CHAPTER 1

What is Statistics?

- Statistics is a way to get information from data.

Other relative info

- Mean (average)
- Median (middle point of a number set) ex. 1,2,3,4,5,6,7,8,9,10 median is 5
- Range= Max – min

Descriptive statistics

- Descriptive statistics deals with methods of organizing, summarizing, and presenting data in a convenient and informative way.
- One form of descriptive statistics uses **graphical techniques**, which allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information.
- Another form of descriptive statistics uses **numerical techniques** to summarize data.
- The mean and median are popular numerical techniques to describe the location of the data.
- The range, variance, and standard deviation measure the variability of the data
- Descriptive Statistics describe the data set that's being analyzed, but doesn't allow us to draw any conclusions or make any interferences about the data.

Inferential Statistics

- Inferential statistics is a body of methods used to draw conclusions or inferences about characteristics of populations based on sample data
- We can sample a much smaller number of population and infer from the data the number of the answer.
- Is used to draw conclusions or inferences about characteristics of populations based on data from a sample.

Key concepts

- Population
  - A population is the group of all items of interest to a statistics practitioner.
  - Frequently very large; sometimes infinite.
  - Ex. All 5 million Florida voters.
- Sample

  - A sample is a subset of data drawn from the population.
  - Potentially very large, but less than the population
  - Ex. A sample of 765 voters
- Parameters
  - A descriptive measure of a population.
- Statistic
  - A descriptive measure of a sample
  - We use statistic to make inferences about parameters

- Statistical inference
  - Is the process of making an estimate, prediction, or decision about a population based on a sample
  - We use statistics to make inferences about parameters. Therefore, we can make an estimate, prediction, or decision about a population based on sample data. Thus, we can apply what we know about a sample to the larger population from which it was drawn
  - **Rationale**
    - Large populations make investigating each member impractical and expensive.
    - Easier and cheaper to take a sample and make estimates about the population from the sample.
  - **However**
    - Such conclusions and estimates are not always going to be correct.
    - For this reason, we build into the statistical inference "measures of reliability", namely **confidence level** and **significance level.**
  - **Confidence & Significance Levels**
    - The confidence level is the proportion of times that an estimating procedure will be correct.
    - When the purpose of the statistical inference is to draw a conclusion about a population, the significance level measures how frequently the conclusion will be wrong in the long run.
    - 1= con + sig

Statistical Applications in Business

- Statistical analysis plays an important role in virtually *all* aspects of business and economics.

Chapter 2

Definitions

- **variable** - some characteristic of a population or sample
  - Typically denoted with a capital letter: X, Y, Z
  - Ex. Student Grades
- **Values** of the variable are the range of possible values for a variable.
  - Ex. student marks (0..100)
- **Data** are the observed values of a variable
  - Ex. student marks: {67, 74, 71, 83, 93, 55, 48}

Types of Data & Information

- Data fall into three main groups:
  - Interval, Nominal and ordinal Data.
- **Interval data**
  - Are real Numbers ex. heights, weights, prices, etc
  - Also referred to as quantitative or numerical
  - Arithmetic operations can be performed on Interval Data, thus its meaningful to talk about 2*Height, or Price + $1, and so on.
- **Nominal data**
  - The values of nominal data are categories.
  - Ex. responses to questions about marital status, coded as: Single = 1, Married = 2, Divorced = 3, Widowed = 4
  - These data are categorical in nature; arithmetic operations don't make any sense (e.g. does Widowed ÷ 2 = Married?
  - Nominal data are also called qualitative or categorical.
- **Ordinal data**
  - Ordinal Data appear to be categorical in nature, but their values have an order; a ranking to them:
  - It's still not meaningful to do arithmetic on this data.
  - Order is maintained no matter what numeric values are assigned to each category.

Calculations for Types of Data

- All calculations are permitted on interval data.
- Only calculations involving a ranking process are allowed for ordinal data
- No calculations are allowed for nominal data, save counting the number of observations in each category.
- This leads to the hierarchy of data.

Hierarchy of data

- Interval
  - Values are real numbers.

- All calculations are valid.
- Data may be treated as ordinal or nominal.

- Ordinal
  - Values must represent the ranked order of the data.
  - Calculations based on an ordering process are valid.
  - Data may be treated as nominal but not as interval.
- Nominal
  - Values are the arbitrary numbers that represent categories.
  - Only calculations based on the frequencies of occurrence are valid.
  - Data may not be treated as ordinal or interval.

Graphical & Tabular Techniques for Nominal Data

- The only allowable calculation on nominal data is to count the frequency of each value of the variable.
- ***Frequency distribution:*** summarize the data in a table that presents the categories and their counts.
  - Bar chart often used to represent F.D
- **R*elative frequency distribution (f.r/total):*** lists the categories and the proportion of frequency out of all counts.
  - Pie chart used to represent R.F.D

Describing the Relationship between Two Nominal Variables

- To describe the relationship between two nominal variables, we must remember that we are permitted only to determine the frequency of the values.
- As a first step we need to produce a cross-classification table, which lists the frequency of each combination of the values of the two variables

Cross-Classification Table of Frequencies

2 variable

	Newspaper				
Occupation	*G&M*	*Post*	*Star*	*Sun*	Total
Blue collar	27	18	38	37	120
White collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	89	112	81	72	354

1 Variable

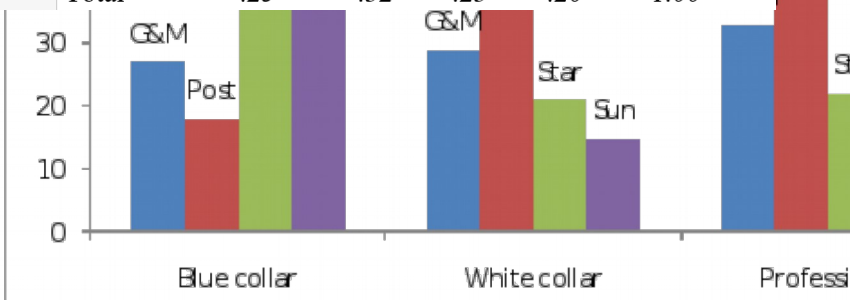
Row Relative Frequencies

variable 2

Newspaper

Occupation	*G&M*	*Post*	*Star*	*Sun*	Total
Blue collar	.23	.15	.32	.31	1.00
White collar	.27	.40	.19	.14	1.00
Professional	.26	.40	.17	.16	1.00
Total	.25	.32	.23	.20	1.00

variable 1



Graphing the Relationship between 2 nominal variables

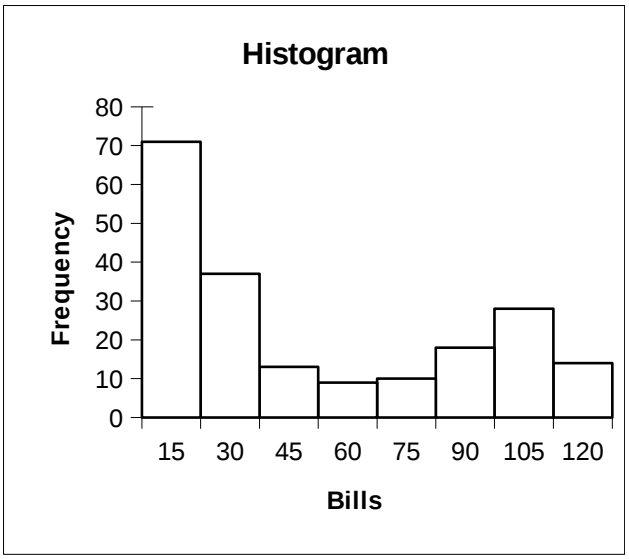
Interpret

- If the two variables are unrelated, the patterns exhibited in the bar charts should be approximately the same. If some relationship exists, then some bar charts will differ from others.

Chapter 3

The histogram

- The graph is called a **histogram**. A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.



Determining the Number of Class Intervals

- Determine the class using the table □□□□□
- Or using the Sturges' formula:
  Number of class intervals = **1 + 3.3 log (n)**

Approximate Number of Classes in Frequency Distributions

Number of Observations	Number of Classes
Less than 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11
1,000 - 5,000	11 - 13
5,000 - 50,000	13 - 17
More than 50,000	17 - 20

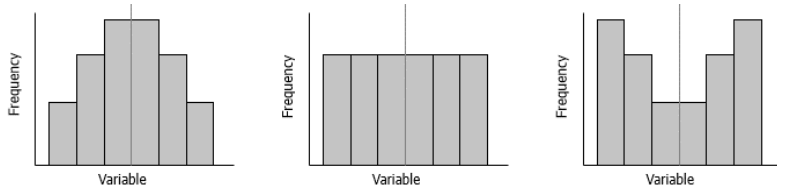
Class interval width

$$\text{Class width} = \frac{\textit{Largest Observation} - \textit{Smallest Observation}}{\textit{Number of Classes}}$$

Building a histogram

- Histogram how to build
  1. Collect Data
  2. Create frequency
     a) Determine the number of class (table or struges formula)
     b) Determine the Class width
- Classes contain observations greater than their lower limits (except for first class) and less than or equal to their upper limits. Ex. (first class =Amounts that are less than or equal to 15, second class= Amounts that are more than 15 but less than or equal to 30, and so on….)

Shapes of histogram

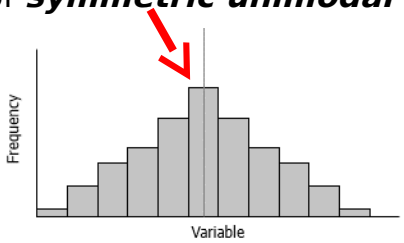
- Symmetry
  - A histogram is said to be ***symmetric*** if, when we draw a **vertical line** down the center of the histogram, the two sides are identical in shape and size:
- 

Skewness
  - A skewed histogram is one with a long tail extending to either the right or the left:

- Modality
  - A ***unimodal*** histogram is one with a single peak, while a ***bimodal*** histogram is one with two peaks:

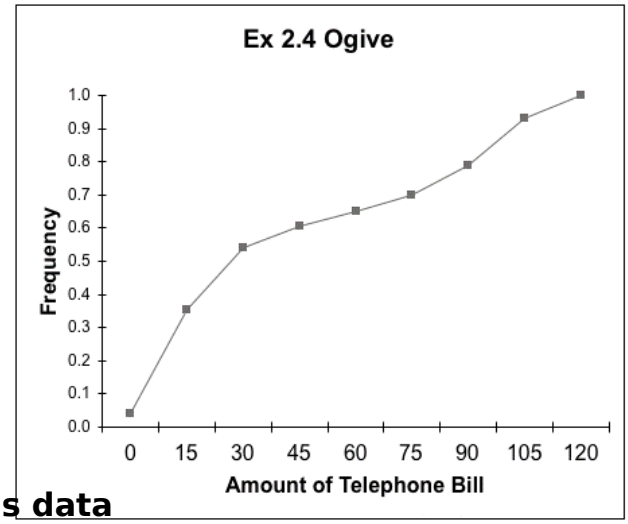
- Bell Shape

- A special type of ***symmetric unimodal*** histogram is one that is bell shaped



Ogive

- is a graph of a ***cumulative frequency distribution*** (keep adding previous frequency to the next)
- We create an ogive in three steps
  1. from the frequency distribution created earlier, calculate ***relative frequencies***
     - Relative Frequency = # of observations in a class / Total # of observations
  2. Calculate ***cumulative relative frequencies*** by adding the current class' relative frequency to the previous class' cumulative relative frequency
  3. Graph the cumulative relative frequencies



Describing time series data

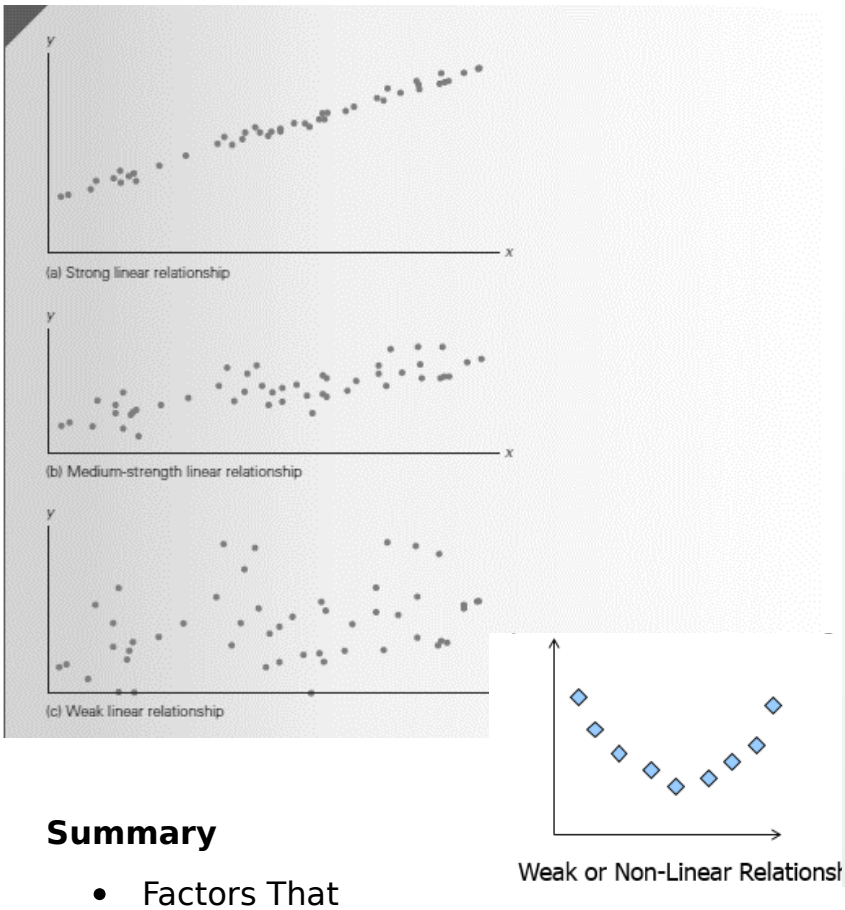
- Besides classifying data by type, we can also classify them according to whether the observations are measured at the same time or whether they represent measurements at successive points in time.
- Observations measured at the same point in time are called ***cross-sectional*** **data**.
- Observations measured at successive points in time are called ***time-series*** **data.**

- Time-series data graphed on a *line chart*, which plots the value of the variable on the vertical axis against the time periods on the horizontal axis.

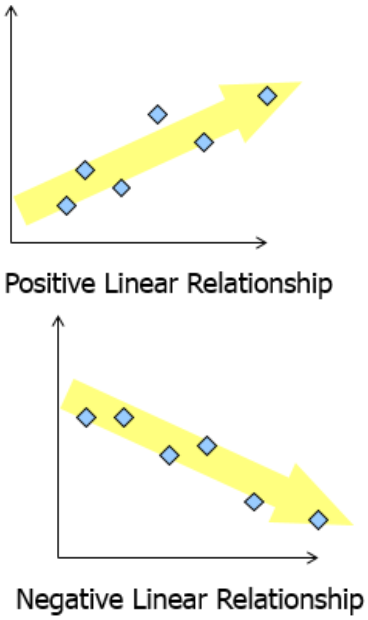
Graphing the Relationship between Two Interval Variables

- Moving from nominal data to interval data, we are frequently interested in how two interval variables are related.
- To explore this relationship, we employ a ***scatter diagram***, which plots two variables against one another.
- The ***independent*** variable is labeled X and is usually placed on the horizontal axis, while the other, ***dependent*** variable, Y, is mapped to the vertical axis.

Linear relationship



Patterns of



Scatter Diagrams

Summary

- Factors That Identify When to Use Frequency and Relative Frequency Tables, Bar and Pie Charts
  1. Objective: Describe a single set of data.
  2. Data type: Nominal
- Factors That Identify When to Use a Histogram, Ogive.
  1. Objective: Describe a single set of data.

2. Data type: Interval

- Factors that Identify When to Use a Cross-classification Table
  1. Objective: Describe the relationship between two variables.
  2. Data type: Nominal
- Factors that Identify When to Use a Scatter Diagram
  1. Objective: Describe the relationship between two variables.
  2. Data type: Interval

	Interval Data	Nominal Data
Single Set of Data	Histogram	Frequency and Relative Frequency Tables, Bar and Pie Charts
Relationship Between Two Variables	Scatter Diagram	Cross-classification Table, Bar Charts

Chapter 4

Numerical Descriptive Techniques

- Measures of **Central Location**
  - Mean, Median, Mode
- Measures of **Variability**
  - Range, Standard Deviation, Variance, Coefficient of Variation
- Measures of **Relative Standing**
  - Percentiles, Quartiles
- Measures of **Linear Relationship**
  - Covariance, Correlation, Determination, Least Squares Line

Measures of Central Location

- **Arithmetic mean**
  - The *arithmetic mean*, *average*, shortened to *mean*, is the most popular & useful measure of central location.
  - It is computed by simply adding up all the observations and dividing by the total number of observations
  - is seriously affected by extreme values called “outliers”.

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{Number of observations}}$$

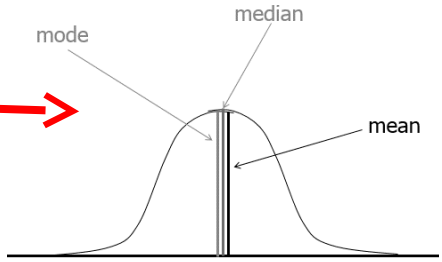
Population Mean = $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

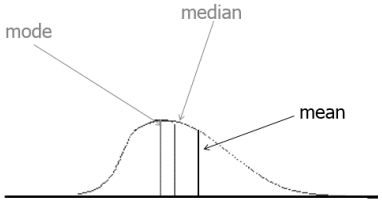
Sample Mean = $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

- Notation
  - Number of observations in a *population*: **N**
  - Number of observations in a *sample*: **n**
  - *Population Mean* is denoted with Greek letter "**mu**":
  - *Sample mean* is denoted with an "**x-bar**":
- **Median**
  - The ***median*** is calculated by placing all the observations in order; the observation that falls in the *middle* is the median.
  - For even number the average of two middle numbers is the media.
    - Ex. 1,2,3,4 median= 2+3/2 = 2.5
- **Mode**
  - The ***mode*** of a set of observations is the value that occurs most *frequently*.
  - A set of data may have one mode (or modal class), or two, or more modes.
  - Mode is a useful for all data types, though mainly used for nominal data.
  - For large data sets the modal *class* is much more relevant than a single-value mode

Mean, Median, Mode

- If a distribution is symmetrical,
  - the mean, median and mode may coincide
- If a distribution is asymmetrical, say skewed to the left or to the right,
  - the three measures may differ





Mean, Median, Mode: Which Is Best?

- The mean is generally our first selection.
- However, there are several circumstances when the median is better. One advantage the median holds is that it is not as sensitive to extreme values as is the mean.

- The mode is seldom the best measure of central location.
- Median is better when there is a relatively small number of extreme observations (either very small or very large, but not both) the median usually produces a better measure of the center of the data.

Mean, Median, & Modes for Ordinal & Nominal Data

- For ordinal and nominal data the calculation of the mean is not valid.
- Median is appropriate for ordinal data. No median for nominal data.
- For nominal data, a mode calculation is useful for determining highest frequency but not "central location"

Measures of Central Location • Summary

- Compute the Mean to
  - Describe the central location of a single set of interval data
- Compute the Median to
  - Describe the central location of a single set of interval or ordinal data
- Compute the Mode to
  - Describe a single set of nominal data

Geometric mean

- When the variable is a growth rate or rate of change, such as the value of an investment over periods of time, we use geometric mean not arithmetic mean.

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)\ldots(1+R_n)} - 1$$

To get the answer = $x(1 + R_g)^n$

- The **geometric mean** is used whenever we wish to find the "average" growth rate, or rate of change, in a variable ***over time.***
- However, **the arithmetic mean** of n returns (or growth rates) is the appropriate mean to calculate if you wish to estimate the mean rate of return (or growth rate) for any *single* period in the future.

Factors identifying when to compute….

Factors That Identify When to Compute the Mean
1. **Objective:** Describe a single set of data
2. **Type of data:** Interval
3. **Descriptive measurement:** Central location

Factors That Identify When to Compute the Median
1. **Objective:** Describe a single set of data
2. **Type of data:** Ordinal or interval (with extreme observations)
3. **Descriptive measurement:** Central location

Factors That Identify When to Compute the Mode
1. **Objective:** Describe a single set of data
2. **Type of data:** Nominal, ordinal, interval

Factors That Identify When to Compute the Geometric Mean
1. **Objective:** Describe a single set of data
2. **Type of data:** Interval; growth rates

Measures of Variability

- **Range**
  - The ***range*** is the simplest measure of variability, calculated as:
  - Range = Largest observation – Smallest observation
  - The advantage of the **range** is its simplicity. The disadvantage is also its simplicity. Because the range is calculated from only two observations, it tells us nothing about the other observations.
- **Variance**
  - **Variance** and its related measure, **standard deviation**, are arguably the most important statistics. Used to measure variability, they also play a vital role in almost all statistical inference procedures.
  - **Population variance** is denoted by σ^2
  - **Sample variance** is denoted by s^2

The variance of population is

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

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

The Variance of Sample is

n-1= Sample size x-bar= sample mean

- *A short-cut formulation* to calculate sample variance directly from the data without the intermediate step of calculating the mean

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

- **Standard Deviation**
  - The standard deviation is simply the square root of the variance
  - The standard deviation can be used to compare the variability of several distributions and make a statement about the general shape of a distribution.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

  - **Empirical Rule** (only if histogram is bell shaped)
    - Calculated as (mean minus – # of standard deviation)
    1. Approximately 68% of all observations fall within one standard deviation of the mean.
    2. Approximately 95% of all observations fall within two standard deviations of the mean.
    3. Approximately 99.7% of all observations fall within three standard deviations of the mean.
  - **Chebysheff's Theorem** (applies to all shapes of histograms)
    - The theorem states that *at least* 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).
    1. Approximately 75% of all observations fall within two standard deviations of the mean
    2. Approximately 88.9% of all observations fall within two standard deviations of the mean

$$1 - \frac{1}{k^2} \quad for \quad k > 1$$

- **Coefficient of Variation**
  - The *coefficient of variation* of a set of observations is the standard deviation of the observations divided by their mean

- It provides a ***proportionate*** measure of variation

Population coefficient of variation = $CV = \frac{\sigma}{\mu}$ Sample coefficient of variation = $CV = \frac{s}{\bar{x}}$

Measure of relative standing and box plots

- Measures of relative standing are designed to provide information about the ***position*** of particular values ***relative*** to the entire data set.
- **Percentile**
  - the P^{th} percentile is the value for which P percent are less than that value and (100-P)% are greater than that value.
- **Quartiles**
  - We have special names for the 25^{th}, 50^{th}, and 75^{th} percentiles, namely *quartiles*.
    - First (lower) decile = 10^{th} percentile
    - First (lower) quartile, Q_1, = 25^{th} percentile
    - Second (middle)quartile,Q_2, = 50^{th} percentile
    - Third quartile, Q_3, = 75^{th} percentile
    - Ninth (upper) decile = 90^{th} percentile
- **Location of Percentiles**

$$L_P = (n+1)\frac{P}{100}$$

where L_P is the location of the P^{th} percentile

- **Interquartile Range**
  - The quartiles can be used to create another measure of variability, the *interquartile range*, which is defined as follows: Interquartile Range = $Q_3 - Q_1$
  - The interquartile range measures the spread of the middle 50% of the observations

Measures of Linear Relationship

- We now present two numerical measures of linear relationship that provide information as to the **strength & direction** of a linear relationship between two variables (if one exists).
- Covariance and coefficient of correlation
- Covariance

$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N} \quad \text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

μ_x and μ_y = mean of population x bar and y bar= mean of sample

Short cut □□□□ $$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

- When two variables move in the ***same direction*** (both increase or both decrease), the covariance will be a ***large positive number***.
- When two variables move in ***opposite directions***, the covariance is a ***large negative number***.
- When there is ***no particular pattern***, the covariance is a ***small number***.

- **Coefficient of Correlation**
  - The coefficient of correlation is defined as the covariance divided by the standard deviations of the variables

Population coefficient of correlation: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. Sample coefficient of correlation: $r = \frac{s_{xy}}{s_x s_y}$

  - The **advantage** of the coefficient of correlation over covariance is that it has fixed range from -1 to +1
  - If the two variables are very strongly positively related, the coefficient value is close to +1 (strong positive linear relationship).
  - If the two variables are very strongly negatively related, the coefficient value is close to -1 (strong negative linear relationship).
  - No straight line relationship is indicated by a coefficient close to zero.

Parameters and statistics

Parameters and Statistics

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	S^2
Standard Deviation	σ	S
Coefficient of Variation	CV	cv
Covariance	σ_{xy}	S_{xy}
Coefficient of Correlation	ρ	r

CHAPTER 6

Approaches to Assigning Probabilities

- *Classical approach*: based on equally likely events.
- *Relative frequency*: assigning probabilities based on experimentation or historical data.
- *Subjective approach*: Assigning probabilities based on the assignor's (subjective) judgment.

Classical Approach

- If an experiment has n possible outcomes, this method would assign a probability of 1/n to each outcome. It is necessary to determine the number of possible outcomes.

Relative frequency approach

- he probability of an event can be approximated by the *relative frequency* , or proportion of times that the event occurs.

$$\text{PROBABILITY (EVENT) is approximately} \quad \frac{\text{\# of times event occurs}}{\text{\# of experiments}}$$

Subjective Approach

- In the subjective approach we define probability as the degree of belief that we hold in the occurrence of an event

Interpreting probability

- No matter which method is used to assign probabilities all will be interpreted in the relative frequency approach.

Relationships between events

- Events could result from ***combining*** other events in various ways.
- There are several types of combinations and **relationships between events**:
  - Complement event
  - Intersection of events
  - Union of events
  - Mutually exclusive events
  - Dependent and independent events

Complement of an Event

- The *complement of event* A is defined to be the event consisting of all sample points that are "not in A".
- $P(A) + P(A^c) = 1$
- Complement of A is denoted as A^c

Intersection of Two Events

- The ***intersection of events* A and B** is the set of all sample points that are in **both A and B**.
- The ***joint probability*** of A and B is the probability of the intersection of A and B,
- Joint probability is denoted as P(A and B).

Union of Two Events

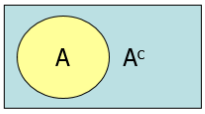
- The ***union of two events*** A and B, is the event containing all sample points that are **in A or B or both**:
- Union of A and B is denoted: **A or B**

Mutually Exclusive Events

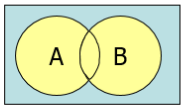
- When two events are ***mutually exclusive*** (that is the two events cannot occur together), their joint probability is 0

Basic Relationships of Probability

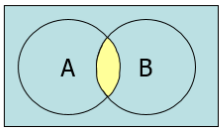
Complement of Event



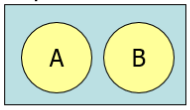
Union of Events



Intersection of Events



Mutually Exclusive Events



Marginal Probabilities

- ***Marginal probabilities*** are computed by adding across rows and down columns; that is they are calculated in the ***margins*** of the table:

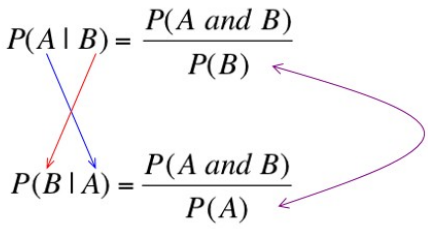
Example:

	B_1	B_2	$P(A_i)$
A_1	**.11**	**.29**	**.40**
A_2	**.06**	**.54**	**.60**
$P(B_j)$	**.17**	**.83**	**1.00**

Conditional Probability

- ***Conditional probability*** is used to determine how two events are related; that is, we can determine the probability of one event ***given the occurrence*** of another related event.
- Conditional probabilities are written as **P(A | B)** and read as "the probability of A *given* B" and is calculated as:

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



Independent and dependent events

- One of the objectives of calculating conditional probability is to determine whether two events are related.
- In particular, we would like to know whether they are ***independent***, that is, if the probability of one event is ***not affected*** by the occurrence of the other event.
- Two events A and B are said to be ***independent*** if

P(A|B) = P(A) or P(B|A) = P(B)

Probability Rules and Trees

- Complement rule
- Multiplication rule
- Addition rule

Complement rule

- As we saw earlier with the complement event, the ***complement rule*** gives us the probability of an event NOT occurring. That is:
- $P(A^C) = 1 - P(A)$

Multiplication Rule

- The ***multiplication rule*** is used to calculate the ***joint probability*** of two events. It is based on the formula for conditional probability.
- If we multiply both sides of the equation by P(B) we have: **P(A and B) = P(A | B)•P(B)**
- P(B|A) means that probability of **A** given that probability of **B** has already occurred
- Likewise, P(A and B) = P(B | A) • P(A)
- Means that Probability of **B** given that probability of **A** has already occurred.

- If A and B are independent events, then P(A and B) = P(A)•P(B)

Addition Rule

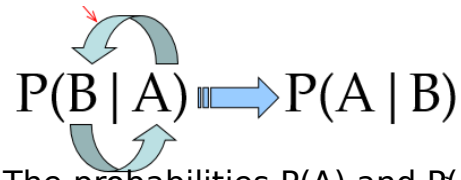
- the ***addition rule*** is used to compute the probability of event A ***or*** B ***or*** both A and B occurring
- P(A or B) = P(A) + P(B) – P(A and B)
- If A and B are mutually exclusive then we omit P(A and B). P(A or B) = P(A) + P(B)

Probability trees

- The probabilities associated with any set of branches from one "node" must add up to 1.00
- the events in an experiment are represented by lines

Bayes' Law

- Bayes' Law is named for Thomas Bayes, an eighteenth century mathematician.
- In its most basic form, if we know P(B | A),
- we can apply Bayes' Law to determine P(A | B)



- The probabilities P(A) and P(A^C) are called *prior probabilities* because they are determined *prior* to the decision about taking the preparatory course.
- The conditional probability P(A | B) is called a *posterior probability* (or revised probability), because the prior probability is revised *after* the decision about taking the preparatory course**.**

Bayer's Law

- If we know P(B|A) and P(A), we can find P(A|B) via the Bayer's law:

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

- $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$