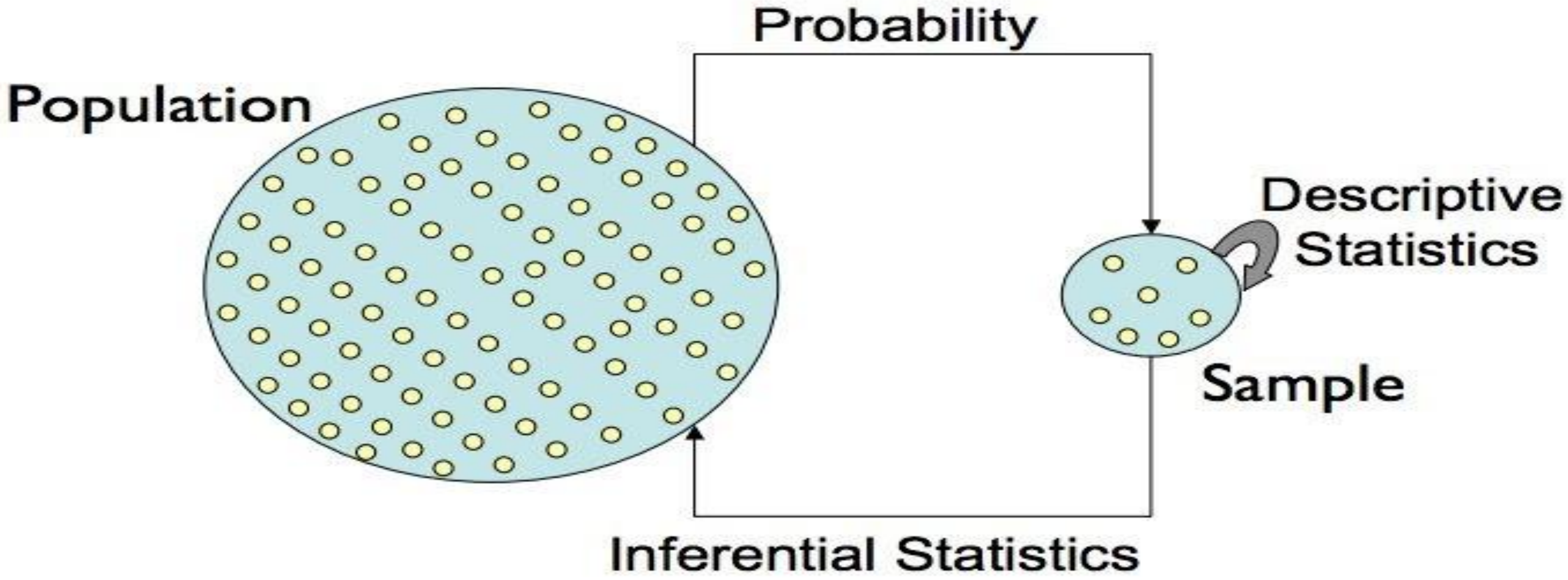
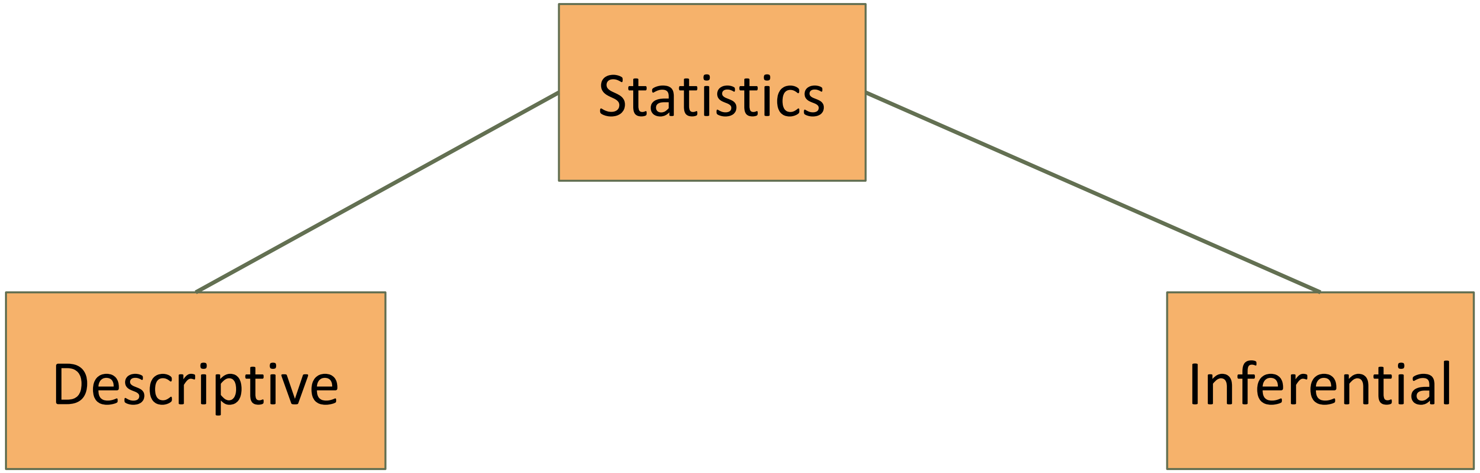
CAP4784
Introduction to Data Analytics

Statistics

The Central Dogma of Statistics



Descriptive vs. Inferential Statistics



Statistics

Descriptive

Inferential

Presenting Organizing and summarizing data
e.g., Mean describes data you have but can't be generalized beyond that

Drawing conclusions about a population from sample data
e.g., t-test, that enable inferences about the population beyond our data

Basic Concepts

- Populations and samples
- Data sets Variables and observations

Populations and Samples

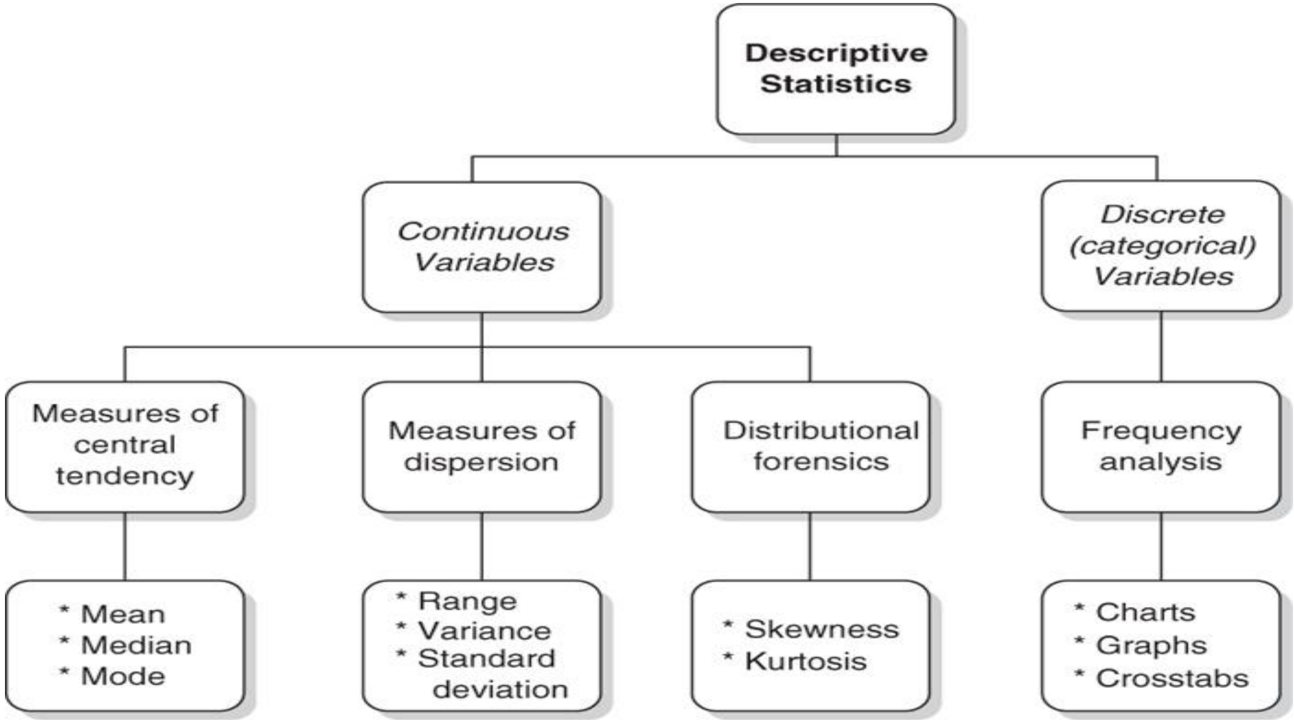
- A **population** includes all of the entities of interest in a study (people, households, machines, etc.)
    - Examples
        - All potential voters in a presidential election
        - All subscribers to cable television
        - All invoices submitted for Medicare reimbursement by nursing homes
- A **sample** is a subset of the population, often randomly chosen and preferably representative of the population as a whole.

Data Sets, Variables, and Observations

- A **data set** is usually a rectangular array of data, with variables in columns and observations in rows
- A **variable** (or field or attribute) is a characteristic of members of a population, such as height, gender, or salary
- An **observation** (or case or record) is a list of all variable values for a single member of a population

Descriptive Statistics

- Descriptive statistics provides ways to capture the properties of a given data set / sample.
  - **Central tendency measures** describe the center around the data is distributed
  - **Variation or variability measures** describe data spread, i.e., how far the measurements lie from the center.



Descriptive Statistics(1)

- **Mean** (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{12+24+47+12+84}{5} = 35.8$$

n is sample size and N is population size

  - ○ Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - ○ Trimmed mean: chopping extreme values

- **Median**:
  - ○ Middle value if odd number of values, or average of the middle two values otherwise
  - ○ Estimated by interpolation (for grouped data):

12, 12, 24, 47, 84 = 24

$$median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

□

12, 12, 24, 47, 61, 84 = 35.5

Descriptive Statistics(2)

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

12, 12, 24, 47, 84 = 12

- **Midrange** is average of largest and smallest values.

$$mean - mode = 3 \times (mean - median)$$

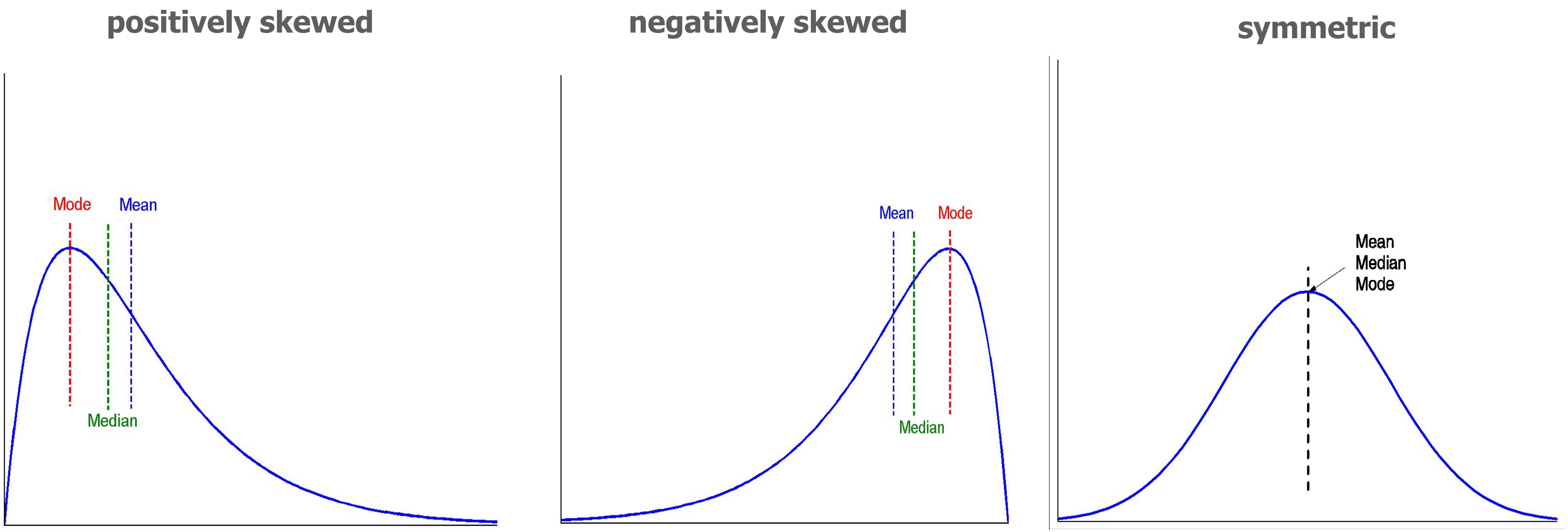
12, 12, 24, 47, 84 = (84 + 12) / 2 = 48

Descriptive Statistics(3)

- Which Measure is Best?
  - Mean is meaningful for symmetric distributions without outliers: e.g., height and weight.
  - Median is better for skewed distributions or data with outliers: e.g., wealth and income.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



positively skewed

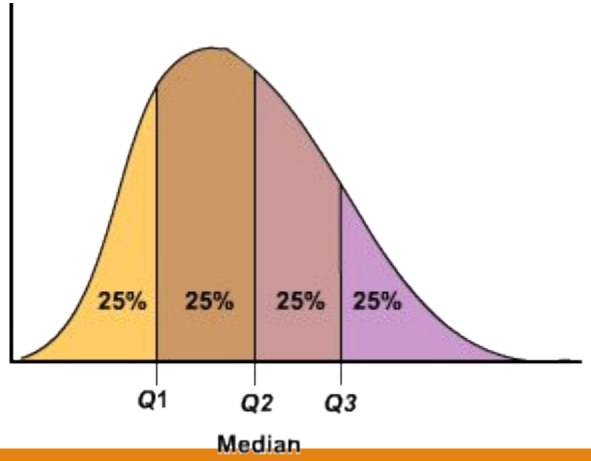
negatively skewed

symmetric

Measuring the Dispersion of Data: Quartiles

- Quartiles
  - A **percentile** is the position of an observation in the dataset relative to the other observations in the data set. Specifically, the percentile represents the percentage of the sample that falls below this observation.
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, median, Q_3, max
  - Outlier: usually, a value higher/lower than 1.5 x IQR

Percentile	Alternate Names	Interpretation
25th percentile	• Lower Quartile (QL) • First Quartile (Q1)	25% of the data falls below this percentile
50th percentile	• Median • Second Quartile (Q2)	50% of the data falls below this percentile
75th percentile	• Upper Quartile (QU) • Third Quartile (Q3)	75% of the data falls below this percentile



Measuring the Dispersion of Data: Variance

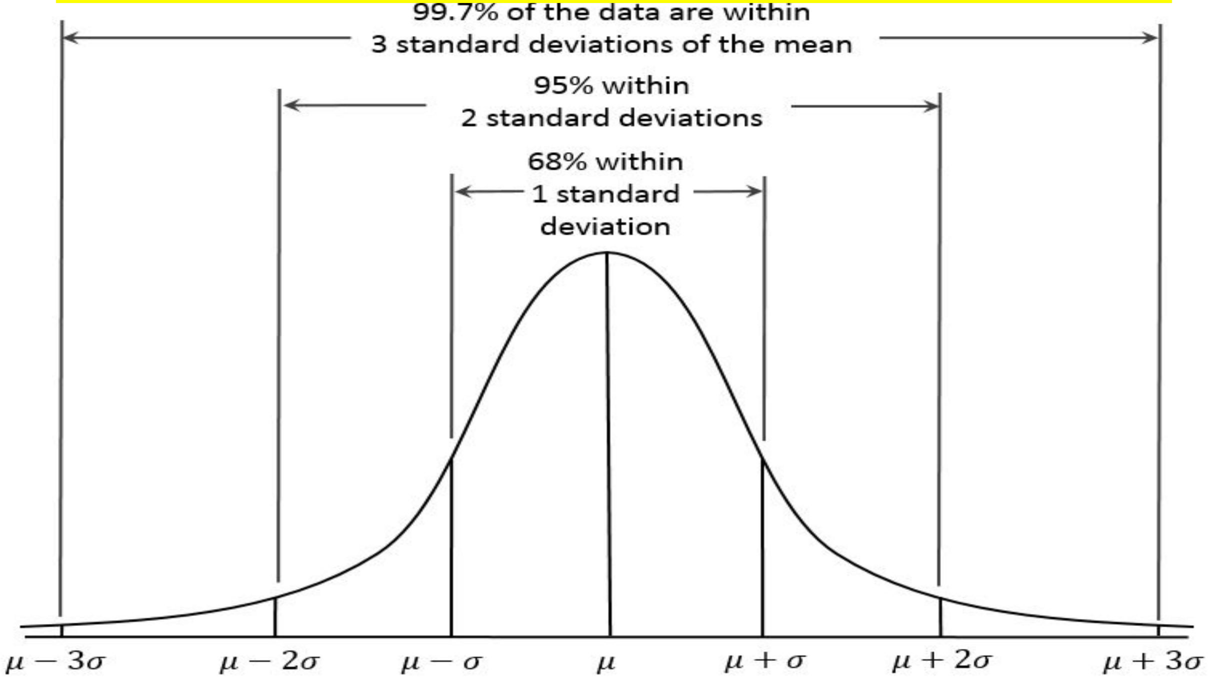
- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
  - Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

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

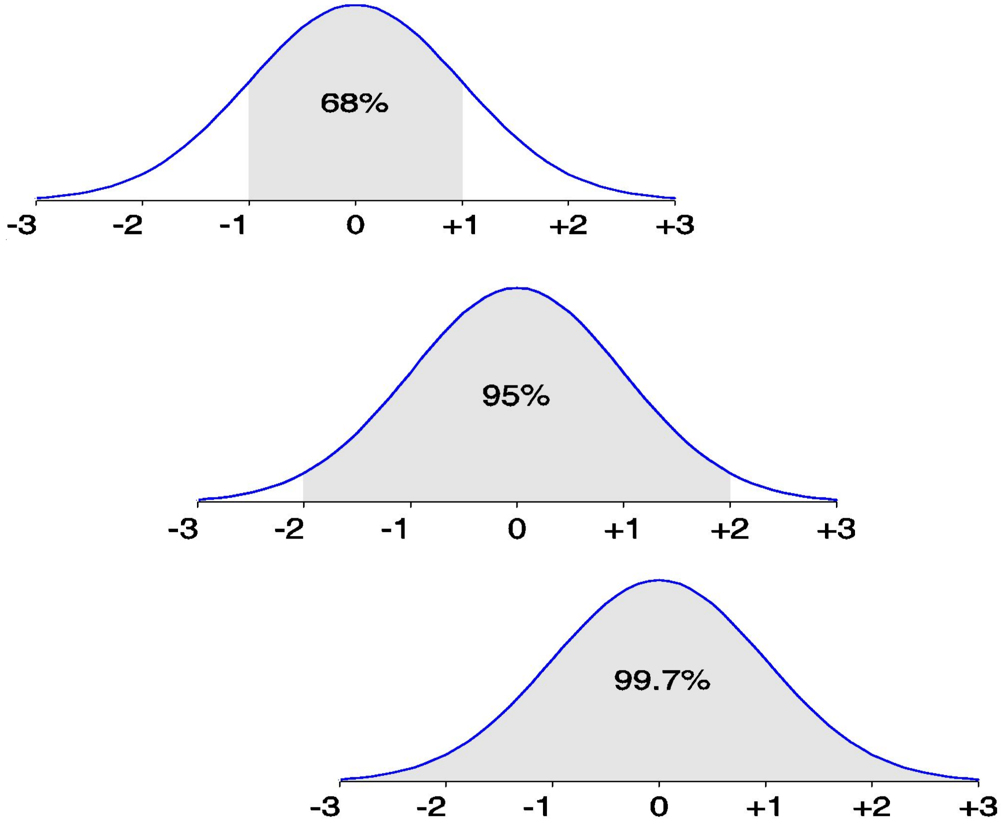
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} {x_i}^2 - \mu^2$$

Properties of Normal Distribution Curve

← — ————Represent data dispersion, spread — ————→



Represent central tendency



Probability

- Probability theory provides a formal framework for reasoning about the likelihood of events.

- The probability p(s) of an outcome s satisfies:
  - *0 <= p(s) <= 1*

$$\sum_{s \in S} p(s) = 1$$

Probability(1)

- A probability is a number between 0 and 1 that measures the likelihood that some event will occur
    - An event with probability 0 cannot occur, whereas an event with probability 1 is certain to occur
    - An event with probability greater than 0 and less than 1 involves uncertainty, and the closer its probability is to 1, the more likely it is to occur
- Probabilities are sometimes expressed as percentages or odds, but these can be easily converted to probabilities on a 0-to-1 scale

Probability vs. Statistics

- Probability deals with predicting the likelihood of future events, while statistics analyzes the frequency of past events.
- Probability is theoretical branch of mathematics on the consequences of definitions, while statistics is applied mathematics trying to make sense of real-world observations.

Probability and Probability Distributions

- A key aspect of solving real business problems is dealing appropriately with uncertainty
  - This involves recognizing explicitly that uncertainty exists and using quantitative methods to model uncertainty
- In many situations, the uncertain quantity is a numerical quantity. In the language of probability, it is called a random variable
- A probability distribution lists all of the possible values of the random variable and their corresponding probabilities

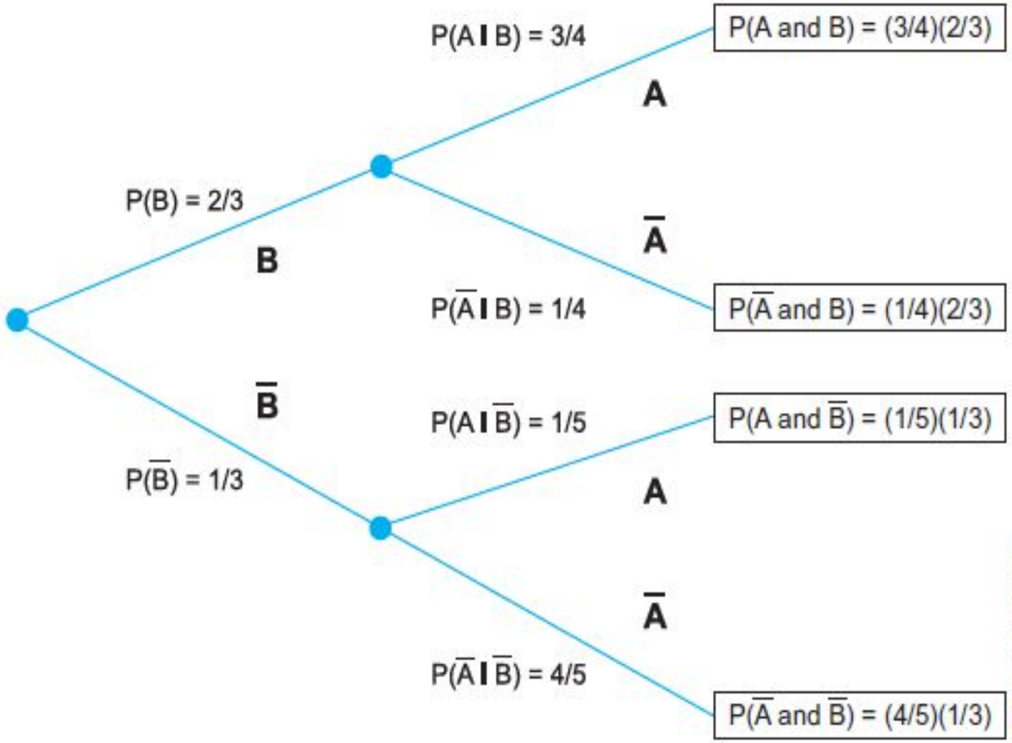
Uncertainty

- Uncertainty and risk are sometimes used interchangeably, but they are not really the same
  - You typically have no control over uncertainty; it is something that simply exists
  - In contrast, risk depends on your position
    - Even if something is uncertain, there is no risk if it makes no difference to you

Example: Assessing Uncertainty at Bender Company

- Objective: To apply probability rules to calculate the probability that Bender will meet its end-of-July deadline, given the information it has at the beginning of July.
- Solution: Let A be the event that Bender meets its end-of-July deadline, and let B be the event that Bender receives the materials it needs from its supplier by the middle of July.
- Bender estimates that the chances of getting the materials on time are 2 out of 3, so that P(B) = 2/3.
- Bender estimates that if it receives the required materials on time, the chances of meeting the deadline are 3 out of 4, so that P(A|B) = 3/4.
- The uncertain situation is depicted graphically in the form of a probability tree.
- The addition rule for mutually exclusive events implies that:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \overline{B}) = 1/2 + 1/15 = 17/30 = 0.5667$$


$P(A \mid B) = 3/4$
$P(A \text{ and } B) = (3/4)(2/3)$
A
$P(B) = 2/3$
B
$\bar{A}$
$P(\bar{A} \mid B) = 1/4$
$P(\bar{A} \text{ and } B) = (1/4)(2/3)$
$\bar{B}$
$P(A \mid \bar{B}) = 1/5$
$P(A \text{ and } \bar{B}) = (1/5)(1/3)$
$P(\bar{B}) = 1/3$
A
$\bar{A}$
$P(\bar{A} \mid \bar{B}) = 4/5$
$P(\bar{A} \text{ and } \bar{B}) = (4/5)(1/3)$

© Cengage Learning

Compound Events and Independence

- Suppose half of the students are female (event A)
- Half of the students are above median (event B)
- What is the probability a student is both A & B?
- Events A and B are independent iff
  - Independence (zero correlation) is good to simplify calculations but bad for prediction.

Rule of Complements

- The simplest probability rule involves the complement of an event
- If A is any event, then the complement of A, denoted by
  - A-bar or in some books by A^c, is the event that A does not occur
- If the probability of A is P(A), then the probability of its complement is given by the equation below.

$$P(\overline{A}) = 1 - P(A)$$

Addition Rule

- Events are mutually exclusive if at most one of them can occur—that is, if one of them occurs, then none of the others can occur.
- Exhaustive events means they exhaust all possibilities—one of the events must occur.
- The addition rule of probability involves the probability that at least one of the events will occur.
  - When the events are mutually exclusive, the probability that at least one of the events will occur is the sum of their individual probabilities:

$$P(\text{at least one of } A_1 \text{ through } A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

Conditional Probability and the Multiplication Rule

- A formal way to revise probabilities on the basis of new information is to use conditional probabilities.
- Let A and B be any events with probabilities P(A) and P(B). If you are told that B has occurred, then the probability of A might change.
  - The new probability of A is called the conditional probability of A given B, or P(A|B).
  - It can be calculated with the following formula:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Conditional Probability and the Multiplication Rule

- The numerator in this formula is the probability that both A and B occur. This probability must be known to find P(A|B)
- However, in some applications, P(A|B) and P(B) are known. Then you can multiply both sides of the equation by P(B) to obtain the multiplication rule for P(A and B):

$$P(A \text{ and } B) = P(A|B)\,P(B)$$

Probabilistic Independence

- There are situations where the probabilities P(A), P(A|B), and P(A|B) are equal. In this case, A and B are probabilistically independent events
    - This does not mean that they are mutually exclusive
    - It means that knowledge of one event is of no value when assessing the probability of the other

- When two events are probabilistically independent, the multiplication rule simplifies to: $P(A \text{ and } B) = P(A)P(B)$

- To tell whether events are probabilistically independent, you need empirical data

Equally Likely Events

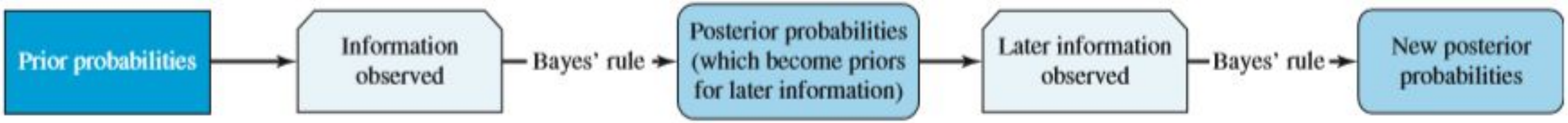
- In many situations, outcomes are equally likely (e.g., flipping coins, throwing dice, etc.)
- Many probabilities, particularly in games of chance, can be calculated by using an equally likely argument
- However, many other probabilities, especially those in business situations, cannot be calculated by equally likely arguments, simply because the possible outcomes are not equally likely.

Subjective versus Objective Probabilities

- Objective probabilities are those that can be estimated from long-run proportions
- The relative frequency of an event is the proportion of times the event occurs out of the number of times the random experiment is run
  - A famous result called the law of large numbers states that this relative frequency, in the long run, will get closer and closer to the “true” probability of an event
- However, many business situations cannot be repeated under identical conditions, so you must use subjective probabilities in these cases
  - A subjective probability is one person’s assessment of the likelihood that a certain event will occur.

Bayes’ rule

- Bayes’ rule is a formal mathematical mechanism for updating probabilities as new information becomes available
  - The original probabilities are called prior probabilities. Then information is observed and Bayes’ rule is used to update the prior probabilities to posterior probabilities
  - The actual updating mechanism can be done in two ways: with frequencies (counts) or with probabilities


Prior probabilities
Information observed
Bayes’ rule
Posterior probabilities (which become priors for later information)
Later information observed
Bayes’ rule
New posterior probabilities

Bayes' rule(1)

- Bayes' rule: Probability approach
  - For any possible outcome O, we let P(O) be the probability of O
  - If we want to indicate that new information, I, is available, we write the probability as P(O|I). This is called a conditional probability
  - The typical situation is that there are several outcomes such as "good market" and "bad market"
    - In general, denote these outcomes as O1 to On, assuming there are n possibilities.

Bayes’ rule(2)

- We start with prior probabilities $P(O_1)$ to $P(O_n)$, n probabilities that sum to 1
- Next, we observe new information, I, such as a market prediction, and we want the posterior probabilities $P(O_1 |I)$ to $P(O_n|I)$, an updated set of n probabilities that sum to 1
- We assume that the “opposite” conditional probabilities, $P(I| O_1)$ to $P(I| O_n)$, are given. In Bayesian terminology, these are called likelihoods
  - Unfortunately, these likelihoods are not what we need in the decision tree.
  - Bayes’ rule is a formal rule for turning these conditional probabilities around

Bayes’ rule(3)

- Bayes’ rule is given by $P(O_i|I) = \dfrac{P(I|O_i)P(O_i)}{P(I|O_1)P(O_1) + \cdots + P(I|O_n)P(O_n)}$

- The denominator in Bayes’ rule is the probability P(I) of the information outcome. It is sometimes called the law of total probability.

$$P(I) = P(I|O_1)P(O_1) + \cdots + P(I|O_n)P(O_n)$$

- In the case where there are only two Os, labeled as O and Not O, Bayes’ rule takes the following form:

$$P(O|I) = \frac{P(I|O)P(O)}{P(I|O)P(O) + P(I|\text{Not } O)P(\text{Not } O)}$$