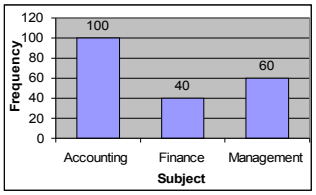
Chapter 1: What is Statistic?

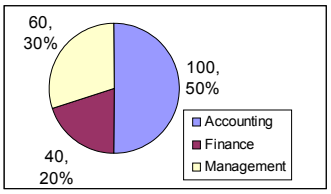
- **Statistics** is a way to get information from data. Data → Statistics → Information.
- A **Population** is the group of **all** items of interest to a statistics practitioner. E.g. All York students.
- **Parameter –** A descriptive measure of a population (mean, proportion, etc.).
- A **Sample** is a set of data drawn from population. E.g. a sample of 500 students completes a survey.
- **Statistic** – A descriptive measure of a sample (mean, proportion, etc.).
- Because populations are very large, expensive, and very time consuming, we use statistics to make inferences about parameters, i.e., we can make an estimate prediction, or decision about a population based on sample data.
- **Descriptive Statistics** are methods of organizing, summarizing, and presenting data in a convenient and informative way. These methods include: Graphical Techniques and Numerical Techniques.
- **Inferential Statistics** is set of methods used to draw conclusions or inferences about characteristics of **populations** based on data from a **sample**.
- **Statistical Inference** is the process of making an estimate, prediction, or decision about a population based on a sample.
- Conclusions and estimates are not always going to be correct, so we build into statistical inference a measure of reliability: the confidence level and the significance level.
- **Confidence Level (1-α)** is the proportion of times that an estimating procedure will be correct.
- **Significance Level (α)** measures how frequently the conclusion will be wrong in the long run.
- E.g. the poll is considered accurate within 3.4 percentage points, 19 times out of 20. Our Confidential level is 95% (19/20 =0.95), while our significance level is 5%.

Chapter 2: Graphical and Tabular Descriptive Techniques

- **Descriptive Statistics** involves arranging, summarizing, and presenting a set of data in such a way that useful information is produced (Graphical Techniques and Numerical Techniques).
- A **Variable** is some characteristic of a population or sample. E.g. Student grades (X, Y, Z …)
- The **Values** of the variable are the possible values for a variable. E.g. Student marks (0 - 100)
- **Data** are the observed values of a variable. E.g. Student mark: 75, 50, 92, 82, 45, 66 …
- There are three types of data: **Interval**, **Nominal**, and **Ordinal** data.
- **Interval Data** are real numbers (Quantitative or numerical), such as heights, incomes, distance …
- The values of **Nominal Data** are categories (Qualitative or Categorical), e.g. sex: female = 1 male = 2
- **Ordinal Data** appear to be categorical in nature, but their values have an order (Ranking), e.g. poor = 1, fair good = 2, good = 3, very good = 4, excellent = 3, etc.
- Data → Not Categorical → **Interval** (real number; all calculation; treated as ordinal or nominal)
  → Categorical → Order → **Ordinal** (ranked order; Calculation only based on ordering process; May be treated as nominal but not interval)
  → No Order → **Nominal** (arbitrary numbers represent categories, Only calculate based on frequency; May not be treated as nominal or not interval
- **Single set of Nominal Data:** Use Frequency distribution and Relative Frequency Tables, Bar and Pie Charts. E.g.

Subject	Frequency	Relative Frequency
Accounting	100	50%
Finance	40	20%
Management	60	30%
Total	200	100%

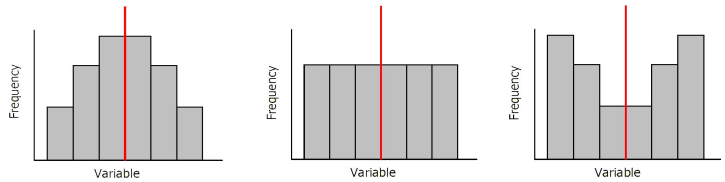




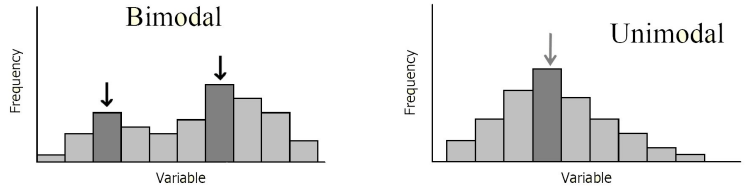
- **Histogram** is most important graphical method for **single set of Interval Data**; it doesn't only summarize interval data but also help explain probabilities.
    - Number of classes intervals = 1 + 3.3 log (n)
    - Class width = (Largest observation + Smallest observation) / Number of Classes.

Shapes of Histograms:

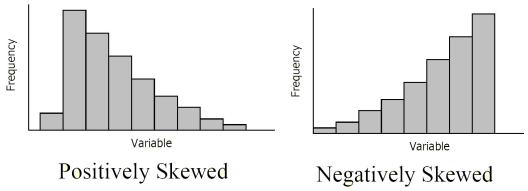
- **Symmetry**: A histogram A histogram is said to be symmetric if, when we draw a **vertical line** down the center of the histogram, the two sides are identical in shape and size:



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

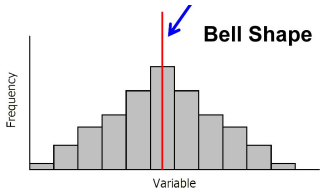


Skewness: A skewed histogram is one with a long tail extending to either the right or the left.



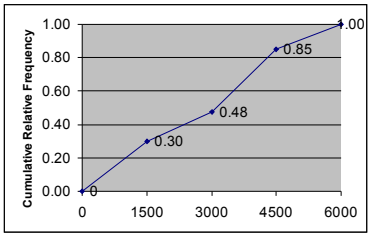
Positively Skewed Negatively Skewed

Bell Shape: A special of symmetric unimodal histogram.



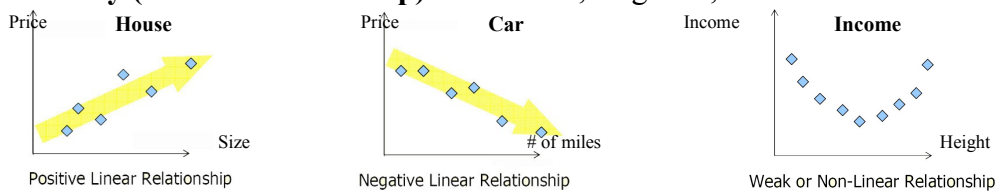
- **Ogive** is a graph of a cumulative relative frequency distribution.

Class Limits	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1500	60	60/200 = 0.30	60/200 = 0.30
1500-3000	35	35/200 = 0.18	0.30+0.18 = 0.48
3000-4500	75	75/200 = 0.38	0.48+0.38 = 0.85
4500-6000	30	30/200 = 0.15	0.85+0.15 = 1.00
Total	200		

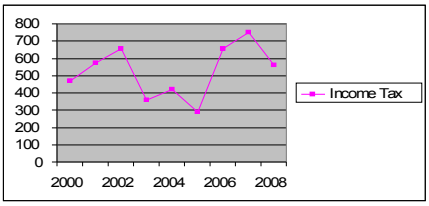


We estimate that about 48% of the students have first income lower than $3000. ….85%..... $4500

- **Two Nominal Variable** → Contingency Table and Bar Chart (two-dimensional)
- **Two Interval Variable** → Scatter diagram to explore the relationship between 2 interval data. The Independent variable (X) → Horizontal axis. The Dependent variable (Y) → Vertical axis.
- **Linearity (Linear Relationship)** → Positive, Negative, or weak or non-linear relationship.



- Cross-Sectional Data → Observations measured at the same point in time.



- **Time-Series Data** → Observations measured at successive points in time → Line chart

Chapter 3: Art and Science of Graphical Presentations

- Graphical Excellence is achieved when
  1. Large data sets are presented concisely and coherently. If small → Table; 1 or 2 → Sentence
  2. The message being presented by the graph is clear to readers. I.e. chart to replace 1000 words
  3. The comparison of two or more variables is aided. I.e. one-variable graph → little information
  4. The substance of the data, not the form of the graph is important.
  5. There is no distortion and deception of the data and findings.

- It's important to be able to critically evaluate the graphically presented information because graphical techniques create a visual impression which is easy to distort.
- Be wary of graphs without a scale on one axis; avoid being influence by caption.
  - Understand the information being presented. Focus on the numerical values the graph represents: absolute values or relative values (e.g. percentages, deltas)?
- Are the horizontal or vertical axes distorted in any way?
  - Note the scales on both axes
  - Graphs with unmarked axes
  - The gaps along axes, the size of bar or chart

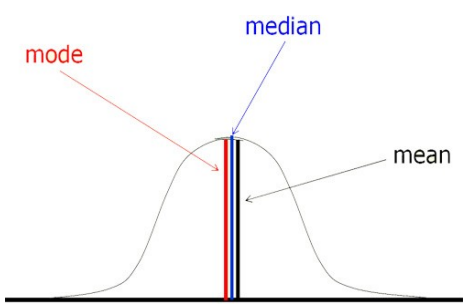
- Writing a Report:
  1. State your objective or purpose of the statistical analysis.
  2. Describe the Experiment – ensure the proper conduct of experiment.
  3. Results - describe using words, tables, and charts – be honest; don't mislead reader
  4. Discussion of limitations of statistical techniques.
  5. (Discuss problems with the analysis – include violations.

- Oral Presentation:
  1. Know your audience: What kind of information they will be expecting? What is their level of statistical knowledge?
  2. Restrict your points to the main study objectives: Don't go into the details of your analysis
  3. Stay within time limits: Respect your audience
  4. Use graphs Use the graphical excellence ideas here to explain complex ideas
  5. Provide handouts.

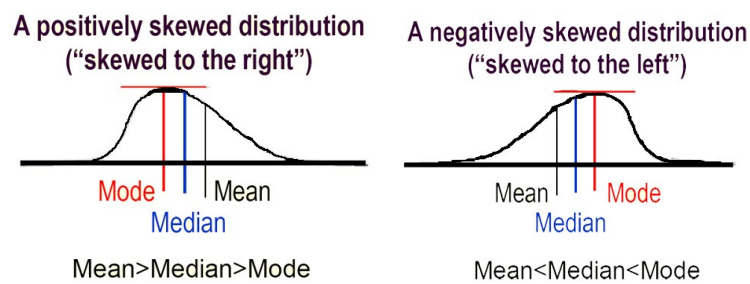
Chapter 4: Numerical Descriptive Techniques

Measures of Central Location				
	Population	Sample	Example	Note
Size	N	n	N = 6: {1 5 7 8 2 7}	
Arithmetic Mean	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{1+5+7+8+2+7}{6} = 5$	- Seriously affected by extreme value – outliers. Billionaire → average income - Valid: Only Interval data
Median	- Sort in order - Middle number or - average of middle two number		1 2 5 7 7 8 Median = (5 + 7) / 2 = **6**	- Half of observation is smaller than 6 and half is greater than 6 - Valid: Ordinal or Interval (extreme)
Mode	- A set of observation occur most frequently. - May have one, two, or more		7 occurs twice Mode = 7	- mainly useful for nominal data - large data set is more relevant - All occurs once: 0 or All
Geometric Mean	$R_g = \sqrt{(1+R_1)(1+R_2)...(1+R_n)} - 1$ → Use when the variable is a growth rate or rate of change. E.g. 2 year investment: 1st year: 100% growth; 2nd year: 50% loses from 1st year $R_g = \sqrt{(1+1)(1-0.5)} - 1 = 1 + 1 = 0$			
MAD	Mean Absolute Deviation : $MAD = \frac{\sum \lvert x - \bar{x} \rvert}{n} = \frac{4+0+2+3+3+2}{6} = 2.33$			

If a distribution is symmetrical,
the mean, median and mod may coincide



If the distribution is non symmetrical, and skewed to the left or the right, the three measure differ.



k	Interval	Chebysheff Theorem $1 - \frac{1}{k^2}$		Empirical Rule (Bell shape)
1	$\bar{x} - 1s, \bar{x} + 1s$	At least 0%	$(1 - 1 / 1^2) = 0$	Approximately **68%**
2	$\bar{x} - 2s, \bar{x} + 2s$	At least 75%	$(1 - 1 / 2^2) = 0.75$	Approximately **95%**
3	$\bar{x} - 3s, \bar{x} + 3s$	At least 88.9%	$(1 - 1 / 3^2) = 0.889$	Approximately **99.7%**

E.g. Histogram is Bell shape, Mean = 10% , Standard Deviation = 8%, According to Empirical Rules:
- Approximately **68%** is between 2 and 18 …. App. **99.7%** is between -14 and 34. ← (10 – 3x8), (10+3x8)

E.g. Positive Skewed Histogram of salaries, Mean = $280, Standard Deviation = $30, according to Chebysheff:
- At least **75%** of the salaries lie between $220 and $340 …. At least **88.9%** is between $190 and $370

Measures of Variability			
Range	Largest observation – Smallest observation	Example: N = 6: {1 5 7 8 2 7 } Range: 8 – 1 = **7**	- Advantage is simplicity - Disadvantage: two completely different dataset may have the same range.
Variance	Population: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$ Sample: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ Shortcut of Sample Variance: $s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$	n / X / x^2 1 / 1 / 1 2 / 5 / 25 3 / 7 / 49 4 / 8 / 64 5 / 2 / 4 6 / 7 / 49 30 / 192	$s^2 = \frac{1}{6-1}\left[192 - \frac{(30)^2}{6}\right]$ $= \frac{1}{5}\left[192 - \frac{(30)^2}{6}\right]$ $= \frac{1}{5}[192 - 150] = 8.4$
Standard Deviation	Population: $\sigma = \sqrt{\sigma^2}$ Sample: $s = \sqrt{s^2}$	==> $s = \sqrt{8.4} = 2.9$ When we compare two standard deviations with similar means, smaller number means more consistent, stable, etc. while larger standard deviation show higher risks, less consistency.	
Coefficient of Variation	Population: $CV = \frac{\sigma}{\mu}$ Sample: $cv = \frac{s}{\bar{x}}$	Mean = 5, Standard deviation = 2.9 ==> $cv = \frac{2.9}{5} = .58$ This coefficient provides a proportionate measure of variation. Smaller → consistency; larger→inconsistency	
	• If data are symmetric, with no serious outliers, use range and standard deviation. • If comparing variation across two data sets, use coefficient of variation.		

- **Measures of Relative Standing** are designed to provide information about the position of particular values relative to the entire data set. The **Percentile (Pth)** is the value for which P percent are less than that value and (100-P)% is greater than that value. E.g. you score in the 60th percentile means 60% of other score were below you, and 40% of scores were above you.
    - First (lower) decile = 10th percentile
    - First (lower) quartile, Q1, = 25th percentile
    - Second (middle) quartile,Q2,= 50th percentile = **Median**
    - Third quartile, Q3, = 75th percentile
    - Ninth (upper) decile = 90th percentile

Interquartile range = $Q_3 - Q_1$ (measure the spread of the middle 50% of observation)

Location of a Percentile: $L_p = (n+1)\frac{P}{100}$, where L_p is the location of the Pth percentile

E.g. {0 1 2 3 4 5 6 **7 8** 9}→ L_{75} = (10+1) (75/100) = **8.25** → The third quartiles or 75th percentile = 7 + (8-7) **0.25** = **7.25**

Measures of Linear Relationship provide information as to the strength and direction of a linear relationship between two variables (if one exists).

Covariance

Popular:

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Shortcut for Sample Covariance:

$$s_{xy} = \frac{1}{n-1}\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right]$$

E.g. x is hours of study and y is grade

n	x	y	x^2	y^2	xy
1	0	40	0	1600	0
2	5	55	25	3025	275
3	8	65	64	4225	520
4	10	70	100	4900	700
5	15	85	225	7225	1275
6	20	100	400	10000	2000
Σ	58	415	814	30975	4770

$$s_{xy} = \frac{1}{6-1}\left[4770 - \frac{58 \times 415}{6}\right] = =151.6667$$

$$s_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] = \frac{1}{6-1}\left[814 - \frac{(58)^2}{6}\right] = 50.6667$$

$$s_y^2 = \frac{1}{n-1}\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right] = \frac{1}{6-1}\left[30975 - \frac{(415)^2}{6}\right] = 454.1667$$

$$s_x = \sqrt{s_x^2} = \sqrt{50.6667} = 7.1181$$

$$s_y = \sqrt{s_y^2} = \sqrt{454.1667} = 21.3112$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{151.6667}{7.1181 \times 21.3112} = 0.9998$$

Coefficent of Correlation

Popular: $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ $\quad -1 \le \rho \le +1$

Sample: $r = \dfrac{s_{xy}}{s_x s_y}$ $\quad -1 \le r \le +1$

r = + 1 Strong positive linear relationship
r = 0 No linear relationship
r = – 1 Strong negative linear relationship
r = 0.56 Moderately strong positive
r = – 0.1 Weak negative …..

Chapter 6: Probability

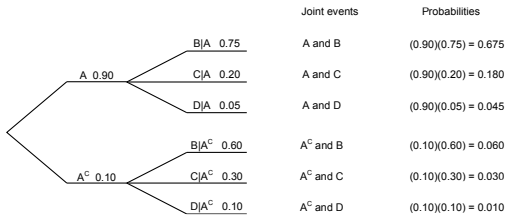
- A **Random Experiment** is an action or process that leads to one of several possible outcomes. The listed outcomes must be **exhaustive** (All possible outcomes included), and the outcomes must be **mutually exclusive** (No two outcomes can occur at the same time.)
- **Sample Space** of a random experiment is a list of exhaustive and mutually exclusive outcomes. The probability of any outcome is between 0 and 1, and the sum of the probabilities of all the outcomes equals 1. $S = \{O_1, O_2, \ldots, O_k\}$ $0 \leq P(O_i) \leq 1$ $\sum P(O_i) = 1$
- There are three ways to assign a probability to an outcome:
  - **Classical Approach:** counting approach used effectively in games of chance.
  - **Relative Frequency:** assigning probabilities based on experimentation or historical data.
  - **Subjective Approach:** assigning probabilities based on the assignor's judgment.
- An **Event** is a collection or set of one or more simple events in a sample space. A **Simple Event** is individual outcome of a sample space.
- The **Probability of an Event** is the sum of the probabilities of the simple events that constitute the event.
- The **Complement of Event** A is defined to be the event consisting of all sample points that are "not in A" or A^c. $P(A) + P(A^c) = 1$
- The **Intersection of Event** (A and B) A and B is the set of all sample points that are in both A and B. The Probability of the intersection is called the **Joint Probability**: P (A and B).
- When two events are **Mutually Exclusive** (that is the two events cannot occur together), their joint probability is 0.
- **Marginal Probabilities** are computed by adding across rows and down columns.

Joint Probabilities

	B_1	B_2	$P(A_i)$
A_2	.20	.35	.55
A_2	.15	.30	.45
$P(B_i)$	.35	.65	**1.00**

Marginal Probabilities

- The **Conditional Probability** is used to determine how two events are related. We can determine the probability of event A given the occurrence of event B: $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$
- If two events A and B are **Independent** – P(A) is not affected by P(B), then P(A|B) = P(A)
- The **Union** of events A and B is the event that occurs when either A or B or both occur.
- **Complement Rule: $P(A^c) = 1 - P(A)$**
- **Conditional Probability:** $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$
- **Multiplication Rule: P(A and B) = P(A|B) • P(B) or P(A and B) = P(B|A) • P(A)**
- **Multiplication Rule for Independent Events: P(A and B) = P(A) • P(B)**
- **Addition Rule: P(A or B) = P(A) + P(B) – P(A and B)**
- **Addition Rule for Mutually Exclusive Events: P (A or B) = P(A) + P(B)**
- **Probability Trees:**



Chapter 7: Random Variables and Discrete Probability Distributions

- A **Random Variable** is a function or rule that assigns a number to each outcome of an experiment. There are two types of random variables: **Discrete Random Variable** (Integer or countable, 1,2,3…) and **Continuous Random Variable** (Real number or uncountable, time, distance, etc.)
- A **Probability Distribution** is a table, formula, or graph that describes the values of a random variable and the probability associated with these values. → P(x) or P(X=x)
- Requirement for a **Distribution of a Discrete Random Variable:** $0 \leq P(x) \leq 1$ and $\sum P(x) = 1$
- **Population Mean** is the weighted average of all of its values. The weights are the probabilities. This parameter is also call **Expected Value** of X: $E(X) = \mu = \sum xP(x)$
- **Population Variance:** $V(X) = \sigma^2 = \sum (x-\mu)^2 P(x) = \sum x^2 P(x) - \mu^2$ (Shortcut)
- **Population Standard Deviation:** $\sigma = \sqrt{\sigma^2}$

E.g

# of Course	# of Student	X	P(X)
0	5	0	0.05
1	10	1	0.10
2	20	2	0.20
3	40	3	0.40
4	15	4	0.15
5	10	5	0.10

X	P(x)	X • P(x)	X^2	X^2 • P(x)
0	0.05	0.00	0	0.00
1	0.10	0.10	1	0.10
2	0.20	0.40	4	0.80
3	0.40	1.20	9	3.60
4	0.15	0.60	16	2.40
5	0.10	0.50	25	2.50
		2.80		9.40

$E(X) = \mu = \sum xP(x) = \mathbf{2.8}$

$V(X) = \sigma^2 = \sum x^2 P(x) - \mu^2 = 9.4 - (2.8)^2 = \mathbf{1.56}$

$\sigma = \sqrt{\sigma^2} = \sqrt{1.56} = 1.249$

- Laws Expected Value:
  - $E(c) = c$
  - $E(X + c) = E(X) + c$
  - $E(cX) = c\,E(X)$
- Laws Variance:
  - $V(c) = 0$
  - $V(X + c) = V(X)$
  - $V(cX) = c^2\,V(X)$

E.g. Monthly sales have a mean of is \$25,000 and standard deviation of \$4,000, profits 30% of sales subtract \$6,000.

- Mean of Monthly Profits:
  E(Profit) = E[.30 (Sales) – 6000] = E[.30 (25000)] – 6000 = .30 • E(25000) – 6000 = \$1,500
- Standard Deviation of Monthly Profits:
  V (Profit) = V[.30 (Sales) – 6000] = 0.30^2 • V(Sales) = 0.09 X 4000^2 = 1440000
  ➔ Standard Deviation of Monthly Profits: $\sigma = \sqrt{\sigma^2} = \sqrt{1440000} = 1200$

- **Binomial Distribution** is the probability distribution that results from doing a binomial experiment. **Binomial Experiments** have the following properties:
    1. Fixed number trials, represented by n.
    2. Each trial has two possible outcomes, a *success* and a *failure*.
    3. P(success) = p P(failure) = 1 – p
    4. The trials are independent, which means that the outcome of one trial does not affect the outcomes of any other trials.
- **Binomial Probability Distribution:** $P(x) = \frac{n!}{x!(n-x)} p^x (1-p)^{n-x}$
- The **Binomial Table** gives cumulative probabilities for $P(X \leq k)$

→ $P(X = k) = P(X \leq k) - P(X \leq [k - 1])$

E.g. $P(X = 5) = P(X \leq 5) - P(X \leq 4)$

$P(X < 5) = P(X \leq 4)$

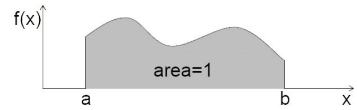
$P(3 \leq X \leq 6) = P(X \leq 6) - P(X \leq 2)$

$P(X > 5) = P(X \geq 6) = 1 - P(X \leq 5)$

- **Binomial Random Variable:**
    - **Mean or Expected Value:** $\mu = np$
    - **Variance:** $\sigma^2 = np(1-p)$
    - **Standard Deviation:** $\sigma = \sqrt{np(1-p)}$

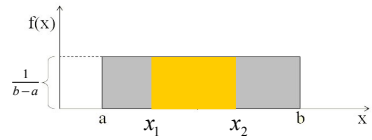
Chapter 8: Continuous Probability Distribution

- A **Continuous Random Variable** is one that can assume an uncountable number of values. So, we cannot list the possible value because infinity number, and the probability of each individual values is 0, i.e., **Point Probability is Zero,** e.g. P(X = 5) = 0
- A function f(x) is called **Probability Density Function** (over range $a \le x \le b$) if it meets the following requirements:
  - $F(x) \ge 0$ for all x between a and b
  - The Total area under the curve between a and b is 1



The **Uniform Probability Distribution** (the Rectangular Probability Distribution) is described by the function: $f(x) = \dfrac{1}{b-a}$ *where* $a \le x \le b$

$P(x_1 \le x \le x_2)$ = width X height = $P(x_1 \le x \le x_2) = (x_2 - x_1) \times \dfrac{1}{b-a}$



- The **Normal Distribution** is the most important of all probability distributions because its crucial role in statistical inference. It look like Bell Shaped, and Symmetrical around the mean.
- The Probability desity function of a **Normal Random Variable** is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < \infty \quad , \quad e = 2.71828 \quad \pi = 3.14159$$

Normal Distribution is fully defined by standard deviation (σ) and Mean (μ).
Unlike the range of the uniform distribution ($a \le x \le b$), Normal Distribution range from minus infinity to plus infinity ($-\infty < x < \infty$)

- A normal distribution whose mean is zero (μ=0) and standard deviation is one (σ=1) is called the **Standard Normal Distribution.** Any normal distribution can be converted to a standard normal distribution with simple algebra which makes calculations much easier.
- Normal Distribution with same standard deviation, **increasing mean <u>shift</u> the curve to the <u>right</u>.**

Normal Distribution with same means, **increasing standard deviation <u>flattens</u> the curve.**

- We can calculate Normal Probabilities by converting normal random variable to a **Standard Normal Random Variable:** $Z = \dfrac{X-\mu}{\sigma}$

E.g. μ = 50, σ = 10

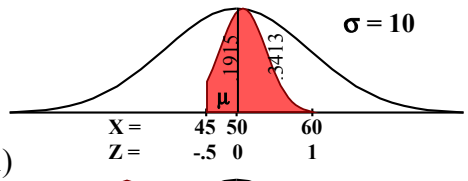
$$P(45 < X < 60) = P(\frac{45-50}{10} < \frac{X-\mu}{\sigma} < \frac{60-40}{10}) = P(-.5 < Z < 1)$$

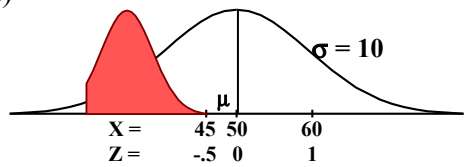
$$P(-.5 < Z < 1) = .1915 + .3413 = .5328$$

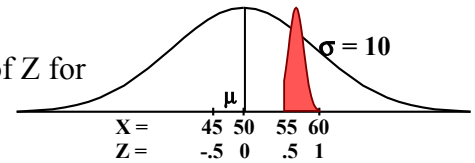
$$P(Z < -.5) = P(Z > .5) = 0.5 - .1915 = .3085$$

$$P(Z > 1) = 0.5 - .3413 = .1587$$

$$P(55 < X < 60) = P(.5 < Z < 1) = .3413 - .1915 = .1498$$





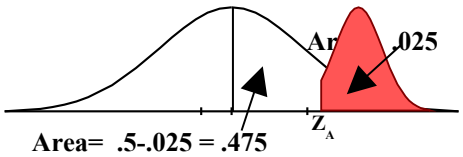


- **Finding Value of Z:** often we're asked to find some value of Z for a given probability, i.e., given an area (A) under the curve, what is the corresponding value of Z (Z_A).

We can find Z_A by "Reverse look up" on table for .475 and we find:

$Z_{.025}$ = 1.96 P(-1.96 < Z < 1.96) = .95

$Z_{.05}$ = 1.645 P(-1.645 < Z < 1.645) = .90



$\mathbf{Z_{.01} = 2.33}$ $P(-2.33 < Z < 2.33) = .98$

Chapter 5: Data Collection and Sampling

Three Methods of Collection Data

- **Direct Observation** – measurements representing a variable of interest are observed and recorded, without controlling any factor that might influence their values. It is inexpensive but difficult to product useful information because result can be affect by other hidden factors. E.g. Aspirin → Heart Attacks (Health conscious)
- **Experiments** – measurements representing a variable of interest are observed and recorded, while controlling factors that might influence their values, usually more expensive and direct observation. E.g. Regular Aspirin and Not Regular Aspirin → Heart Attacks.
- **Surveys** – solicit or ask for information from people. E.g. Gallup polls and Harris Survey, pre-election polls. Majority of surveys are conducted for private use. E.g. Marketing surveys.
- Key Survey parameter is the Response Rate (i.e. the proportion of all people selected who complete the survey)
- Surveys may be administered in a variety of ways
    - Personal Interview – High expected response rate, lower misunderstanding of question, but expensive and easy to be bias if interviewer not train properly.
    - Telephone Interview – Less expensive, But lower response rate
    - Self Administered Questionnaire – By mail, inexpensive, But lower response rate and high number of incorrect response (misunderstanding of question)
- Key Questionnaire Design Principles
    - Keep the questionnaire as short as possible.
    - Ask short, simple, and clearly worded questions.
    - Demographic Questions: 2 schools of thought
        - Keller recommends: Start with demographic questions to help respondents get started comfortably
        - Other experts: recommend these questions be asked at the end).
    - Use dichotomous (yes|no) & multiple choice questions
    - Use open-ended questions cautiously
    - Avoid using leading-questions
    - Pretest a questionnaire on a small number of people.
    - Think about the way you intend to use the collected data when preparing the questionnaire.

- Statistical inference permits us to draw conclusions about a population based on a sample.
- **Sampling** – selecting a sub-set of a whole population is often done for reasons of:
    - cost (it's less expensive to sample 1,000 TV viewers than 100 million TV viewers)
    - practicality (e.g. performing a crash test on every automobile produced is impractical).
- In any case, the **sampled population** and the **target population** should be similar to one another.
- A sampling plan is just a method or procedure for specifying how a sample will be taken from a population. Our focus will be on these three methods:
    - **Simple Random Sampling (SRS)** is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.

        - Use of a random numbers table or a software package
          Example: drawing three names from a hat containing all the names of the students in the class. Any group of three names is as equally likely as picking any other group of three names.
    - **Stratified Random Sampling** is a stratified random sample is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.
        - Example: Male students = 2000, female = 4000,
          → Sample = 300 ← male =100 (33%) + female = 200 (66%)
    - **Cluster Sampling** is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).
        - This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.
        - Two stage process:
            - Researcher draws a random sample of clusters
            - Researcher draws a random sample of elements within each selected cluster
        - Can involve multiple stages, with clusters within clusters
        - Cluster sampling may increase sampling error due to similarities among cluster members.

- **The larger the sample size is, the more accurate we can expect the sample estimates to be.**

- Two major types of error can arise when a sample of observations is taken from a population:
- **Sampling error** refers to:
    - differences between the sample and the population that exist only because of the observations that happened to be selected for the sample
    - the differences in results for different samples (of the same size) is due to sampling error
    - **Increasing the sample size will reduce this type of error.**
    - Example:
        - Two samples of size 10 of 1,000 households
        - If we happened to get the highest income level data points in our first sample and all the lowest income levels in the second, this delta is due to sampling error.

- **Non-sampling error** are more serious
    - due to mistakes made in the acquisition of data
    - due to the sample observations being selected improperly.
    - **Increasing the sample size will not reduce this type of error.**
- Three types of nonsampling errors:
    - **Errors in data acquisition** arises from the recording of incorrect responses due to:
        - incorrect measurements being taken because of faulty equipment
        - mistakes made during transcription from primary sources
        - inaccurate recording of data due to misinterpretation of terms
        - inaccurate responses to questions concerning sensitive issues
    - **Nonresponse errors** refers to error (or bias) introduced when

        - responses are not obtained from some members of the sample, i.e. the sample observations that are collected may not be representative of the target population
        - Response Rate (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey and sources of non-response error.
    - **Selection bias** occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample
    - Example: 1936 Literary Digest Poll predicted a Republican candidate, Alfred Landon would defeat Democrat incumbent, Franklin D. Roosevelt, by a 3 to 2 margin. Outcome: Roosevelt defeated Landon by a support of 62% of electorate.
    - 2 mistakes made in sampling procedure
        - Ten million sample ballots sent out to prospective voters. Most of the names were taken from the Digest's subscription list and from telephone directories (subscribers to magazine & people with telephones tended to be wealthier than average, and such people then, as today, tended to vote Republican)
        - Only 2.3 million ballots were returned (resulting in a self-selected sample).

Chapter 9: Sampling Distribution

- A sampling distribution is created by, as the name suggests, sampling.
- The method we will employ on the rules of probability and the laws of expected value and variance to derive the sampling distribution.
- **Sampling Distribution of the Sample Mean**
- $E(\overline{X}) = \mu_{\bar{x}} = \mu = \sum xP(x)$
- $V(\overline{X}) = \sigma_{\bar{x}}^2 = \sigma^2 / n \qquad \sigma_{\bar{x}} = \sigma / \sqrt{n}$ where $\sigma^2 = \sum x^2 P(x) - \mu^2$
- If X is normal, $\overline{X}$ is normal. If X is nonnormal, $\overline{X}$ is approximately normal for sufficiently large sample sizes.
    - Note: the definition of "sufficiently large" depends on the extent of nonnormality of X (e.g. heavily skewed; multimodal)
    - If the population is normal, then $\overline{X}$ is normally distributed for all values of n
    - If the population is non-normal, then $\overline{X}$ is approximately normal only for larger values of n
    - In many practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of $\overline{X}$.
- The standard deviation of the sampling distribution is called the standard error:
- Remember that μ and σ^2 are the parameters of the population of X
- To create the sampling distribution of X, we repeatedly drew samples of size 2 from the population and calculated for each sample Thus, we treat $\overline{X}$ as a brand new random variable, with its own distribution, mean, and variance
- Central Limit Theorem: The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of $\overline{X}$ will resemble a normal distribution.

- Statisticians have shown that the mean of the sampling distribution is always equal to the mean of the population and that the standard error is equal to $\sigma / \sqrt{n}$ for infinitely large populations.

However, if the population is finite the standard error is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

where N is the population size and $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor. If the population size is large relative to the sample size the finite population correction factor is close to 1 and can be ignored.

- As a rule of thumb we will treat any population that is at least 20 times larger than the sample size as large. In practice most applications involve populations that qualify as large. As a consequence the finite population correction factor is usually omitted.
- The sampling distribution can be used to make inferences about population parameters. In order to do so, the sample mean can be standardized to the standard normal distribution using the following formulation: $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

- Another way to state the Probability: $P(\mu - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$

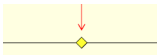
Sample Distribution of a Proportion

- The estimator of a population proportion of successes is the sample proportion. That is,
- we count the number of successes in a sample and compute: $\hat{P} = \frac{X}{n}$
- Using the laws of expected value and variance, we can determine the mean, variance, and
- standard deviation of . (The standard deviation of is called the standard error of the proportion.)
  - $E(\hat{P}) = \mu_{\hat{p}} = p$
  - $V(\hat{P}) = \sigma^2_{\hat{P}} = \frac{p(1-p)}{n}$
  - $\sigma_{\hat{P}} = \sqrt{p(1-p)/n}$
- Sample proportions can be standardized to a standard normal distribution using this formulation:
  - $Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$

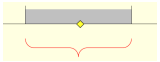
- The final sampling distribution introduced is that of the difference between two sample means.
- This requires: independent random samples be drawn from each of two normal populations
- If this condition is met, then the sampling distribution of the difference between the two sample means, i.e. $\overline{X}_1 - \overline{X}_2$ will be normally distributed.
- Note: If the two populations are not both normally distributed, but the sample sizes are "large" (>30), the distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal.
- The expected value and variance of the sampling distribution of are given by:
- Mean: $E(\overline{X}_1 - \overline{X}_1) = \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$
- Variance: $V(\overline{X}_1 - \overline{X}_1) = \sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ (also called the standard error if the difference between two means)
- Standard deviation: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- We can compute Z (standard normal random variable) in this way: $Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Chapter 10: Introduction to Estimation

- Statistical inference is the process by which we acquire information and draw conclusions about populations from samples. There are two general procedures for making inferences about population: **Estimation** and **Hypothesis**.
- The objective of **Estimation** is to determine the approximate value of a population parameter on the basis of a sample statistic. E.g., sample mean ($\bar{x}$) is employed to estimate the population mean (μ).
- A **Point Estimator** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



Drawbacks: (1) The point probabilities in continuous distributions were virtually zero.
(2) We'd expect that the point estimator gets closer to the parameter value with an increased sample size, but (3) point estimators don't reflect the effects of larger sample sizes.
- An **Interval Estimator** draws inferences about a population by estimating the value of an unknown parameter using an interval.



For example, suppose we want to estimate the mean summer income of a class of business students. For n=25 students, $\bar{X}$ (point estimate) is calculated to be 400 $/week. Interval estimate is: the mean income is between $380 and $420 /week.
- Desirable qualities in Estimators include:
- **Unbiasedness** – An unbiased estimator of a population parameter is an estimator whose expected value is equal to that parameter. E.g. the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ , since: E($\bar{X}$) = μ
- **Consistency** – An unbiased estimator is said to be consistent if the difference between the estimator and the parameter grows smaller as the sample size grows larger. E.g. $\bar{X}$ is a consistent estimator of because: V($\bar{X}$) is σ^2 / n . That is, as n grows larger, the variance of X grows smaller.
- **Relative Efficiency** – If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to be relatively efficient. Both the sample median and sample mean are unbiased estimators of the population mean. However, statisticians have established that the sample median has a greater variance than the sample mean, so we choose $\bar{X}$ since it is relatively efficient when compared to the sample median.
- We can calculate an interval estimator from a sampling distribution by:
  - Drawing a sample of size n from the population
  - Calculating its mean: $\bar{X}$ And, by the central limit theorem, we know that X is normally (or approximately normally) distributed so will have a standard normal (or approximately normal) distribution. $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
- **Confidence Interval Estimator of μ:** **LCL and UCL:** $\bar{X} \pm Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
  - A **larger** confidence level produces a **wider** confidence interval.
  - **Larger** values of variance σ produce **wider** confidence intervals.
  - **Increasing** the sample size **decreases the width** of the confidence interval while the confidence level can remain unchanged. Note: this also increases the cost of obtaining additional data.
- We can control the width of the interval by determining the sample size necessary to produce narrow intervals. Suppose we want to estimate the mean demand "to within 5 units" (W) ; i.e. we

want to the interval estimate to be: $\overline{X} \pm 5$, since: $\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, it follows that

$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5$ So

- Sample Size to Estimate a mean: $n = \left(\frac{Z_{\alpha/2}\sigma}{W} \right)^2$

Chapter 11: Introduction to Hypothesis Testing

- Five critical concepts in hypothesis testing:
  1. There are two hypotheses, the null hypothesis (H_0) and the alternative hypotheses (H_1).
  2. The testing procedure begins with the assumption that the null hypothesis is true.
  3. The goal is to determine whether there is enough evidence to infer that the alternative hypothesis is true.
  4. There are two possible decisions:
     - Reject H_0 – conclude that there is enough evidence to support the alternative hypothesis.
     - Do not Reject H_0 – Conclude that there is not enough evidence to support the alternative hypothesis.
  5. Two possible errors can be made:
     - Type I error: Reject a true null hypothesis. P(Type I error) = α
     - Type II error: Do not reject a false null hypothesis. P(Type II error) = β

	H_0 is true	H_0 is false
Reject H_0	Type I Error P(Type I error) = α	Correct Decision
Do not Reject H_0	Correct Decision	Type II Error P(Type II error) = β

- The **Rejection Region** is a range of values such that if the test statistic falls into that range, we decide to reject the null hypothesis in favor of the alternative hypothesis.
  - Z-test, $\alpha = .05$, for H1 < 100 → Rejection Region: $Z < -Z_{\alpha} < -Z_{.05} < -1.645$
  - Z-test, $\alpha = .01$, for H1 > 100 → Rejection Region: $Z > Z_{\alpha} > Z_{.01} < 2.83$
  - $\alpha = .05$, for H1 ≠ 100 → R. R.: $Z < -Z_{\alpha/2}$ ($-Z_{.025} = -1.96$) or $Z > Z_{\alpha/2}$ ($Z_{.025} = 1.96$)

- The **p-value** of a test is the probability of observing a test statistic at least as extreme as the one computed given that the null hypothesis is true.
- **The smaller the p-value, the more statistical evidence exists to support the alternative hypothesis.**
  - If the p-value is less than 1%, there is overwhelming evidence that supports the alternative hypothesis.
  - If the p-value is between 1% and 5%, there is a strong evidence that supports the alternative hypothesis.
  - If the p-value is between 5% and 10% there is a weak evidence that supports the alternative hypothesis.
  - If the p-value exceeds 10%, there is no evidence that supports the alternative hypothesis.
- **If the p-value is less than α, we judge the p-value to be small enough to reject the null hypothesis. If the p-value is greater than α, we do not reject the null hypothesis.**

- **Conclusions of a Test of Hypothesis:**
  - If we reject the null hypothesis, we conclude that there is enough evidence to infer that the alternative hypothesis is true.

- If we do not reject the null hypothesis, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true.
- Remember: The alternative hypothesis is the more important one. It represents what we are investigating.

- **The Probability of a Type II Error:** It is important that that we understand the relationship between Type I and Type II errors; that is, how the probability of a Type II error is calculated and its interpretation.
- A Type II error occurs when a false null hypothesis is not rejected.

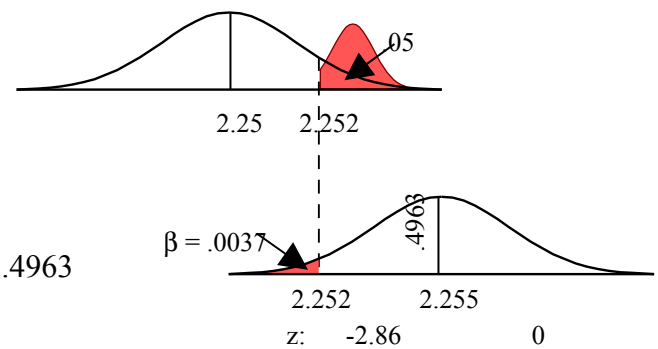
E.g. $\alpha = .05$ $n = 20$ $\sigma = .005$ $H_0 : \mu = 2.25$ $H_1 : \mu > 2.25$

Rejection region: $\dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} > Z_\alpha$ $\bar{x} > Z_\alpha(\sigma / \sqrt{n})$

$\dfrac{\bar{x} - 2.25}{.005 / \sqrt{20}} > Z_{.05} > 1.645$ → $\bar{x} > 2.252$

$\beta = P(\bar{x} < 2.252 \text{ given } \mu = 2.255)$

$\beta = P\left(\dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \dfrac{2.252 - 2.255}{.005 / \sqrt{20}}\right) = P(z < -2.68) = .5 - .4963$



$\beta = .0037$

- **Decreasing the significance level α, increases the value β of and vice versa.**
- Consider this diagram again. Shifting the critical value line to the right (to decrease) will mean a larger area under the lower curve for … (and vice versa)
- if the probability of a Type II error (β) is judged to be too large, we can reduce it by
  - increasing α, and/or
  - increasing the sample size, n.
- **Increasing the sample size n and/or increasing α, the value of β will decrease.**

- The power of a test is defined as $1 - \beta$. It represents the probability of rejecting the null hypothesis when it is false.
- When more than one test can be performed in a given situation, it is preferable to use the test that is correct more often. If one test has a higher power than a second test, the first test is said to be more powerful and the preferred test.

Chapter 12: Inference about a Population

- When the population standard deviation is unknown and the population is normal, the test statistic for testing hypotheses about μ is: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

- Confidence Interval Estimator of μ when σ is unknown: $\bar{x} \pm t_{\alpha/2} \dfrac{s}{\sqrt{n}}$

Which is Student distributed with **v = n – 1** degree of freedom
Rejection Region:
  - $t < -t_{\alpha,v}$ $\quad v = n - 1$
  - $t > t_{\alpha,v}$ $\quad v = n - 1$
  - $t < -t_{\alpha/2,v}$ or $t > t_{\alpha/2,v}$ $\quad v = n - 1$

- The Student t distribution is robust, which means that if the population is nonnormal, the results of the t-test and confidence interval estimate are still valid provided that the population is "not extremely nonnormal".
- To check this requirement, draw a histogram of the data and see how "bell shaped" the resulting figure is: If a histogram is extremely skewed and could be considered "extremely nonnormal", the t-statistics would be not be valid in this case.

Estimating the Totals of Finite Populations

- Large populations are defined as "**populations that are at least 20 times the sample size**"
- We can use the confidence interval estimator of a mean to produce a confidence interval estimator of the population total: $N\left(\bar{x} \pm t_{\alpha/2} \dfrac{s}{\sqrt{n}}\right)$ Where N is the size of the finite population.

Inference about a Population Variance

- If we are interested in drawing inferences about a population's variability, the parameter we need to investigate is the population variance: σ^2
- The sample variance (s^2) is an unbiased, consistent and efficient point estimator for σ^2.

Moreover, the statistic, $X^2 = \dfrac{(n-1)s^2}{\sigma^2}$, has a chi-squared distribution, with n–1 degrees of freedom.

- Confidence Interval Estimator of σ^2:
  - Lower Confidence Limit (LCL): $\dfrac{(n-1)s^2}{X^2_{\alpha/2,\,n-1}}$
  - Upper Confidence Limit (UCL): $\dfrac{(n-1)s^2}{X^2_{1-\alpha/2,\,n-1}}$
- Rejection Region
  - $X^2 < X^2_{1-\alpha,\,n-1}$
  - $X^2 > X^2_{\alpha,\,n-1}$

- $X^2 < X^2_{1-\alpha/2,\, n-1}$ or $X^2 > X^2_{\alpha/2,\, n-1}$

- Decreasing the sample size decrease the test statistic and increase the p-value of the test.
- Increase the sample size narrows the intervals.

Inference about a Population Proportion

- When data are nominal, we count the number of occurrences of each value and calculate
- proportions.
- Thus, the parameter of interest in describing a population of nominal data is the population
- proportion p.
- This parameter was based on the binomial experiment.
- Recall the use of this statistic: $\hat{P} = \dfrac{X}{n}$ where p-hat ($\hat{P}$) is the sample proportion and x successes in a sample size of n items.
- When np and n(1–p) are both greater than 5, the sampling distribution of $\hat{P}$ is approximately normal with mean: μ = p. and Standard deviation: $\sigma = \sqrt{p(1-p)/n}$
- **Test Statistic for p:** $z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$
- **The confidence interval estimator for p is given by:** $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
- **Selecting the Sample Size:** $n = \left(\dfrac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{W} \right)^2$
- Estimating the Total Number of Successes in a Large Finite Population: $N\left(\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)$

Chapter 13: Inference About Comparing Two Populations

- In order to test and estimate the difference between two population means, we draw random samples from each of two populations.
- **Sampling Distribution of $\bar{x}_1 - \bar{x}_2$:**
  1. $\bar{x}_1 - \bar{x}_2$ is normally distributed if the original populations are normal or approximately normal if the populations are non-normal and the sample sizes are large (n1, n2 > 30)
  2. The expected value of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$
  3. The variance of $\bar{x}_1 - \bar{x}_2$ is $\sigma_1^2 / n_1 + \sigma_2^2 / n_2$ and the standard error is: $\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$
- When **Population Variances are known** and the sampling distribution of is normal or approximately normal we can build the test statistic (z-test) and the Interval estimator:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- However, the z-statistics is rarely used because the population variances σ_1^2 and σ_2^2 are always unknown. Instead we use a t-statistic. We consider two cases for the unknown population variances: when we believe they are equal and conversely when they are not equal.
- Since the population variances are unknown, we can not know for certain whether they are equal, but we can conduct F-test σ_1^2 / σ_2^2 of to determine whether the two population variances differ.
- Test Statistic and Interval estimator for $\mu_1 - \mu_2$ when the two population variances are **equal** $\sigma_1^2 = \sigma_2^2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_1}\right)}} \quad \text{where } S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_1}\right)}$$

- Rejection Region: $t > t_{\alpha,v}$, $t < -t_{\alpha,v}$, ($t > t_{\alpha/2,v}$ or $t < -t_{\alpha/2,v}$)

$v = n_1 + n_2 - 2$

- Test Statistic and Interval estimator for $\mu_1 - \mu_2$ when the two population variances are **not equal** $\sigma_1^2 \neq \sigma_2^2$**:**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \qquad (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{d.f.: } v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Rejection Region: $t > t_{\alpha,v}$, $t < -t_{\alpha,v}$, ($t > t_{\alpha/2,v}$ or $t < -t_{\alpha/2,v}$)

- The number of degrees of freedom associated with the equal-variances test statistic and confidence interval estimator is always greater than or equal to the number of degrees of freedom associated with the unequal-variances test statistic and confidence interval estimator for given sample sizes n_1 and n_2:

$$n_1 + n_2 - 2 \geq \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- Larger numbers of degrees of freedom have the same effect as having larger sample sizes, and larger sample sizes yield more information by producing more powerful tests (lower type II error probabilities) and narrower confidence interval estimation. Therefore, whenever there is insufficient evidence that the variances are unequal, it is preferable to perform the equal variances t-test.
- **Hypothesis Testing and Estimating a Ratio of two Variances**
  - $H_0 : \mu_D = 0 \qquad where \quad \mu_D = \mu_1 - \mu_2$
  - $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \quad or \quad H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \quad or \quad H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} < 1$
  - Test Statistic: $F = \dfrac{S_1^2}{S_2^2}$
  - Rejection Region: $F > F_{\alpha, n_1-1, n_2-1}$, $\quad F < \dfrac{1}{F_{\alpha, n_2-1, n_1-1}}$, $\quad F > F_{\alpha/2, n_1-1, n_2-1}$ or $F < \dfrac{1}{F_{\alpha/2, n_2-1, n_1-1}}$
  - Confidence Interval Estimator: $LCL = \left(\dfrac{s_1^2}{s_2^2}\right)\dfrac{1}{F_{\alpha/2, n_1-1, n_2-1}} \quad UCL = \left(\dfrac{s_1^2}{s_2^2}\right)F_{\alpha/2, n_2-1, n_1-1}$

- **Matched Pairs Experiment:** an observation in one sample is matched with an observation in a second sample, this is called a matched pairs experiment.
- **Hypothesis Test:**
  - $H_0 : \mu_D = 0 \qquad where \quad \mu_D = \mu_1 - \mu_2$
    $H_1 : \mu_D \neq 0 \quad or \quad H_1 : \mu_D > 0 \quad or \quad H_1 : \mu_D < 0$
  - Rejection Region: $t > t_{\alpha, n_D-1}$, $\quad t < -t_{\alpha, n_D-1}$, $\quad (t > t_{\alpha/2, n_D-1}$ or $t < -t_{\alpha/2, n_D-1})$
  - Test Statistic: $t = \dfrac{\bar{x}_D - \mu_D}{s_D / \sqrt{n_D}}$
  - Confidence Interval Estimator: $\bar{x}_D \pm t_{\alpha/2} \dfrac{s_D}{\sqrt{n_D}}$

- **Inference about the difference between two population proportions:** As mentioned previously, with nominal data, calculate proportions of occurrences of each type of outcome. Thus, the parameter to be tested and estimated in this section is the difference between two population proportions: $p_1 - p_2$.
- To draw inferences about the parameter $p_1 - p_2$, we take samples of population, calculate the sample proportions and look at their difference.
  - $\hat{p}_1 = \dfrac{x_1}{n_1} \qquad and \qquad \hat{p}_2 = \dfrac{x_2}{n_2}$
- **Sampling Distribution of $\hat{p}_1 - \hat{p}_2$**
  - The statistic $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed if the sample sizes are large enough so that $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2, and\ n_2(1-\hat{p}_2)$ are great than or equal to 5.
  - The Mean of $\hat{p}_1 - \hat{p}_2$ is $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
  - The Variance of $\hat{p}_1 - \hat{p}_2$ is $V(\hat{p}_1 - \hat{p}_2) = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$

The standard error is $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Hypothesis Test for two population proportions

- **Case 1: Test Statistic for** $p_1 - p_2$
  - $H_0 : (p_1 - p_2) = 0$

    $H_1 : (p_1 - p_2) \neq 0 \quad or \quad H_1 : (p_1 - p_2) > 0 \quad or \quad H_1 : (p_1 - p_2) < 0$
  - Rejection Region: $Z > Z_\alpha \quad , \quad Z < -Z_\alpha \quad , \quad (Z > Z_{\alpha/2} \quad or \quad Z < -Z_{\alpha/2})$
  - Test Statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$
- **Case 2: Test Statistic for** $p_1 - p_2$
  - $H_0 : (p_1 - p_2) = D$

    $H_1 : (p_1 - p_2) \neq D \quad or \quad H_1 : (p_1 - p_2) > D \quad or \quad H_1 : (p_1 - p_2) < D$
  - Rejection Region: $Z > Z_\alpha \quad , \quad Z < -Z_\alpha \quad , \quad (Z > Z_{\alpha/2} \quad or \quad Z < -Z_{\alpha/2})$
  - Test Statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$
- **Confidence Interval Estimator for Both Case 1 & 2:**
  - $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- This formular is valid when $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2, and\ n_2(1-\hat{p}_2)$ are great than or equal to 5.

Chapter 15: Analysis of Variance

- Analysis of variance is a technique that allows us to compare two or more populations of interval data. Analysis of variance is:
  - an extremely powerful and widely used procedure.
  - a procedure which determines whether differences exist between population means.
  - a procedure which works by analyzing sample variance.
- Independent samples are drawn from k populations. Note: These populations are **referred** to as treatments. It is not a requirement that $n_1 = n_2 = \ldots = n_k$.
- **x** is the **response variable**, and its values are **responses**.
- $\mathbf{x_{ij}}$ refers to the ith observation in the jth sample. E.g. $x_{3,5}$ is the 3rd observation of the 5th sample.
- The **Grand Mean** is the mean of all observations: $\bar{\bar{x}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} / n$ where k is the number of populations $n = n_1 + n_2 + \ldots + n_k$

The unit that we measure is the **experimental unit**. Population classification criterion is called a **factor**. Each population is a **factor level**.

The null Hypothesis and alternative hypothesis:

- $H_0 : \mu_1 = \mu_2 = \mu_3$ $\qquad H_1 :$ at least two means differ

Test Statistic:

- If the alternative hypothesis is true, there would be large differences between some of the sample means. The statistic that measures the proximity of the sample means to each other is called the **between-treatments variation,** denoted SST, stands for **Sum of Squares for Treatments.** $SST = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2$

A larger SST indicates large variation between sample means which supports H_1.

If the sample means are close to each other, all the sample means would be close to the grand mean, and as a result, SST would be small. In fact, SST achieves its smallest value (zero) when all the sample means are equal. A small value of SST supports the null hypothesis.

- To determine whether SST is large enough to indicate the population means differ (support H_1), we need to know the **within-treatments variation**, denoted by **SSE (Sum of Squares for Error)**.

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 \quad or \quad SSE = \sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2 + \ldots + \sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2$$

$$SSE = \sum_{j=1}^{k}(n_j - 1)S_j^2 \quad or \quad SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2$$

- The **Mean Square for Treatment:** $MST = \dfrac{SST}{k-1}$
- The **Mean Square for Error:** $MSE = \dfrac{SSE}{n-k}$
- The **Test Statistic:** $F = \dfrac{MST}{MSE}$ **Rejection Region:** $F > F_{\alpha,k-1,n-k}$
- If SST is large, F will be larger.

- Required conditions: 1) the populations tested are normally distributed, And 2) The variances of all the populations tested are equal.
- We don't use multiple *t* tests (compare two population: $\mu_1 - \mu_2$, $\mu_2 - \mu_3$ and $\mu_1 - \mu_3$) instead *F* test (Analysis of Variance) because multiple tests increase the probability of making Type I errors.
- We still need the *t* test to make inferences about $\mu_1 - \mu_2$, because the analysis of variance allows us to test only a different only, i.e, $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$, and it requires that the population variances are equal. However, if we simply want to determine whether a difference between two means exists, we can use the analysis of variance.
- The advantage of using the analysis of variance is that we can partition the sum of squares, which enables us to measure how much variation is attributable to differences between populations and how much variation is attributable to differences within populations.
- Relationship Between the *F* Statistic and the *t* Statistic: $F = t^2$
- **An**alysis **o**f **Va**riance **(ANOVA)** table:

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square
Treatments	k – 1	$SST = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$	$MST = \frac{SST}{k-1}$
Error	n – k	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$ $SSE = \sum_{j=1}^{k}(n_j - 1)S_j^2$	$MSE = \frac{SSE}{n-k}$
Total	n – 1	SS (Total)	

Test Statistic: $F = \frac{MST}{MSE}$

Rejection Region: $F > F_{\alpha,k-1,n-k}$

Example: Job offer for MBA major

Finance	Marketing	Management
3	1	8
1	5	5
4	3	4
1	4	6
9	13	23

$n_1 = 4, \quad n_2 = 4, \quad n_3 = 4 \qquad s_1^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} = 2.25, \quad s_2^2 = 2.92 \quad s_3^2 = 3.92$

$\bar{x}_1 = 2.25, \quad \bar{x}_2 = 3.25, \quad \bar{x}_3 = 5.75 \qquad \bar{\bar{x}} = \frac{9+13+23}{4+4+4} = 3.75$

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: At least two means differ.

Rejection region: $F > F_{\alpha,k-1,n-k} = F_{.05,2,9} = 4.26$

$SST = \sum n_j(\bar{x}_j - \bar{\bar{x}})^2 = 4(2.25–3.75)^2 + 4(3.25 – 3.75)^2 + 4(5.75 –3.75)^2 = 26.00$

$SSE = \sum (n_j - 1)s_j^2 = (4 –1)(2.25) + (4 – 1)(2.92) + (4 – 1)(2.92) = 24.25$

ANOVA table

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square
Treatments	k – 1 = 3–1 = 2	$SST = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 26.00$	$MST = \frac{SST}{k-1} = \frac{26}{3-1} = 13$
Error	n – k = 12–3=9	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = 24.25$ $SSE = \sum_{j=1}^{k}(n_j - 1)S_j^2 = 24.25$	$MSE = \frac{SSE}{n-k} = \frac{24.25}{12-3} = 2.6944$

Total	n – 1 = 12–1 = 11	SS (Total) = 50.25

The Value of Test Statistic: $F = \frac{MST}{MSE} = \frac{13}{2.69} = 4.82$

Conclusion: Reject H_0, there is enough evidence to conclude that there are differences in the number of job offers between the three MBA majors.

Chapter 16: Chi-Squared Tests

- Two Techniques for Chi-Squared Tests: **Goodness-of-fit test** and **Contingency table**.
- A **goodness-of-fit test** applied to data produced by a multinomial experiment, a generalization of a binomial experiment and is used to describe one population of data.
- The second uses data arranged in a contingency table to determine whether two classifications of a population of nominal data are statistically independent.
- This test can also be interpreted as a comparison of two or more populations.
- In both cases, (goodness-of-fit test and whether two classifications of a population of nominal data are statistically independent), we use the chi-squared (χ^2) distribution.
- The **Multinomial Experiment**: Unlike a binomial experiment which only has two possible outcomes (e.g. heads or tails), a multinomial experiment:
  - Consists of a fixed number *n* of trials.
  - Each trial can have one of k outcomes, called cells.
  - Each probability p_i remains constant for each trial.
  - Our usual notion of probabilities holds, namely: $p_1 + p_2 + \ldots + p_k = 1$, and
  - Each trial is **independent** of the other trials.

Chi-squared Goodness-of-Fit Test

- We test whether there is sufficient evidence to reject a **specified set** of values for p_i.
- The null hypothesis and alternative hypothesis:
  - H_0: $p_1 = a_1, p_2 = a_2, \ldots, p_k = a_k$ (where $a_1, a_2, \ldots, a_k$ are the values of interest)
  - H_1: At least one p_i is not equal to its specified value
- The test builds on comparing actual frequency and the expected frequency of occurrences in all the cells.
- Chi-squared goodness of fit test statistic is $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$
- Rejection Region: $\chi^2 > \chi^2_{\alpha,k-1}$ degree of freedom: v = k – 1
- Required Conditions: In order to use this technique, the sample size must be large enough so that the expected value for each cell is 5 or more (i.e. $n \times p_i \geq 5$). If the expected frequency is less than five, combine it with other cells to satisfy the condition.

Example:

History distribution: A = .3, B = .3, C = .2, D = .2

New Observed A = 38, B = 50, C = 38, D = 24

(n = 150, k =4, α = 5%)

$H_0: p_1 = .3,\ p_2 = .3,\ p_3 = .2,\ p_4 = .2$

H_1: At least one p_i is not equal to its specified value.

Rejection region: $\chi^2 > \chi^2_{\alpha,k-1} = \chi^2_{.05,3} = 7.81473$

Co.	Observed Frequency	Expected Frequency	Delta	Summation Component
	f_i	e_i	$(f_i - e_i)$	$(f_i - e_i)^2/e_i$
A	38	.3 (150) = 45	-7	$(-7)^2/45$= 1.0889
B	50	.3 (150) = 45	5	$5^2/45$= 0.5556
C	38	.2 (150) = 30	8	$8^2/30$= 2.1333
D	24	.2 (150) = 30	-6	$(-6)^2/30$= 1.2000
Total	150	150		4.9778

Value of Test Statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 4.9778$

Conclusion: Do not Reject H_0, There is not enough evidence to infer that the market shares are distributed differently from market shares in the past.

Chi-squared Test of a Contingency Table

- The Chi-squared test of a contingency table is used to determine whether there is enough evidence to infer that two nominal variables are related, and to infer that differences exist among two or more populations of nominal variables.
- In order to use these techniques, we need to classify the data according to two different criteria.
- The null hypothesis and alternative hypothesis:
  - H0: The two variables are independent
  - H1: The two variables are dependent
- Chi-squared contingency table test statistic is $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$
- Expected Frequencies for a Contingency Table: $e_{ij} = \frac{Row\ i\ total \times Column\ j\ total}{Sample\ size}$
- Rejection Region: $\chi^2 > \chi^2_{\alpha,v}$ degree of freedom: v = (r – 1)(c – 1)
- In a contingency table where one or more cells have expected values of less than 5, we need to combine rows or columns to satisfy the rule of five. Note: By doing this, the degrees of freedom must be changed as well.

Example:
Employee group and responses to new pension plan:
Is there enough evidence to infer that the responses differ among the three groups of employees?

Responses	Low managers	Middle managers	Top managers	
For	67	32	11	110
Against	63	18	9	90
	130	50	20	200

H_0 : The two variables (responses and employee group) are independent (no relationship between 2 variables)
H_1 : The two variables are dependent

Cell i	Observed Frequency	Expected Frequency	Delta	Summation Component
	f_i	e_i	$(f_i - e_i)$	$(f_i - e_i)^2/e_i$
1	67	110(130)/200 = 71.50	-4.5	$(-4.5)^2/71.5 = 0.2832$
2	32	110(50)/200 = 27.50	4.5	$4.5^2/27.5 = 0.7364$
3	11	110(20)/200 = 11.00	0	0
4	63	90(130)/200 = 58.50	4.5	$4.5^2/58.5 = 0.3462$
5	18	90(50)/200 = 22.50	-4.5	$(-4.5)^2/22.5 = 0.9000$
6	9	90(20)/200 = 9.00	0	0
Total				2.2658

Rejection region: $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{.05,2} = 5.99147$

Value of Test Statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 2.2658$

Conclusion: Do not Reject H_0, there is not enough evidence to infer that responses differ among the three groups of employees.

Chapter 17: Simple Linear Regression and Correlation

- The problem objective is to analyze the relationship between interval variables; regression analysis is the first tool we will study.
- Regression analysis is used to predict the value of one variable (the dependent variable) on the basis of other variables (the independent variables). Dependent variable: denoted Y. Independent variables: denoted $X_1, X_2, \ldots, X_k$
- If we are interested **only** in determining whether a relationship **exists**, we employ **correlation analysis**.
- Simple Linear Regression examines the relationship between two variables. Mathematical equations describing these relationships are also called **models**, and they fall into two types: deterministic or probabilistic.
- **Probabilistic Model**: a method used to capture the **randomness** that is part of a real-life process.
- To create a probabilistic model, we start with a deterministic model that **approximates the relationship** we want to model and add a **random term** that measures the **error** of the deterministic component.
- E.g. do all houses of the same size (measured in square feet) sell for exactly the same price?

- **Deterministic Model**: an equation or set of equations that allow us to fully determine the value of the dependent variable from the values of the independent variables.
- E.g. The cost of building a new house is about \$75 per square foot and most lots sell for about \$25,000. Hence the approximate selling price (y) would be: y = \$25,000 + (\$75/ft^2)(x) (where x is the size of the house in square feet). In this model, the price of the house is completely determined by the size.
- In real life however, the house cost will vary even among the same size of house. (e.g. décor options, cabinet upgrades, lot location…).
- We can represent the price of a house as a function of its size in this Probabilistic Model: y = 25,000 + 75x + ε [where ε (Greek letter epsilon) is the random term (a.k.a. error variable]. It is the difference between the actual selling price and the estimated price based on the size of the house. Its value will vary from house sale to house sale, even if the square footage (i.e. x) remains the same.

- A straight line model with one independent variable is called a **first order linear model** or a **simple linear regression model**.

It is written as $y = \beta_0 + \beta_1 x + \varepsilon$ where y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line (defined as rise/run)
ε = Error variable

- **Least Squares Line Coefficients:** $\hat{y} = b_0 + b_1 x$ where $b_1 = \dfrac{S_{xy}}{S_x^2}$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$S_{xy} = \frac{1}{n-1}\left[\sum xy - \frac{\sum x \sum y}{n}\right]$$

$$S_x^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

- The deviations between the actual data points and the line are called **Residuals** ($e_i = y_i - \hat{y}_i$)
- The residuals are observations of the error variable. The objective method of producing a straight line that minimizes the sum of the squared differences between the points and the line (residuals). The minimized sum of squared deviations is called the **sum of squares for error** (SSE).
- For these regression methods to be valid the following four conditions for the error variable (ε) must be met:
  - The probability distribution of ε is normal.
  - The mean of the distribution is 0; that is, E(ε) = 0.
  - The standard deviation of ε is σ_ε, which is a constant regardless of the value of x.
  - The value of ε associated with any particular value of y is independent of ε associated with any other value of y.

Assessing the Model

- The least squares method will always produce a straight line, even if there is no relationship between the variables, or if the relationship is something other than linear. Hence, in addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. They are based on the sum of squares for errors (SSE).
- **Sum of Squares for Error:** $SSE = \sum (y_i - \hat{y}_i)^2 = (n-1)(S_y^2 - \frac{S_{xy}^2}{S_x^2})$
- **Standard Error of Estimate:** $S_\varepsilon = \sqrt{\frac{SSE}{n-2}}$
- If S_ε is large, some of the errors will be large – the model's fit is poor. If S_ε is small, the errors tend to be close to the means (0) – the model fits well. We judge the value of S_ε by comparing it to the sample mean of the dependent variable ($\bar{y}$).

- **Testing the Slope:** If no linear relationship exists between the two variables, we would expect the regression line to be horizontal, that is, to have a slope of zero. We want to see if there is a linear relationship, i.e. to see if the slope (β_1) is something other than zero.
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$ (There is a linear relationship)
    $H_1 : \beta_1 < 0$ (There is a **negative** linear relationship)
    $H_1 : \beta_1 > 0$ (There is a **positive** linear relationship)
  - Test statistic: $t = \frac{b_1 - \beta_1}{S_{b_1}}$ where $S_{b_1} = \frac{S_\varepsilon}{\sqrt{(n-1)S_x^2}}$

    Confidence Interval Estimator of β_1**:** $b_1 \pm t_{\alpha/2} S_{b_1}$ v = n – 2
  - Rejection Region: $t < -t_{\alpha,n-2}$, $t < t_{\alpha,n-2}$, $t < -t_{\alpha/2,n-2}$ *or* $t > t_{\alpha/2,n-2}$
- Note: Even if the alternative hypothesis is true, it may be that a linear relationship exists or that a nonlinear relationship exists but that the relationship can be approximated by a straight line.

- **Coefficient of Determination:** It is also useful to measure the strength of the relationship. This is done by calculating the coefficient of determination (R^2): $R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$

- The coefficient of determination is the square of the coefficient of correlation ($R^2 = r^2$)
- Variation in y = SSE + SSR
- **Sum of Squares Error (SSE)** measures the amount of variation in y that remains unexplained (i.e. due to **error**)
- **Sum of Squares Regression (SSR)** measures the amount of variation in y explained by variation in the **independent variable x**. (The greater the explained variation, the better the model is.)
- **R^2** measures the proportion of the variation in y that is explained by the variation in x (Explained variation / Variation in y).
- For example, R^2 has a value of .6483. This means 64.83% of the variation in the auction selling prices (y) is explained by the variation in the odometer readings (x). The remaining 35.17% is unexplained, i.e. due to error.
- Unlike the value of a test statistic, **the coefficient of determination does not have a critical value that enables us to draw conclusion**. In general the higher the value of R^2, the better the model fits the data. I.e., $R^2 = 1$ (Perfect match between the line and the data points), $R^2 = 0$ (There are no linear relationship between x and y)

- **Coefficient of Correlation:** $r = \sqrt{R^2} = \frac{S_{xy}}{S_x S_y}$
- We can use the **coefficient of correlation** to test for a linear relationship between two variables. If r = –1 (negative association) or r = +1 (positive association) every point falls on the regression line. If r = 0 there is no linear pattern.
- The population coefficient of correlation is denoted ρ (rho). We can conduct the t-test of the coefficient of correlation as an alternate means to determine whether two variables are linearly related.
  - $H_0 : \rho = 0$
  - $H_1 : \rho \neq 0$ (there is a linear relationship)
  - Test Statistic: $t = r\sqrt{\frac{n-2}{1-r^2}}$
  - Rejection Region: $t < -t_{\alpha/2,n-2}$ *or* $t > t_{\alpha/2,n-2}$

- If we're interested in discovering the relationship between two variables, or if we've conducted an experiment where we controlled the values of the independent variable, the *t* test for β_1 should be applied. If we're interested only in determining whether two random variables that are bivariate normally distributed are linearly related, the *t* test of ρ should be applied.

Using the Regression Equation

- **Point Prediction:** We can use regression equation ($\hat{y} = b_0 + b_1 x$)
- **Prediction Interval** is used to predict one particular value of the dependent variable, given a specific value of the independent variable: $\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)S_x^2}}$ where $\hat{y} = b_0 + b_1 x_g$

- **Confidence Interval Estimator of the Expected Value of y:** $\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)S_x^2}}$

- The prediction interval provides a prediction for a value of y. The confidence interval estimator of the expected value of y is an estimator of the population mean for a given x.

Example:

Interest Rate	8.5	7.8	76	7.5	8.0	8.4	8.8	8.9	8.5	8.0
# of borrowers (1000)	115	111	185	201	206	167	155	117	133	150

a) Find the least squares regression line

	x	y	xy	x^2	y^2
1	8.5	115	977.5	72.25	13,225
2	7.8	111	865.8	60.84	12,321
3	7.6	185	1,406.0	57.76	34,225
4	7.5	201	1,507.5	56.25	40,401
5	8.0	206	1,648.0	64.00	42,436
6	8.4	167	1,402.8	70.56	27,889
7	8.8	155	1,364.0	77.44	24,025
8	8.9	117	1,041.3	79.21	13,689
9	8.5	133	1,130.5	72.25	17,689
10	8.0	150	1,200.0	64.00	22,500
Σ	**82.0**	**1540**	**12543.4**	**674.56**	**248400**

$\sum x = 82 \qquad \sum y = 1540 \qquad \sum xy = 12543.4$

$\sum x^2 = 674.56 \qquad \sum y^2 = 248400$

$\bar{x} = \sum x / n = 82/10 = 8.2$

$\bar{y} = \sum y / n = 1540/10 = 154$

$$s_x^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1} = .24$$

$$S_{xy} = \frac{\sum xy - \left(\sum x \sum y\right)/n}{n-1} = \frac{12543.4 - (82)(1540)/10}{9} = -9.4$$

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{-9.4}{.24} = -39.1667$$

$b_0 = \bar{y} - b_1\bar{x} = 154 - (-39.1667)(8.2) = 475.1669$

The sample regression line:

$\hat{y} = b_0 + b_1 x$

$\hat{y} = 475.1669 - 39.1667x$

Interpret coefficient:

The slope $b_1 = -39.1667$ tells us that for each additional 1 percentage point increase in interest rates, the number of borrowers decreases on average by 39166.7.

The y-intercept $b_0 = 475.1669$ tells us when interest rate (x) equals to 0, the number of borrowers ($\hat{y}$) equals to 475.166, but this intercept has no meaning because interest rate did not include any number with zero in our sample.

b) Calculate the standard error of estimate (S_ε)

$$s_y^2 = \frac{\sum y^2 - \left(\sum y\right)^2 / n}{n-1} = \frac{248400 - (1540)^2 / 10}{9} = 1248.8889$$

$$SSE = (n-1)\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) = (10-1)\left(1248.8889 - \frac{(-9.4)^2}{.24}\right) = 7926.5001$$

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{7926.5001}{10-2}} = 31.4772$$

It appears that the standard error of estimate (S_ε) is small because $S_\varepsilon = 31.4772$ is smaller than $\bar{y} = 154$, so the model seems to fit well because smaller S_ε means most points variables fall on the regression line.

c) Is there a linear relationship between two variables?

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Rejection Region:

$t < -t_{\alpha/2,n-2} < -t_{.025,8} < -2.306 \quad or \quad t > t_{\alpha/2,n-2} > t_{.025,8} > 2.306$

Test statistic:

$$S_{b_1} = \frac{S_\varepsilon}{\sqrt{(n-1)S_x^2}} = \frac{31.4772}{\sqrt{(10-1).24}} = 21.4175$$

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{(-39.1667) - 0}{21.4175} = -1.8287$$

Conclusion: Do not Reject H_0, there is not enough evident to infer that there is a linear relationship between the number of borrowers and the interest rate.

Confidence Interval Estimator of β_1**:**

$b_1 \pm t_{\alpha/2} S_{b_1} = -39.1667 \pm 2.306(21.4175)$

-39.1667 ± 49.3888

$LCL = -88.5555$

$UCL = 10.2221$

We estimate with 95% confidence the slope coefficient lies between –88.5555 and 10.2221.

d) Determine the coefficient of determination

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{(-9.4)^2}{(.24)(1248.8889)} = 0.2948$$

This statistic tells us that 29.48% of the variation in the number of borrowers is explained by the variation in interest rate. The remaining 70.52% is unexplained.

e) **Determine Coefficient of Correlation:**

$r = \sqrt{R^2} = \sqrt{0.2948} = -0.5430$

The coefficient of correlation (r) is – 0.5430, which tells us the negative relationship is moderately strong.

f) Estimate with 95% confident the number of **borrower** when the interest rate is 8%. (Prediction Interval)

$\hat{y} = b_0 + b_1 x_g = 475.1669 - 39.1667(8) = 161.8333$

$$\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)S_x^2}}$$

$$161.8333 \pm (2.306)(31.4772)\sqrt{1 + \frac{1}{10} + \frac{(8-8.2)^2}{(10-1).24}}$$

161.8333 ± 76.1662

$LCL = 85.6671$

$UCL = 237.9995$

We estimate with 95% confidence the number of borrower **lies between** 85.6671 and 237.9995.

g) Predict with 95% confident the **mean (average, expected value)** number of **borrowers** when the interest rate is 8%. (Confidence Interval Estimator)

$\hat{y} = b_0 + b_1 x_g = 475.1669 - 39.1667(8) = 161.8333$

$$\hat{y} \pm t_{\alpha/2,n-2} S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)S_x^2}}$$

$$161.8333 \pm (2.306)(31.4772)\sqrt{\frac{1}{10} + \frac{(8-8.2)^2}{(10-1).24}}$$

161.8333 ± 24.989

$LCL = 136.8443$

$UCL = 186.8223$

We estimate with 95% confidence the number of **borrowers lie** between 136.8443 and 2186.8223.