

Population vs. Sample

- **Population**
  - ✓ The collection of all the elements of a universe of interest, i.e., target of interest.
- **Sample**
  - ✓ A set of elements taken from the population under study, i.e., a subset of the population.
  - ✓ Since it is impossible and/or impractical to examine the entire population in most cases, a sample is used instead.
  - ✓ If a sample is selected randomly, i.e., every element in the population has the same chance to be chosen, then the sample is called random sample.

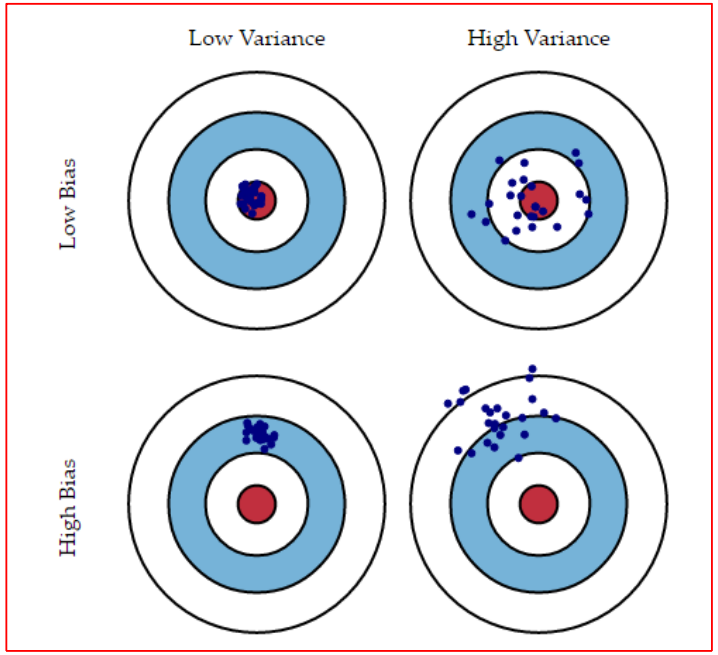


Population vs. Sample

- **Examples:**
  - ✓ The preferences of voters for a president candidate
  - ✓ The voltage at three particular points in the guidance system for a spacecraft:
    - ➢ Presumably, this population would possess characteristics similar to the three systems in the sample (conceptual population)
  - ✓ Measurements on patients in a medical experiment represent a sample from a conceptual population consisting of all patients similarly afflicted today, as well as those who will be afflicted in the near future (conceptual population)

Goal of Statistics

- The study of statistics is concerned with
  - ✓ The ***design of experiments*** or sample surveys to obtain a specified quantity of information at minimum cost
  - ✓ The optimum use of this information in ***making an inference about a population***.
    - ➢ An inference about a population based on information contained in a sample from that population and to provide an associated measure of ***goodness for the inference***.

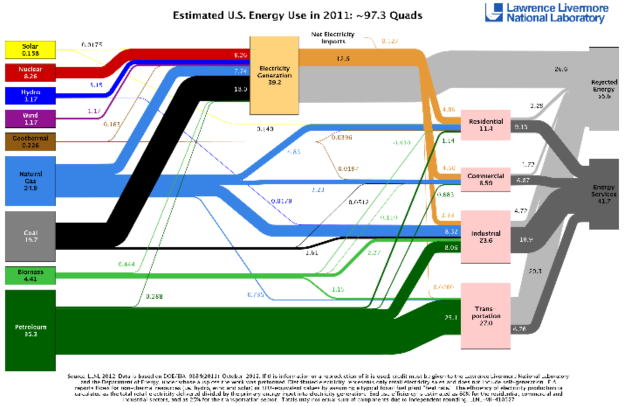


Descriptive Statistics

- We characterize a person by using height, weight, color of hair and eyes, and other descriptive measures of the person’s physiognomy.

- Characterizing a population that consists of a set of measurements is important.
  - ✓ The characterizations must be meaningful so that knowledge of the descriptive measures enables us to clearly visualize the set of numbers.
  - ✓ The characterizations possess practical significance so that knowledge of the descriptive measures for a population can be used to solve a practical, non-statistical problem.

- There are two methods to characterize the population:
  - ✓ Graphical Methods (visualization)
  - ✓ Numerical Methods (inference)

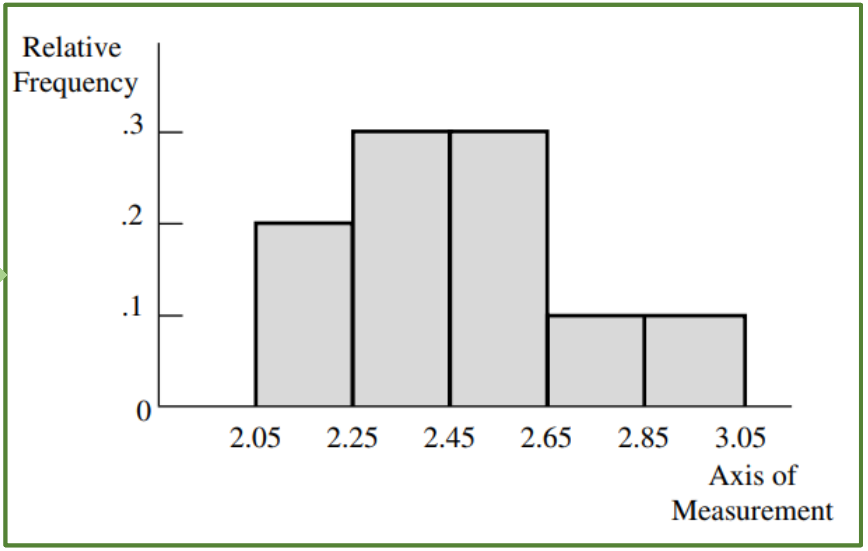


Relative Frequency Histogram

- An individual population (or any set of measurements) can be characterized by *a relative frequency distribution*, which can be represented by a relative frequency histogram.
  - ✓ A graph is constructed by subdividing the axis of measurement into intervals of equal width.
  - ✓ A rectangle is constructed over each interval, such that the height of the rectangle is proportional to the fraction of the total number of measurements falling in each cell.

Characterize the ten measurements:

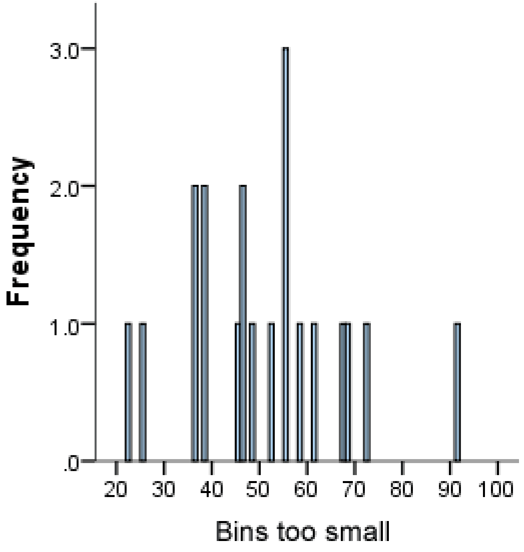
2.1, 2.4, 2.2, 2.3, 2.7, 2.5, 2.4, 2.6, 2.6, and 2.9

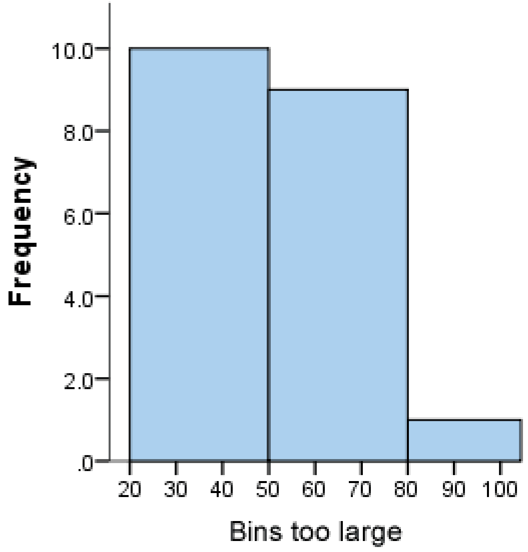


How to Construct a Relative Frequency Histogram

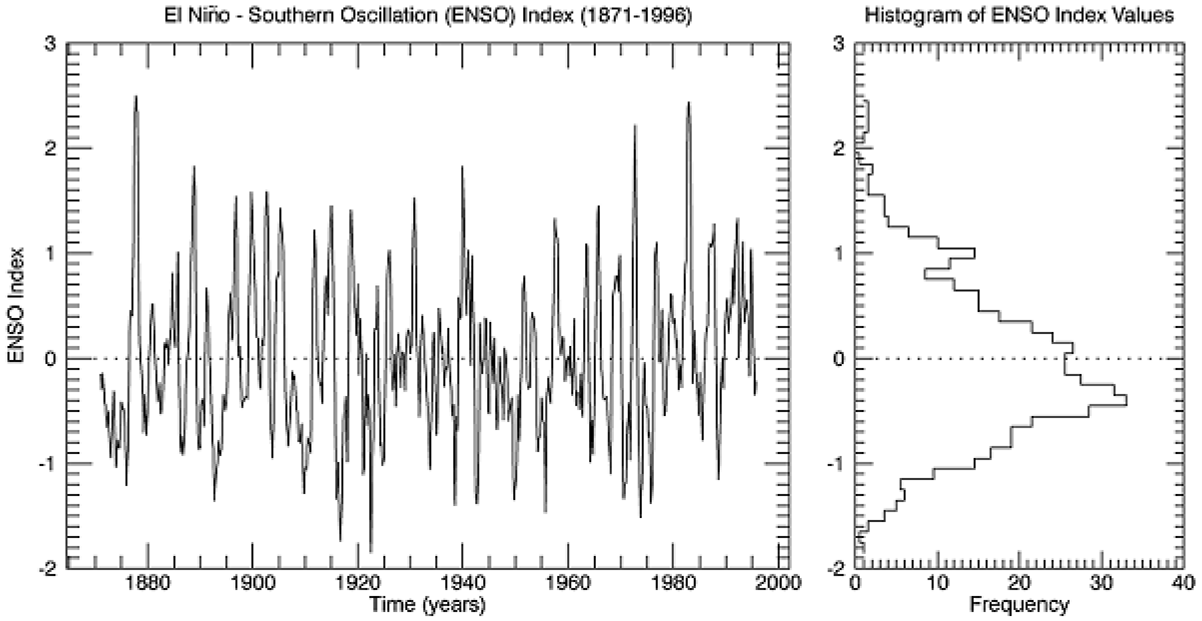
- Points of subdivision of the axis of measurement should be chosen so that it is impossible for a measurement to fall on *a point of division*.

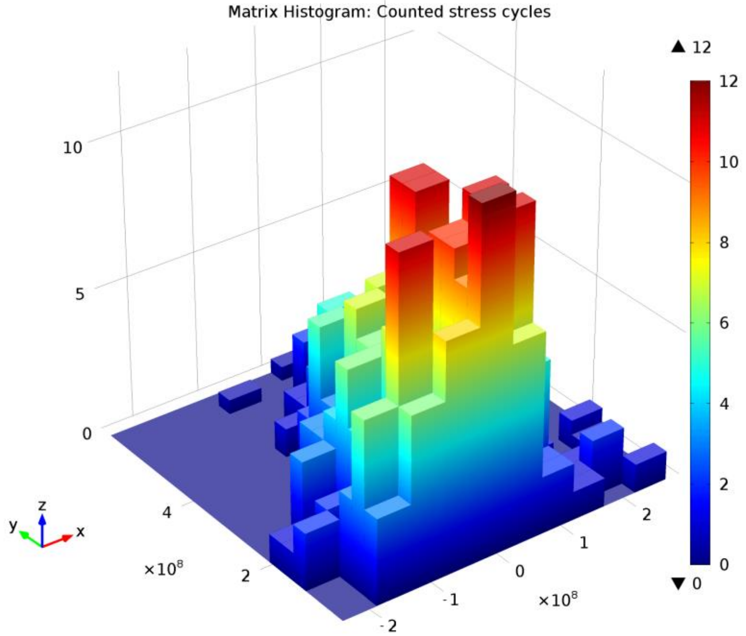
- The second guideline involves the width of each interval and consequently, the minimum number of intervals needed to describe the data.
  - ✓ Generally speaking, we wish to obtain information on the form of the distribution of the data.





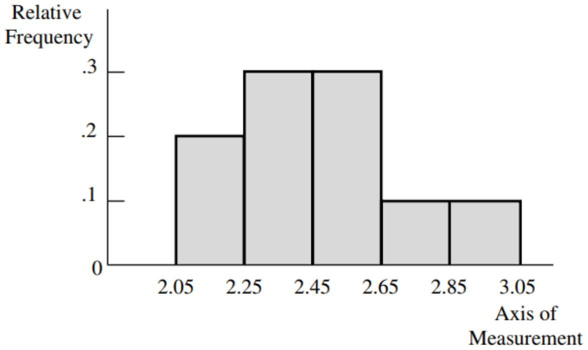
Relative Frequency Histogram

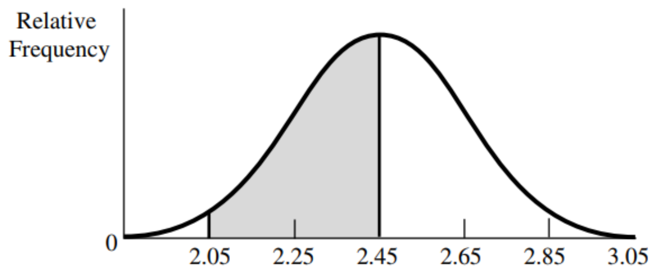




Meaning and Limitation of Histogram

- The relative frequency distribution provides meaningful summaries of the information contained in a set of data.
  - ✓ This is primarily due to the probabilistic interpretation that can be derived from the relative frequency histogram.
  - ✓ If a measurement is selected at random from the original data set, the probability that it will fall in a given interval is proportional to the area under the histogram lying over that interval.

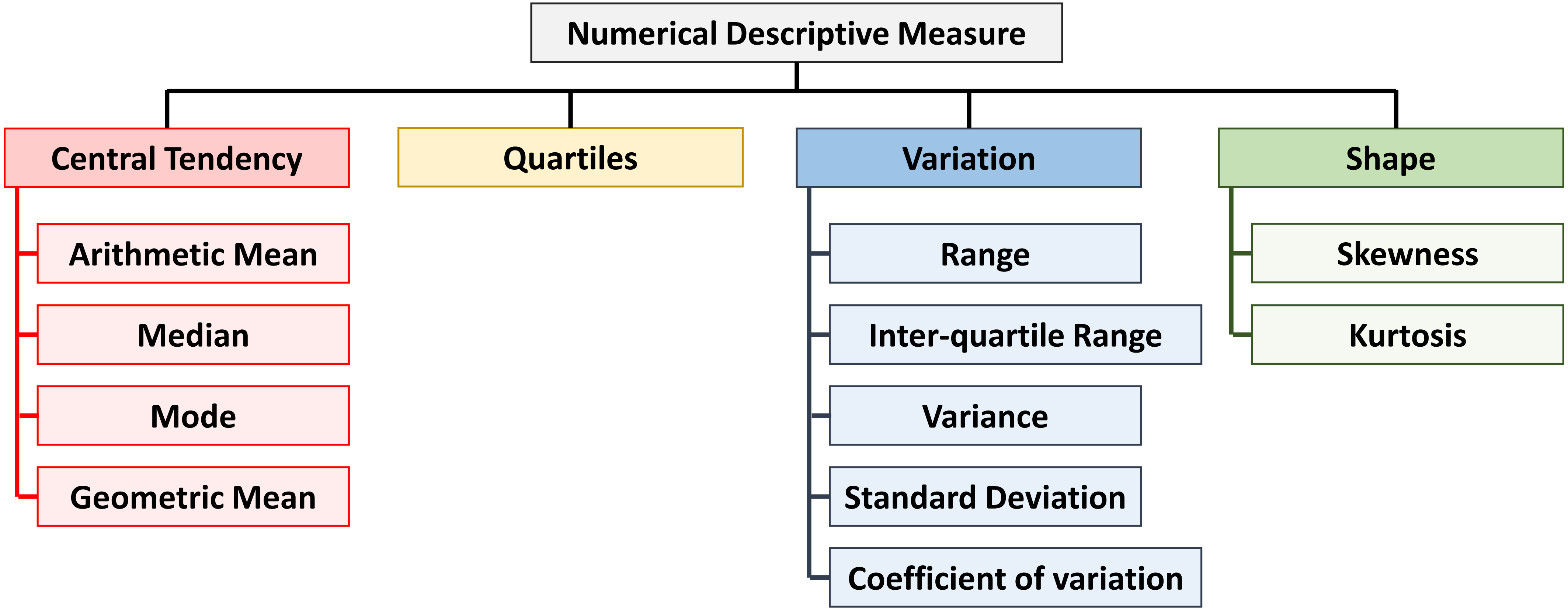




- The relative frequency histograms presented provide useful information regarding the distribution of sets of measurement, but histograms are usually *not adequate for the purpose of making inferences*.

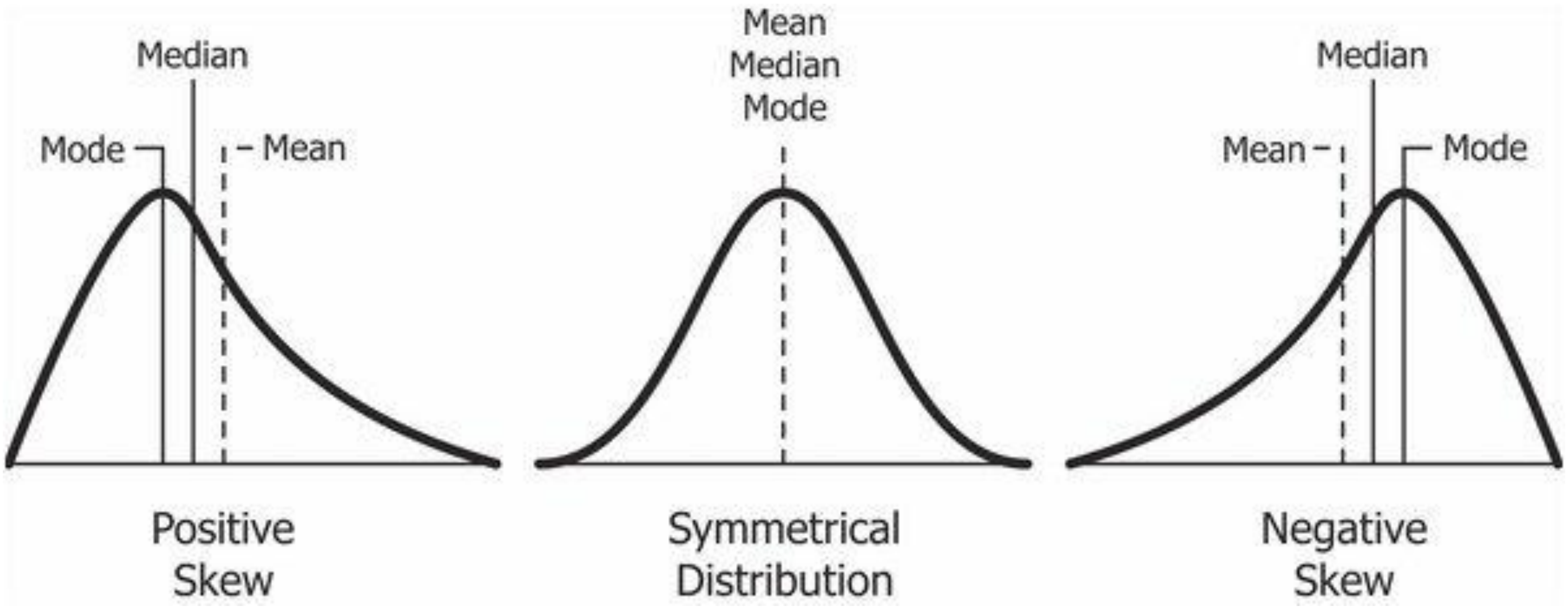
Numerical Descriptive Measure

- Numerical descriptive measure
  - ✓ Numbers that have meaningful interpretations and that can be used to describe the frequency distribution for any set of measurements



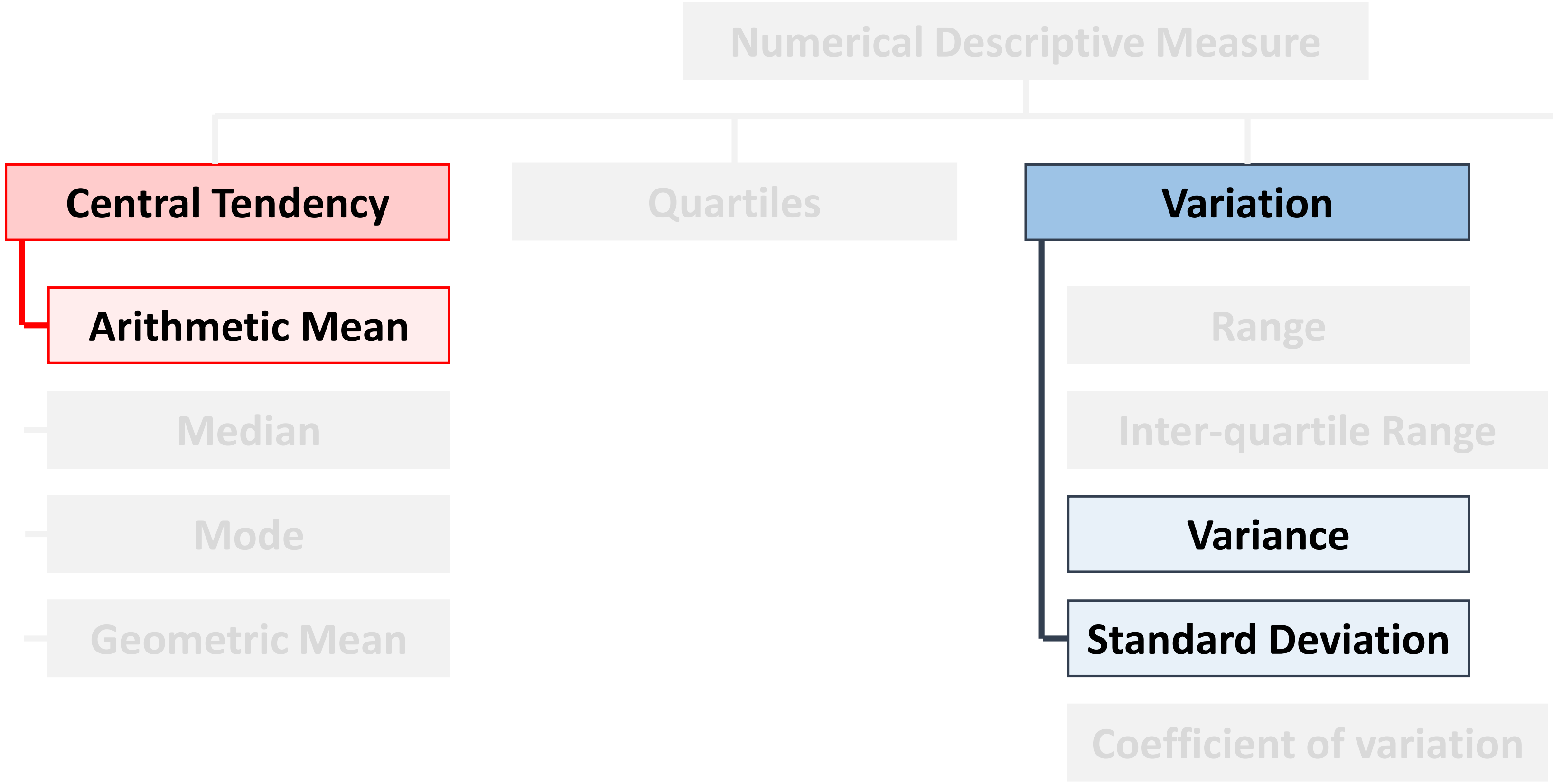
Numerical Descriptive Measure

- Numerical descriptive measure
  - ✓ Numbers that have meaningful interpretations and that can be used to describe the frequency distribution for any set of measurements



Numerical Descriptive Measure

- Numerical descriptive measure
  - ✓ Numbers that have meaningful interpretations and that can be used to describe the frequency distribution for any set of measurements



Measure of Central Tendency

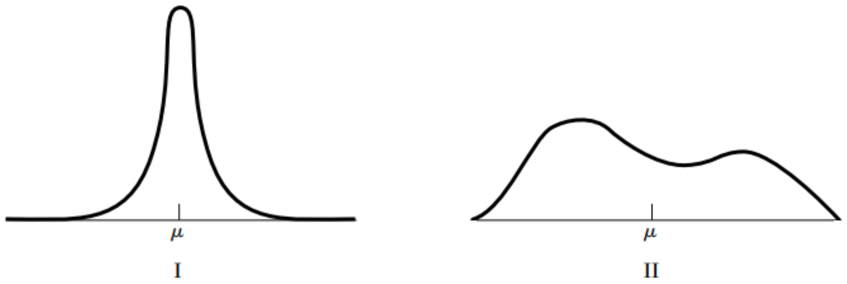
DEFINITION 1.1

The *mean* of a sample of n measured responses $y_1, y_2, \ldots, y_n$ is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The corresponding *population* mean is denoted μ

- ✓ We usually cannot measure the value of the population mean, μ; rather, μ is an unknown constant that we may want to estimate using sample information
- ✓ The mean of a set of measurements only locates the center of the distribution of data; by itself, it does not provide an adequate description of a set of measurements.
  - ➢ To describe data adequately, we must also define measures of data variability



Measure of Central Dispersion (Variation)

DEFINITION 1.2

The *variance* of a sample of n measurements $y_1, y_2, \dots, y_n$ is the sum of the square of the differences between the measurements and their mean, divided by $n - 1$. Symbolically, the sample variance is

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

The corresponding *population* variance is denoted by the symbol σ^2

- ✓ Notice that we divided by $n - 1$ instead of by n in our definition of s^2.
  - ➢ The theoretical reason for this choice of divisor is provided later, where we will show that s^2 defined this way provides a "better" estimator for the true population variance, σ^2.
- ✓ The larger the variance of a set of measurements, the greater will be the amount of variation within the set.

Measure of Central Dispersion (Variation)

DEFINITION 1.3

The *standard deviation* of a sample of measurements is the positive square root of the variance; that is,

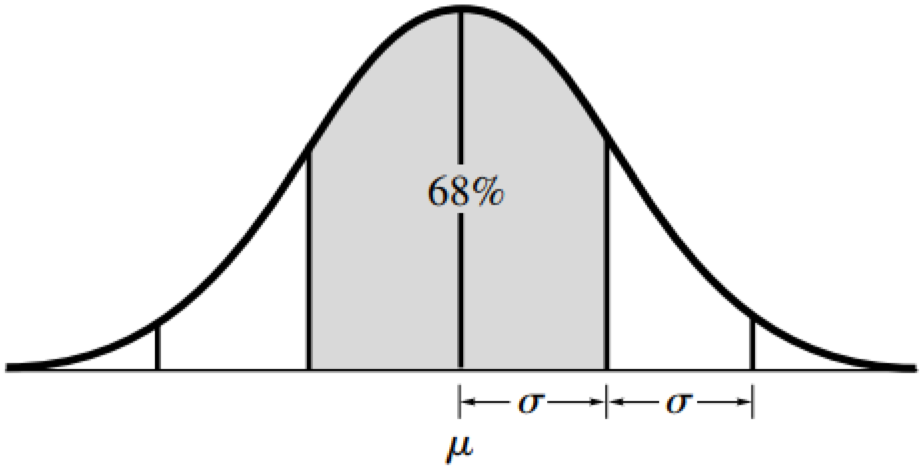
$$s = \sqrt{s^2}$$

The corresponding *population* standard deviation is denoted by $\sigma = \sqrt{\sigma^2}$.

- ✓ Although it is closely related to the variance, the standard deviation can be used to give a fairly accurate picture of data variation for a single set of measurements.

Mound-Shaped Data Distribution

- Many distributions can be approximated by a bell-shaped frequency distribution known as a normal curve.



- Data possessing mound-shaped distributions have definite characteristics of variation, as expressed in the following statement.

  ✓ $\mu \pm \sigma$ contains approximately 68% of the measurements.
  ✓ $\mu \pm 2\sigma$ contains approximately 95% of the measurements.
  $\mu \pm 3\sigma$ contains almost all of the measurements.

Inference Process

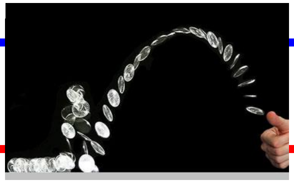
- The mechanism instrumental in making inferences can be well illustrated by analyzing our own intuitive inference-making procedures.
- Suppose that two candidates are running for a public office in our community and that we wish to determine whether our candidate, Jones, is favored to win.
  - ✓ Phone calls to check the probability of winning of Jones
  - ✓ Check if the relative frequency is > 0.5
  - ✓ Results: sample size 20, all favor Jones
- What is our conclusion?
- Reasoning
  - ✓ *It is not impossible* to draw 20 out of 20 favoring Jones when less than 50% of the electorate favor him, *but it is highly improbable*.
    - ➢ We conclude that he would win : the relative frequency is > 0.5
- This example illustrates *the potent role played by probability in making inferences*.

Probability vs. Statistics

- Probabilists assume that they know the structure of the population of interest and use the theory of probability to compute the probability of obtaining a particular sample.
- Statisticians use probability to make the *trip in reverse*—from the sample to the population
- Basic to inference making is the problem of calculating the probability of an observed sample.
  - ✓ As a result, probability is the mechanism used in making statistical inferences.
  - ✓ This is why we learn the probability first!!!!

***Probability theory* compute the probability of the sample measurements**

Probability Model
θ (parameters) :
characteristics of a model



Sample Data
$y = (y_1, \ldots, y_n)$:
Measurements

***Statistics* infer the causes (i.e., parameters) that generated the observed data (samples)**

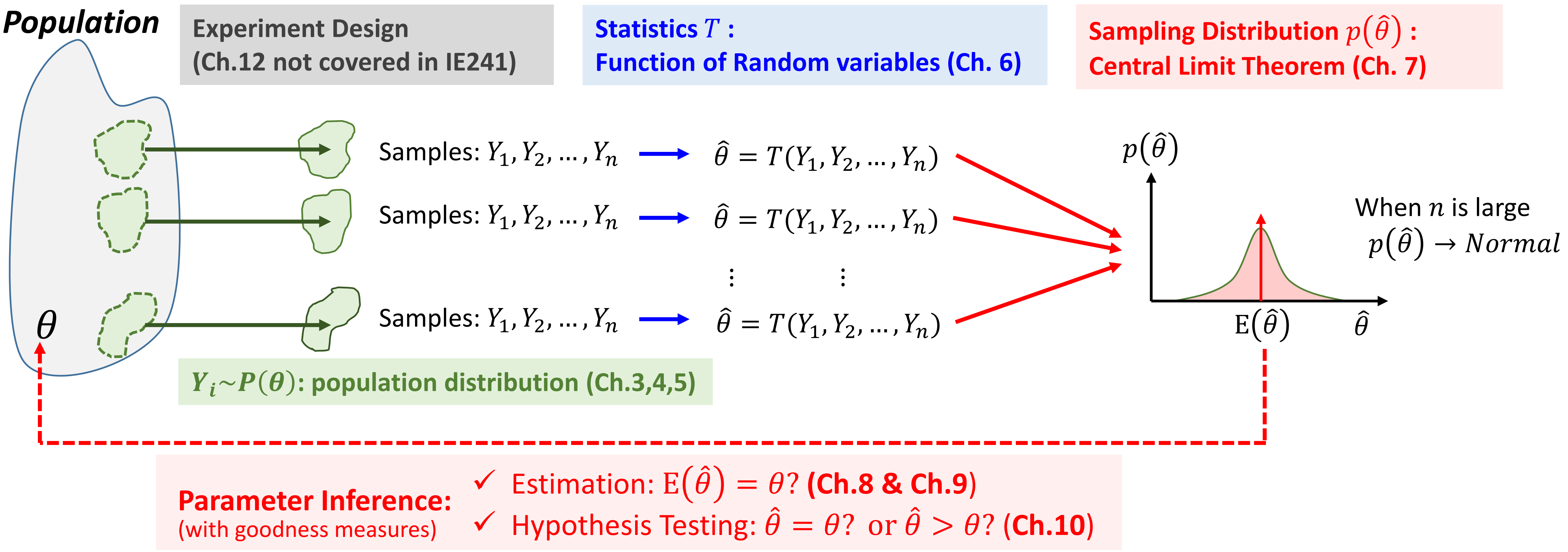
θ : Probability of having a head for each coin tossing

$\text{ex.}:\ (\text{Head}, \text{Head}, \text{Tail}, \ldots)$

Why Theory is Necessary?

- A theory is a model or an approximation to reality.
- It’s built on simplifying assumptions.
- But it aims to provide good and useful information about reality.

Overview of IE241



- **Probability Theory (Ch.2 ~ Ch.6) plays an important role in inference** by computing the probability of the occurrence of the sample and connects the computed probability to the most probable target parameter.
- ***Estimator*** $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$ for a target parameter θ is a function of the random variables observed in a sample and therefore itself is a random variable.
- Sampling distribution $p(\hat{\theta})$ can be used to evaluate the goodness of the ***estimator*** (confidence interval) and the errors (i.e., α and β errors) of ***hypothesis testing***.

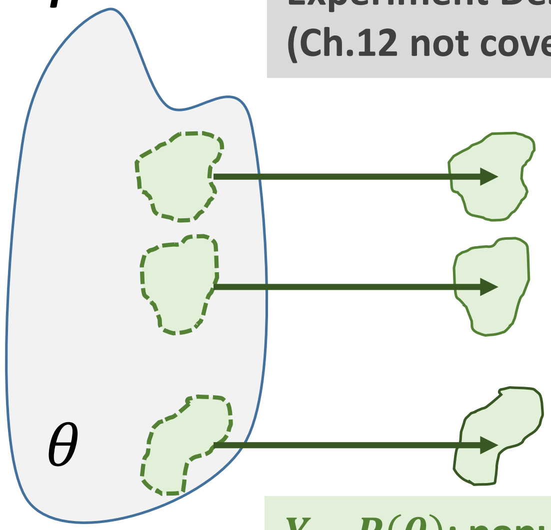
Overview of IE241

Population

Experiment Design
(Ch.12 not covered in IE241)

Statistics T :
Function of Random variables (Ch. 6)

Sampling Distribution $p(\hat{\theta})$:
Central Limit Theorem (Ch. 7)



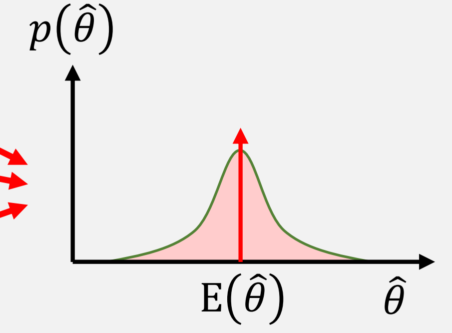
Samples: $Y_1, Y_2, \dots, Y_n$ → $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$

Samples: $Y_1, Y_2, \dots, Y_n$ → $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$

$\vdots$ $\vdots$

Samples: $Y_1, Y_2, \dots, Y_n$ → $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$

$Y_i \sim P(\theta)$: population distribution (Ch.3,4,5)



Decision making problems

$$\max_{x}.\ \mathrm{E}_{\theta}[f(x;\theta)]$$