Buss-1020-hd-notes-updated-ranked-3-out-of-744

Quantitative Business Analysis (University of Sydney)

Week 1 – Types, levels, and sources of data

Week 1 LO's:

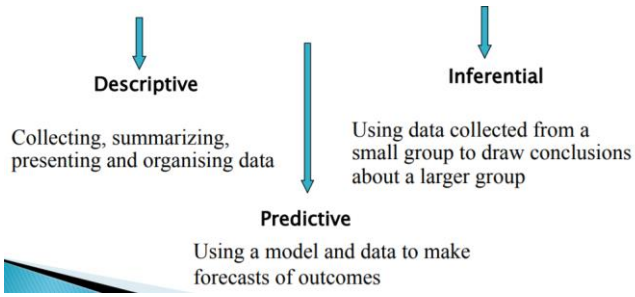
- Introduction to the Unit of Study
- **Introduction to Business Analytics and Statistics: Text pages 1-36**
  - **Business Analytics and Statistics**
  - Basic Vocabulary
  - Types of Variables
  - Data Collection
  - Sampling

Business analytics and statistics:

Framework for conducting statistical analyses: DCOVA

- Define the problem or objective and the data required
- Collect the required data in an appropriate manner
- Organise the data – 'clean it', prepare it for analyses, tabulate and summarize it
- Visualise the data
- Analyse the data

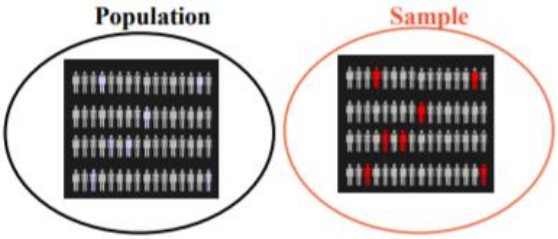
Three different branches of statistics:



- Descriptive: Summarizes and visualizes business data
  - Collect data e.g. survey
  - Summarize data e.g. sample average
  - Present data e.g. tables and graphs
- Inferential: Draw conclusions from business data → drawing conclusions about a large group of individuals based on a smaller group.
  - Estimation e.g. estimate population average amount spent using sample average spent
  - Hypothesis testing e.g. test claim that the population average amount spent in one group is larger than that in another group
- Predictive: Make forecasts about business activities → making reliable predictions based on a sample of data and a model.
  - Predict amount a new customer will spend, based on their attributes
  - Use a statistical model
- The three branches are used improve business processes:
  - Through better understanding of the data they generate
  - Through allowing better decisions to be made

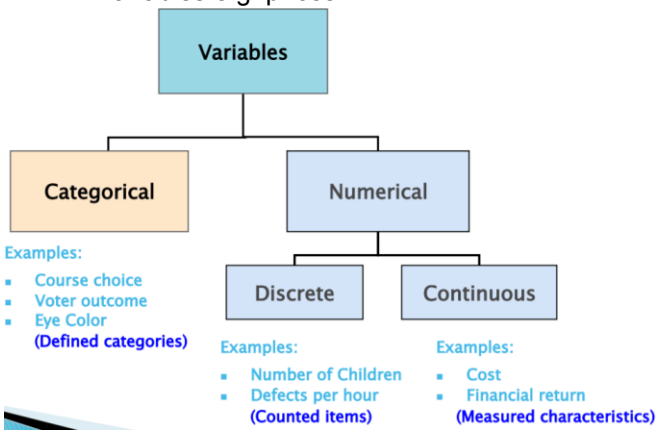
Basic vocabulary:

- Variable – characteristics of an item or individual → data on a variable is what we analyse when using a statistical method
- Data – observed values or outcomes of one or more variables
- Operational definition – a clearly defined meaning of a variable that is universally accepted by all associated with the analysis

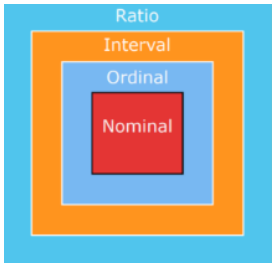
- Population vs. sample:
  - Population (N) – consists of all items or individuals where a conclusion is drawn (large group)
  - Sample (n) – portion of population selected for analysis (small group)
  - 

- Parameter – numerical measure that describes a relevant characteristic of a population
- Statistic – numerical measure that describes a characteristic of a sample → a statistic is often used to estimate a parameter

Types of variables:

- Categorical (qualitative) variable: have values that can only be placed into categories e.g. "yes" or "no" for male or female
- Numerical (quantitative) variable: have values that represent actual number quantities e.g:
  - Discrete variables arise from a counting process – can be finite or infinite e.g. number of HD students in BUSS1020
  - Continuous variables arise from a measuring process – can be assigned any real value within a given interval and can have infinitely many values e.g. return on stock is 12.53%
  - Sometimes, discrete variables with many outcomes are treated like continuous variables e.g. prices
- 

Levels of data management (measurement scales):

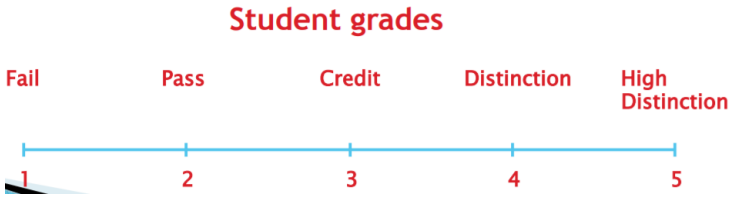
- Usage potential of various levels of data:
  - Ratio → highest level of measurement
  - Interval
  - Ordinal
  - Nominal → lowest level of measurement
- 

Nominal level data (for categorical data):

- Labels are used to distinguish different categories that have no order
- E.g. – employment classification:
  - 1. Teacher
  - 2. Construction worker
  - 3. Manufacturing worker
  - 4. Lawyer
  - 5. Doctor
  - 6. Other

Ordinal level data (for categorical data):

- Labels are used to classify AND to indicate rank or order
  - Often represents an underlying scale: e.g. quality
  - Differences between levels are not comparable
- E.g.2 – the lab tutorial session was:
  - 1. Helpful
  - 2. Somewhat helpful
  - 3. Moderately helpful
  - 4. Very Helpful
  - 5. Extremely helpful
- E.g.2:

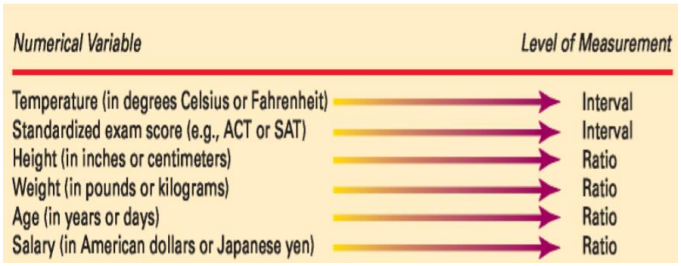


Interval level data (for numerical data):

- Data are numerical and differences between values have a consistent meaning
  - The location of zero is a matter of convenience or convention: not a natural or fixed data point
  - No “true” 0 (0 doesn’t mean literally nothing i.e. when the temperature is 0, it doesn’t mean we have no temperature)
- Examples:
  - Temperature, calendar time, monetary utility, scaled marks

Ratio level data (for numerical data):

- Highest level of measurement (can carry the most data)
- Same properties as interval data AND zero has a true meaning: represents absence of the thing being measured
- Measurement examples: height, weight, volume, price, profit, loss, revenue, expenses, financial ratios, return, inventory turnover, demand, supply, cost, counts
- Interval and Ratio Scales



Data collection:

Sources of data:

- Primary sources – analyst in firm collects and presents the data
- Secondary sources – analyst is not the data collector:
    - Data distributed by an organisation or individual → financial data on a company provided by investment services
        - Industry or market data from market research firms and trade associations
        - Stock prices, weather conditions, sports statistics in newspapers/online
    - A designed experiment:
        - Consumer testing of different versions of a product
        - Quality testing different supplier's material
        - Market testing product promotion effectiveness
        - Testing different web page designs
    - A survey → political/internet polls that determine customer satisfaction with a recent product or service experience
    - An observational study → market researchers using focus groups to elicit unstructured responses to open-ended questions
        - Time taken for customers to be served in a fast food outlet
        - Measuring volume of traffic at intersection to justify an advertisement
    - Automated and streaming data collection"
        - Mobile phone and data usage
        - GPS data
        - Social media feeds
- Useful summary that I prefer:
    - Data sets that cover large spatial areas and/or long time periods are usually distributed by an organization. In a designed experiment, the researcher subjects different groups to different conditions and observes the results. In a survey, people are asked questions about their beliefs, attitudes, behaviors, and other characteristics. In an observational study, the researcher collects data by directly observing a behavior, usually in a natural or neutral setting. Data collected by ongoing business activities can be collected from operational and transactional systems that exist in both physical and online settings, but can also be gathered from secondary sources such as third-party social media networks and online apps and website services that collect tracking and usage data.

Data formatting:

- Traditionally, data are stored in easy to use Excel tables where each column is used for each clearly defined variable → this is structured data (we focus on this in BUSS1020)
- A lot of modern data are unstructured – messy and not storable in an Excel file, or stored in several different locations e.g. video streams, email, texts, blogs etc.
    - Such data needs: data linking, preparation, and data cleaning before organisation can analyse
    - Cleaning data involves removing errors such as negative trade volumes/zero heart rates, flagging strange data points such as outliers, and filling in or deleting missing data

Sampling:

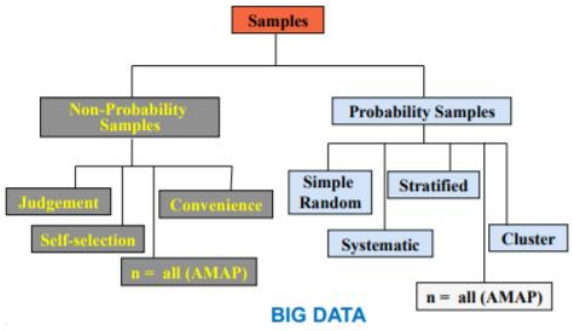
- We use sampling because we often cannot get data for the whole population
- Sample is less time-consuming and costly, and more practical than selecting every item in the population for analysis.
- Need to pick a sampling frame – lists of items that are in the population AND can be sampled
    - E.g.: Population lists, directories, customer databases, social media users, maps
- Inaccurate or biased results can result if parts of the population are excluded

Week 2 – Sampling, organising and visualising data

Week 2 LO's:

- Sampling (ctd from week 1)
- Categorical Data
- Numerical Data
- Principles of graphical excellence

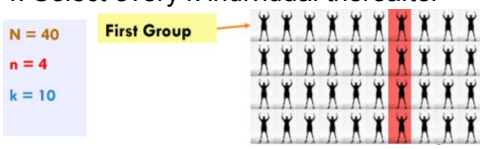
Types of samples:



Non-probability sampling:

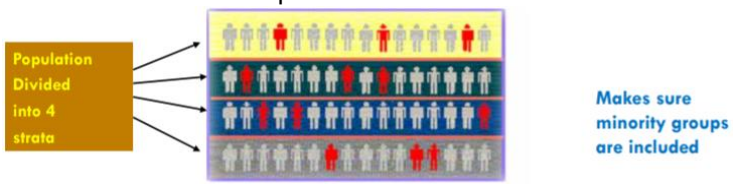
- Items are chosen without regard to their probability of occurrence
- Judgement sample: 'experts' select most appropriate items/people by convenience (but experts likely to have tendency to pick individuals that share same vision as sampler)
- Self-selected sample: individuals choose to participate (but only people with stronger opinions likely to participate)
- Convenience sample: selection is easy/inexpensive/quick e.g. snowball/chain sampling (each person then nominates another, but this can lead to bias trends as they will be connected somehow)
- Quota sample: pre-set quotas of groups chosen by convenience (but sampler may ignore population outside of targeted quota group e.g. 18-25 y/o's, sampler may choose only students and ignore those working in this age group)

Probability sampling:

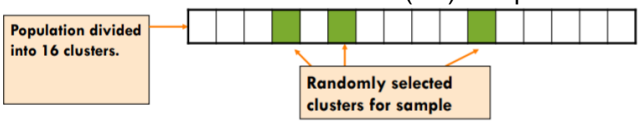
- Items are chosen randomly, sometimes using known probabilities that (closely) match those in population
- Simple random sample (SRS) – every individual/item (in the frame) has equal chance of being selected
  - Sampled items may be with replacement or without replacement
  - Often obtained via a random number generator or software
- Systematic sample:
  - 1. Decide on sample size (n)
  - 2. Divide sampling frame of N individuals into groups of k (k=N/n)
  - 3. A SRS is used to randomly select one individual from the first group
  - 4. Select every k individual thereafter
  - 

    - Suppose we pick person 7 from our first group of 10 using a SRS
    - We will continue to pick person 7 for the next three groups of 10
    - This will then give us a sample size of 4

- Stratified sample:
  - 1. Divide frame into strata according to a characteristic e.g. gender
  - 2. An SRS is selected from each strata, with sample size proportional to each strata's size
  - 3. Samples from each strata are combined into one sample
  - Common technique when sampling voters: e.g. stratify across racial/socio-economic/party affiliation variables
  - This technique ensures proportionate representation by ensuring that minority groups are included in the sample



- Cluster sample:
  - 1. Population is divided into several 'clusters', each already representative of the population
  - 2. A SRS of clusters is selected
  - 3. All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique
  - A common application of cluster sampling involves election exit polls, where certain election districts are selected and (full) sampled

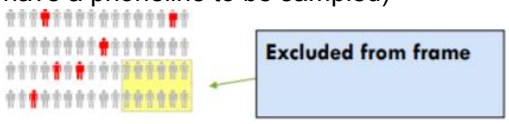


Comparing probability sampling methods:

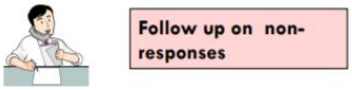
- SRS and systematic sampling:
  - Simple, cheap to use, effective against many types of bias
  - But may not give the best representation of the population's underlying characteristics (may not be able to catch the minority 5% whereas stratified guarantees those 5% has a say)
- Stratified sampling:
  - Ensures representation of individuals across the entire population, possibly in the right proportions
  - Effective against bias
  - Most efficient method, but costly
- Cluster sampling:
  - Quite cost effective
  - Can be less efficient (need large samples to be able to provide significant insight)

Types of survey errors:

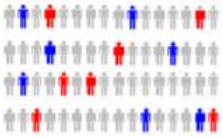
- 1. Coverage error or selection bias
  - Exists if some groups are excluded from the frame/population and have little/no chance of being selected (e.g. Truman election where people in rural areas didn't have a phoneline to be sampled)



- 2. Non-response error or bias
  - People who choose not to respond may be different from those who do respond



- 3. Sampling error
  - Variation from sample to sample will always exist, unless n=all



Random differences from sample to sample

(expected)

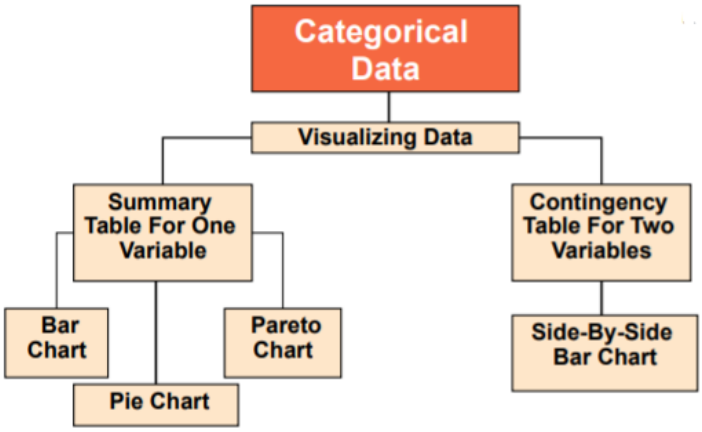
- 4. Measurement error
  - Weaknesses in question design, respondent error (e.g. tax survey show lower than true income values), and interviewer's effects on the respondent (Hawthorne effect)

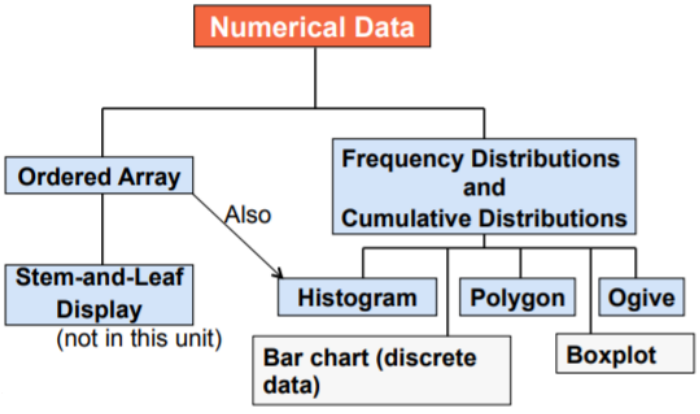
Ambiguous, unclear or leading question

- To evaluate survey worthiness:
  - First question its purpose, then whether it's a probability sample, and finally consider the different possible survey errors

Organising and visualising categorical and numerical data:

- Categorical Data
  - Organising One Variable Categorical Data — DCOVA
  - Visualising One Variable Categorical Data — DCOVA
  - Organising Two Variable Categorical Data — DCOVA
  - Visualising Two Variable Categorical Data — DCOVA
- Numerical Data
  - Organising Numerical Data — DCOVA
  - Visualising One Variable Numerical Data — DCOVA
  - Visualising Two Variable Numerical Data — DCOVA
- Principles of Graphical Excellence — DCOVA





Organising one variable categorical data:

- Summary table – indicates frequency, amount, percentage, or proportion in each category. This shows:
  - The relative frequency of each category
  - Differences between categories
- E.g.1:

Frequency table results for Type:
Count = 316

Type	Frequency	Percent of Total
Growth	227	71.8
Value	89	28.2

- E.g.2:

Frequency table results for Risk:
Count = 316

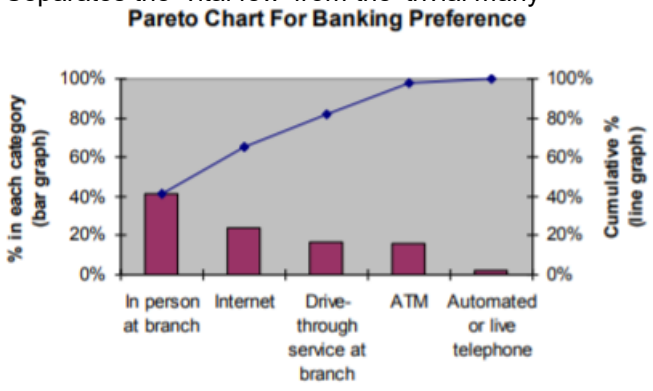
Risk	Frequency	Percent of Total	Cumulative Percent of Total
Low	212	67.1	67.1
Average	91	28.8	95.9
High	13	4.1	100

Retirement Funds.xlsx

  - Type for risk variable is categorical and scale in this case would be ordinal

Visualising one variable categorical data:

- Bar chart – bar separated into categories, and bar length represents amount, frequency, or percentage of values in that category
- Pie chart – shaded circle with one slice each category → slice size represents percentage in each category
- Pareto chart – for categorical, nominal scale data
  - A vertical bar chart where categories shown in descending order of frequency
  - A cumulative polygon is also shown on the same graph
  - Separates the 'vital few' from the 'trivial many'



Organising two variable categorical data:

- Contingency table
  - Can show pattern or relationship between two or more categorical variables
  - Cross tabulates or tallies jointly the responses of the categorical variables

	Average	High	Low	Total
Growth	74	10	143	227
Value	17	3	69	89
Total	91	13	212	316

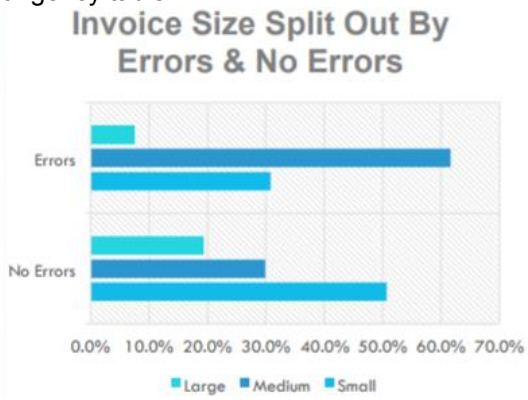
	Average	High	Low	Total
Growth Row % Column %	74 (32.6%) (81.32%)	10 (4.41%) (76.92%)	143 (63%) (67.45%)	227 (100%) (71.84%)
Value	17 (19.1%) (18.68%)	3 (3.37%) (23.08%)	69 (77.53%) (32.55%)	89 (100%) (28.16%)
Total	91 (28.8%) (100%)	13 (4.11%) (100%)	212 (67.09%) (100%)	316 (100%) (100%)

- E.g. 32.6% of growth funds are average risk funds and 81.32% of average risk funds are growth funds
- Value funds are proportionately more low risk and growth funds are proportionately more average and high risk

Visualising two variable categorical data:

- Side-by-side bar chart – represents data from a contingency table

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%



  - Invoices with errors are much more likely to be of medium size (61.57% vs 30.77% and 7.69%)
  - Much easier than using two pie charts (pie charts that don't clearly show percentage proportions can be misleading)
- Exam note: Categorical proportions can change over time and/or spatially

Organising numerical data:

- Ordered array – sequence of data, in rank order, from the smallest to the largest value

Age of Surveyed University Students	**Day Students**					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

  - Shows range (minimum to maximum value)
  - May help identify outliers
- Frequency distribution – summary table where data is arranged in numerically ordered classes
  - Select an appropriate number of class grouping, a suitable width for each class, and establish boundaries for each class to avoid overlapping
  - Number of groups depends on sample size → generally, a frequency distribution should have at least 5 classes
  - To determine the width of a class interval, divide the range of the data by the number of class desired

- E.g. – the Bureau of Meteorology (BOM) measures the rainfall (in mm.) in July 2013 for 20 Sydney suburbs:
  - Sort raw data in ascending order:
    12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
  - Find range: **58 - 12 = 46**
  - Select number of classes: **5 (usually between 5 and 15)**
  - Compute class interval (width): **10 (46/5 then round up)**
  - Determine class boundaries (limits):
    - **Class 1: 10 to less than 20**
    - **Class 2: 20 to less than 30**
    - **Class 3: 30 to less than 40**
    - **Class 4: 40 to less than 50**
    - **Class 5: 50 to less than 60**
  - Compute class midpoints: **15, 25, 35, 45, 55**
  - Count observations & assign to classes
- Frequency distribution:

Class	Midpoints	Frequency
>10 but less than 20	15	3
> 20 but less than 30	25	6
> 30 but less than 40	35	5
> 40 but less than 50	45	4
> 50 but less than 60	55	2
Total		20

- Relative frequency and percentage distribution:

Class	Frequency	Relative Frequency	Percentage
>10 but less than 20	3	.15	15
> 20 but less than 30	6	.30	30
> 30 but less than 40	5	.25	25
> 40 but less than 50	4	.20	20
> 50 but less than 60	2	.10	10
Total	20	1.00	100

- Cumulative frequency distribution:

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 ≤ X< 20	3	15%	3	15%
20 ≤ X< 30	6	30%	9	45%
30 ≤ X< 40	5	25%	14	70%
40 ≤ X< 50	4	20%	18	90%
50 ≤ X< 60	2	10%	20	100%
Total	20	100%	20	100%

Use of frequency distribution:

- Condenses raw data into a more useful form
- Allows a quick visual interpretation of the data
- Enables determination of the major characteristics of the data, including where the data are concentrated/clustered
- Allows a histogram to be drawn

Frequency distribution tips:

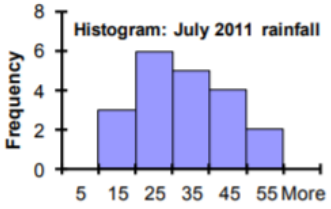
- Different group boundaries may provide different pictures for the same data (especially for smaller samples)
- As the size of the data set increases, the impact of changes to group boundaries is greatly reduced
- When comparing two or more data sets with different sample sizes, should use either a relative frequency or percentage distribution

Visualising one variable numerical data:

- Histogram – no gaps in between bars because a histogram represents a continuous numerical data set (although some bars might be "absent" reflecting no frequencies)
  - A histogram organizes data into groups (bins) so that the bin size reflects percentage of data points in each group
  - Vertical bar chart of a frequency distribution
  - Group boundaries (or midpoints) on horizontal axis.
  - Vertical axis/bar height: frequency, relative frequency, or percentage.

Class	Frequency	Relative Frequency	Percentage
>10 but less than 20	3	.15	15
>20 but less than 30	6	.30	30
>30 but less than 40	5	.25	25
>40 but less than 50	4	.20	20
>50 but less than 60	2	.10	10
Total	20	1.00	100

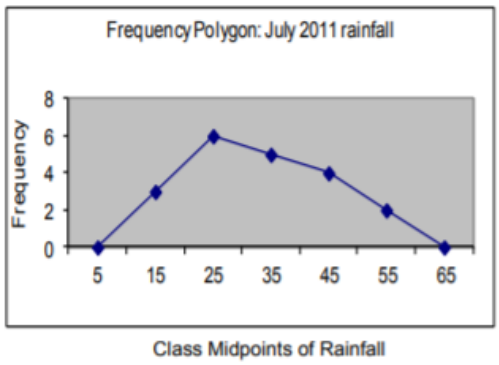
(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



- The polygon
  - A frequency or percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages
  - The cumulative frequency polygon or cumulative percentage polygon, (also known as ogive), displays the variable of interest along the X axis, and the cumulative amount along the Y axis
  - Useful when there are two or more groups to compare
  - Frequency polygon:

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

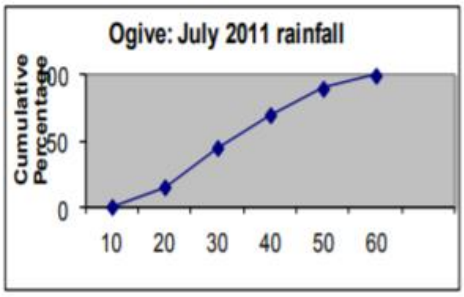
(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



  - Ogive (cumulative % polygon):

Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	0
20 but less than 30	20	15
30 but less than 40	30	45
40 but less than 50	40	70
50 but less than 60	50	90
60 but less than 70	60	100

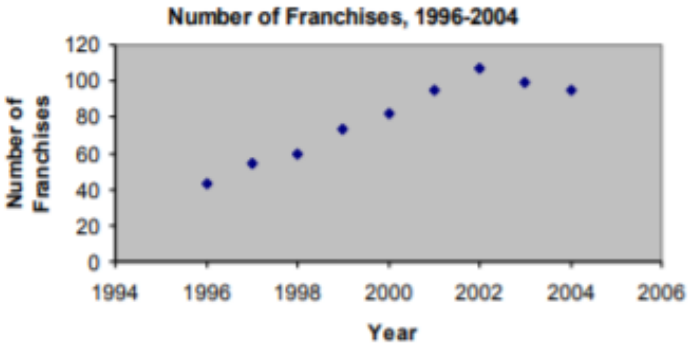
(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.



Visualising two variable numerical data:

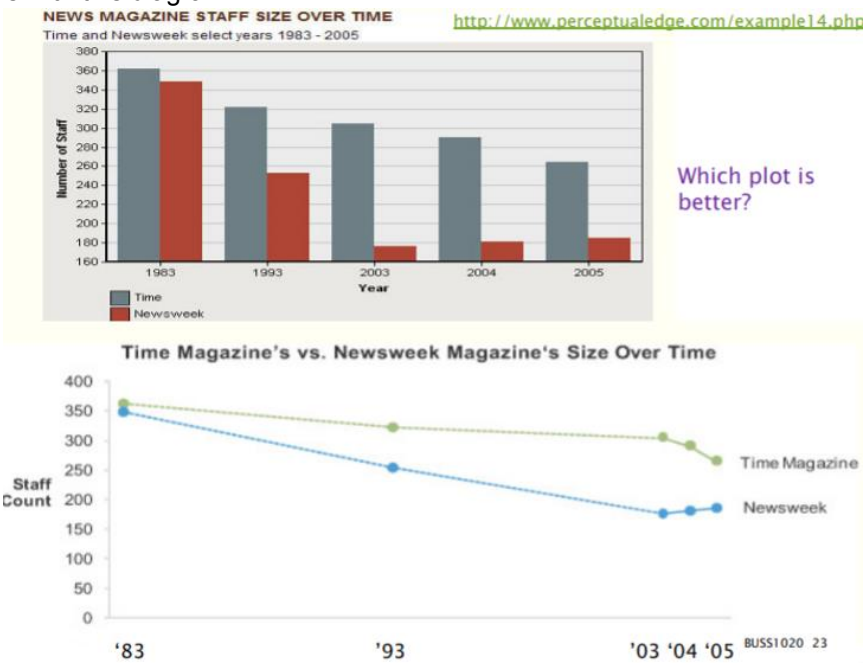
- Scatter plot - data consisting of paired observations on two variables
  - One variable measured on the vertical axis and the other on the horizontal axis
  - Used to examine possible relationships between two numerical variables
- Time-series plot - used to study patterns in a numeric variable over time
  - Numeric variable measured on the vertical axis and time period on the horizontal axis
  - Frequency of observations is often on an issue

Year	Number of Franchises
1996	43
1997	54
1998	60
1999	73
2000	82
2001	95
2002	107
2003	99
2004	95



Principles of graphical excellence:

- Maximise message, minimise noise – not distort the data with 'chart junk' or 'noise'
- The scale on the vertical axis should (usually) begin at zero (0).
- Include an informative title and clearly labelled axis
- Include a reference to the source
- 3D graphs should have a meaningful 3rd dimension. Usually 2D is sufficient.
- Graphs used should be the simplest possible that accurately tells the story – objectively convey the message in the data
- Problems with this diagram:



  - Y-axis doesn't start with 0 for bar graph
  - X-axis in bar graph doesn't show magnitude of time visually (inconsistent points)

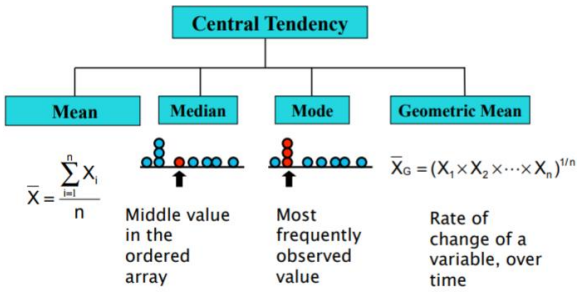
Week 3 - Numerical descriptive measures:

Week 3 LO's:

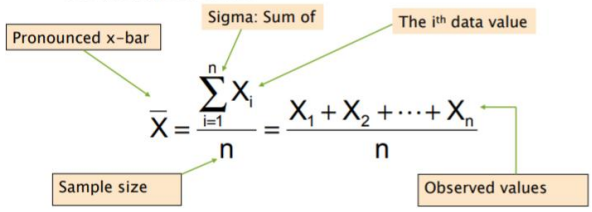
- Central tendency
  - Mean
  - Median
  - Mode
  - Geometric Mean
- Variation and shape
  - Range
  - Variance (Sample)
  - Standard Deviation (Sample)
  - Coefficient of Variation
  - Z-Scores
  - Shape: Skewness and Kurtosis
  - Descriptive Measures Using Excel
  - Quartiles and Boxplot
  - Boxplots using Statcrunch
- Population Measures
  - Mean and Variance
- Rules and Ethical Considerations
  - The Empirical Rule
  - Chebyshev's Rule
  - Ethical Considerations

- Summary of definitions:
  - Central tendency – the extent to which the data values group together around a typical or central value
  - Variation – the amount of dispersion or degree of scattering of values around the central value
  - Shape – the pattern in the distribution of values from the lowest value to the highest value

Central tendency:

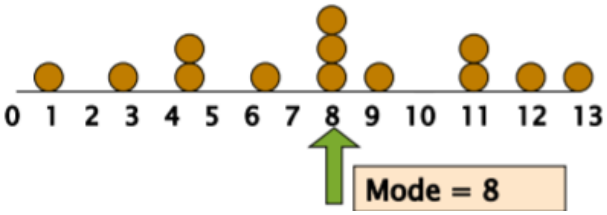


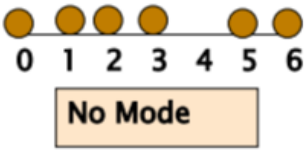
- Mean (arithmetic) – the most common measure of central tendency
  - For a sample of size n:



    - $\sum_{i=1}^{n} X_i$ represents the sum of values of X_i, starting at X_1 and ending with X_n
    - If i=3, it would be sum of values of X starting at X_3 to X_n
  - Can be affected by extreme values (e.g. outliers)
  - Only useful for numerical data
- Median – in an ordered array, the median is the "middle" number (50% above, 50% below)
  - Median position = $\frac{n+1}{2}$ position in the ordered data
    - If the number of values is odd, the median is the middle number
    - If the number of values is even, the median is taken as the average of the two middle numbers
    - Note: this is to determine the position of the median in the ordered array, not the value of the median
  - Useful for all data that has an order
  - Not affected by (a few) extreme values

- Mode – value that has "highest likelihood of occurring"/occurs the most frequently





  - There may be no mode or several modes
  - Not affected by extreme values
  - Used for both numerical and categorical data
- Geometric mean – often used to measure the rate of change of a variable over time
  - $\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$
- Geometric mean rate of return – measures the status of an investment over time
  - $\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$
    - R_t is the rate of return in time period t
  - Example: An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two

$$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$$

50% decrease 100% increase

Arithmetic mean rate of return: $\overline{X} = \frac{(-.5)+(1)}{2} = .25 = 25\%$ **Misleading result**

Geometric mean rate of return:

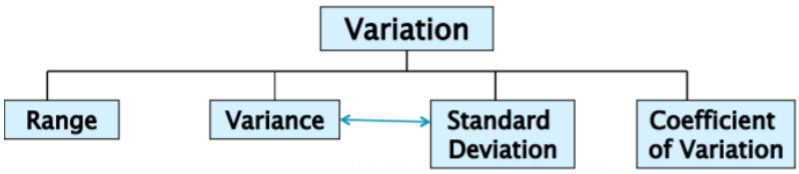
$$\begin{aligned}\overline{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1 \\ &= [(1+(-.5))\times(1+(1))]^{1/2} - 1 \\ &= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More representative result

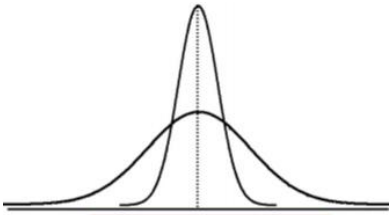
Which measure to choose:

- Mean:
  - Advantage – most commonly used and easy to calculate, most commonly used
  - Disadvantage – sensitive to outliers and can't be used for categorical data
- Median:
  - Advantage – less sensitive to outliers so it is the next most popular measure
  - Disadvantage – does not consider all the information (extreme values are important)
- Sometimes both the mean and the median are used.
- Mode:
  - Advantage – easy to answer the 'popularity' question
  - Disadvantage – usually reported for discrete or categorical data only, and it is not applicable in all the cases
- Geometric mean (return):
  - Advantage – useful to measure and track percentage changes
  - Disadvantage – cannot be applied to numbers with different signs

Variation and shape:



-
  - Give information on the spread/variability/dispersion of the data values



- Range – difference between largest and smallest values
  - The simplest
  - Can be used for all ordered data
  - Can be misleading as it ignores the way data are distributed and is highly sensitive to outliers
- Sample variance – average (approx.) of squared deviations of values from the mean
  - $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$ $\bar{X}$ = arithmetic mean; n = sample size; X_i = i^{th} value of the variable X
- Sample standard deviation (S) – shows variation about the mean
  - Most commonly used measure of variation
  - Is the square root of the variance
  - Use n – 1 not n for sample S and S^2 because this gives a more unbiased indicator of the population standard deviation → this acts as a correction for small samples (n might be small so want to divide by less to account for any large tails/variation)
  - Has the same units as the original data i.e. mean and S could both be in $ units
  - $$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$
  - Example:

Sample Data (X_i): 10 12 14 15 17 18 18 24

n = 8 Mean = $\bar{X}$ = 16

$$S = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{X})^2 + (14-\bar{X})^2 + \cdots + (24-\bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$ → A measure of the "average"/typical scatter around the mean

Coefficient of variation:

- Measures relative variation, compared to the mean
- As a percentage (%)
- Can be used to compare the variability of two or more sets of data measured in different units
- $CV = \left(\frac{S}{\bar{X}}\right) \cdot 100\%$
- E.g:
  - Stock A: Average price last year = $50 and S = $5 so CV = 10%
  - Stock B: Average price last year = $100 and S = $5 so CV = 5%
  - Stock C: Average price last year = $8 and S = $2 so CV = 25%
  - Both Stock A and B have same S, but stock B is less variable relative to its mean price
  - Stock C has a much smaller S than stock A, but has a much higher CV, so is more variable relative to its mean price

Summary characteristics of measures of variation:

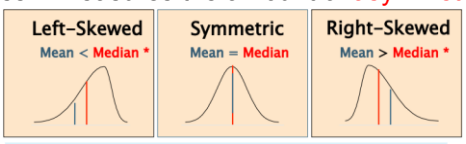
- The more the data are spread out, the greater the range, variance, and standard deviation
- The more the data are concentrated, the smaller the range, variance, and standard deviation
- If the values are all the same (no variation), all these measures will all be zero (0)
- None of these measures are ever negative

Assessing extreme observations: sample Z-score

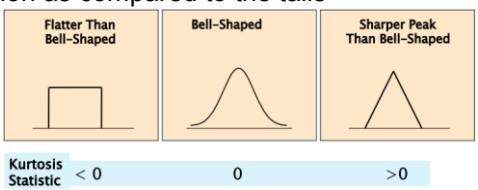
- The Z-score is the number of standard deviations a data value is from the mean.
- To compute the Z-score of a data value, subtract the mean from the value and divide by S
- A data value could be considered extreme (an "outlier") if its Z-score is less than or greater than 3 standard deviations from the mean → Z-score used to identify outliers
- The larger the absolute value of the Z-score, the farther the data value is from the mean
- $Z = \frac{X - \bar{X}}{S}$ where X represents the data value; X is the sample mean; S is the sample standard deviation
- If Z = 0, it means that the number has the same value as the mean

Shape of a distribution: skewness and kurtosis

- Skewness – measures the amount of asymmetry in a distribution



  - Value of 0 will give a perfectly symmetric distribution
  - A tail on the LHS means the distribution is left-skewed → values are dragging mean left (will give a negative skewness statistic)
- Kurtosis – measures the relative concentration/"peakedness" of values in the centre of a distribution as compared to the tails

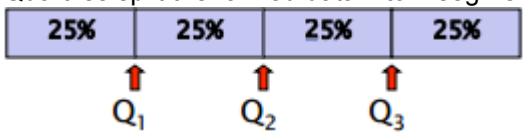


  - Positive value: sharper, more peaked, taller, fatter tails
  - Negative value: less/not peaked, shorter, thinner tails
- $Skewness = \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})^3}{S^3}$; $Kurtosis\ statistic = \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})^4}{S^4} - 3$ (not examinable)

Quartiles and boxplot:

Quartile measures:

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- - The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
  - Q2 is the median
  - Only 25% of the observations are greater than the third quartile Q3
- Locating quartile position in the ranked data:

  First quartile position: $Q_1 = (n+1)/4$ ranked value

  Second quartile position: $Q_2 = (n+1)/2$ ranked value

  Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

  - where n is the number of observed values
- Calculation rules for finding ranked position:
  - If the result is a whole number, then it is the ranked position to use
  - If the result is a fractional half (e.g. 2.5, or 7.5), then average the two corresponding data values
  - If the result is not a whole number or a fractional half, then round the result to the nearest integer to find the ranked position e.g. 1.25 → 1, or 5.75→6
- Example:

  Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

  (n = 9)

  Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data,
  so $Q_1 = (12+13)/2 = 12.5$

  Q_2 is in the $(9+1)/2 = 5^{th}$ position of the ranked data,
  so Q_2 = median = 16

  Q_3 is in the $3(9+1)/4 = 7.5$ position of the ranked data,
  so $Q_3 = (18+21)/2 = 19.5$

  Q_1 and Q_3 are measures of non-central location
  Q_2 = median, is a measure of central tendency

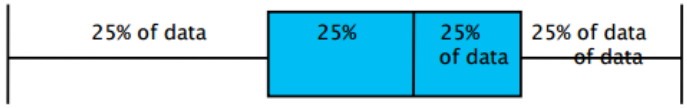
  -
- Interquartile range (IQR):
  - The IQR is Q3 – Q1 (= 19.5 – 12.5 = 7 for e.g. above)
  - IQR measures the spread in the middle 50% of the data
  - IQR is also called the mid-spread
  - It is a measure of variability, which is not influenced by outliers or extreme values
  - Measures like Q1, Q3, and IQR that are not influenced by outliers, so they are called “resistant” or “robust” measures
- The five-number summary:
  - $X_{Smallest}$
  - First Quartile (Q_1)
  - Median (Q_2)
  - Third Quartile (Q_3)
  - $X_{Largest}$

The boxplot:

- Boxplot – A graphical display of the data based on the five-number summary"

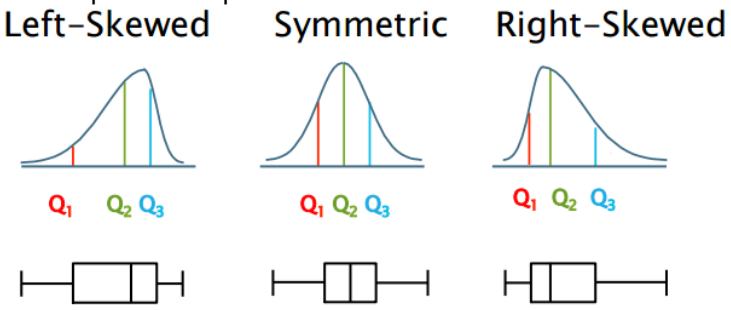
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



$X_{smallest}$ Q_1 Median Q_3 $X_{largest}$

- Distribution shape and boxplot:



  - Left-skewed: mean < median:
    - How I think about it: there are more extreme values on the left, away from the median value, which is what pulls the mean left → the 50% of values below the median is further away from median (than the 50% above the median), so this pulls the mean below the median
  - Symmetric: mean = median
  - Right-skewed: mean > median

Population measures

- Descriptive statistics describe a sample, not the population
- Summary measures of a population, i.e. parameters, are denoted with Greek letters e.g. μ, σ^2

Measure	Population Parameter	Sample Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	S^2
Standard Deviation	σ	S

	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	

- Exam note: n (or n – 1) is replaced by N (= population size) for population parameters
- Population mean:

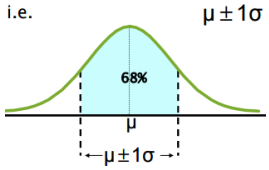
$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

- Population variance

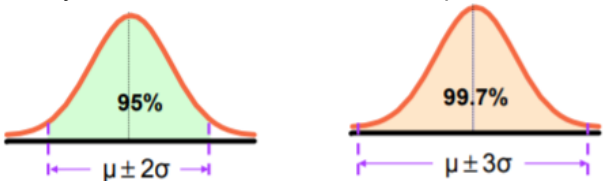
$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

The Empirical Rule:

- Approximates the way data varies in a bell-shaped/normal distribution
- Approximately 68% of the data in a bell-shaped distribution are within ± one standard deviation of the mean



- Approximately 95% of the data in a bell-shaped distribution lies within ± two standard deviations of the mean (i.e. $\mu \pm 2\sigma$)
- Approximately 99.7% of the data in a bell-shaped distribution lies within $\mu \pm 3\sigma$



- E.g.: Suppose that variable SAT scores is bell-shaped with a mean of 500 and standard deviation of 90, then approximately:

68% of all test takers scored between 410 and 590 (500 ± 90).

95% of all test takers scored between 320 and 680 (500 ± 180).

99.7% of all test takers scored between 230 and 770 (500 ± 270).

Chebyshev's Rule:

- Regardless of how the data are distributed, at least $(1 - \frac{1}{k^2})$ x 100% of the values will fall within k standard deviations of the mean (for k > 1)

At least	within
$(1 - 1/2^2) \times 100\% = 75\%$ ..... k=2	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% = 89\%$ ..... k=3	$(\mu \pm 3\sigma)$

- Suppose that a financial return variable is not bell-shaped distributed, but has mu = 0.5% and sigma = 1%. Then:
  - (k=2) At least 75% of all returns will lie between -1.5% and 2.5% (0.5 ± 2*1).
  - (k=3) At least 89% of all returns will lie between -2.5% and 3.5% (0.5 ± 3*1).
  - (k=3) At most 11% of all returns will lie outside -2.5% and 3.5%.

Ethical considerations:

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts

Week 4 - Basic probability:

Week 4 LO's:

- Probability concepts
- Basic probability rules
- Conditional probability
- Bayes' theorem
- Counting rules

Basic probability concepts:

- Probability – the chance, likelihood or possibility that an uncertain event will occur (always between 0 and 1)
- Impossible Event – an event that has no chance of occurring (probability = 0)
- Certain Event – an event that will definitely occur (probability = 1)

3 approaches to assessing probability:

- Three approaches to assess the probability of an uncertain (discrete or category) event
- A priori – based on prior knowledge of the process of the number of ways an event can occur
  - All outcomes are equally likely
  - $$\text{probability an event occurs} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of outcomes}}$$
  - E.g.: You are ranked number 50 in a tennis tournament of 100 players, who are ranked from 1 to 100. When randomly selecting your 1st round tennis opponent, what is the chance your opponent is in the top 10 rankings
    $$\frac{X}{T} = \frac{\text{10 top 10 players}}{\text{99 possible opponents}} = \frac{10}{99}$$
    - $= 0.101$
- Empirical probability – estimated from observed data of ways the event that has occurred
  - $$\text{probability of occurrence} = \frac{\text{number of ways the event } \textit{has} \text{ occurred}}{\text{number of trials}}$$
  - E.g.1: What is the probability your financial asset's price increases from today to tomorrow?
    $$\frac{X}{T} = \frac{\text{121 days of increases}}{\text{250 days considered}} = \frac{121}{250}$$
    - $= 0.484$
  - E.g.2: Find the probability of selecting a male taking statistics from the following table:

	Taking Stats	Not Taking Stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

    - Probability of male taking stats $= \frac{\text{number of males taking stats}}{\text{total number of people}} = \frac{84}{439} = 0.191$
- Subjective probability – based on an individual's past experience, personal opinion, and/or analysis of a particular situation → an opinion/guess not based on priori or observation
  - E.g.
    - What is the probability that Facebook dominates social media for the next 5 years?
    - What is the probability that a new start-up business remains solvent for 10 years?
    - What is the probability that Australia win the cricket test series in India?
    - What is the probability that a specific credit card transaction is fraudulent?

Events – each possible outcome of a variable

- Simple event – an event described by a single characteristic
  - E.g.: A customer purchases a product, or a die lands on 6
- Joint event – an event described by two or more characteristics
  - E.g.: A customer purchases a product and pays more than $100, or the roulette ball lands on red and an odd number
- Complement of an event (denoted A') – all events that are not event A
  - E.g.: The customer does not purchase the product

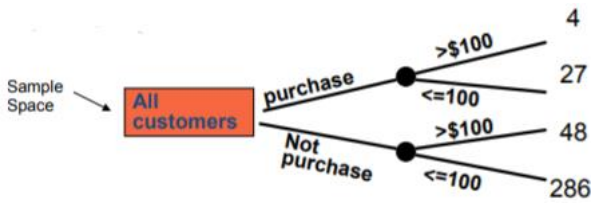
Sample space:

- The collection of all possible events – different, mutually exclusive, and collectively exclusive
  - E.g.: All 6 faces of a die, or all 52 cards in a deck

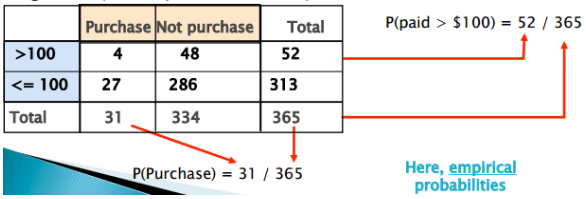
Visualising events:

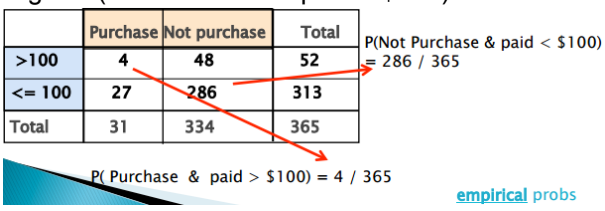
Contingency table:	Decision tree:

	Purchase	Not purchase	Total
> $100	4	48	52
<= $100	27	286	313
Total	31	334	365



Probabilities/events of a sample space

- Simple/marginal probabilities – the probability of a simple event:
  - E.g.1: P(Purchase a specific product)
  - E.g.2: P(total paid > $100)
  - 

- Joint probability – the probability of an occurrence of two or more events (a joint event)
  - E.g.1: P(Purchase & paid > $100)
  - E.g.2: P(Not Purchase & paid < $100)
  - 

- Mutually exclusive events – events that cannot occur simultaneously
  - E.g. – randomly choosing a day from 2018. A = day in Jan. and B = day in Feb.:
    - Events A and B are mutually exclusive.
- Collectively exhaustive events – one of the events must occur and this set of events covers the entire sample space
  - E.g. – randomly choosing a day from 2018. A = Weekday, B = Weekend, C = Jan., and D = Spring:
    - Events A, B, C, D are collectively exhaustive (but not mutually exclusive e.g. a weekday can be in January or Spring)
    - Events A and B are collectively exhaustive and mutually exclusive

Computing marginal and joint probabilities:

- Computing a marginal (or simple) probability:
  - MARGINAL PROBABILITY

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k) \quad (2)$$

where $B_1, B_2, \ldots, B_k$ are k mutually exclusive and collectively exhaustive events, defined as follows:

Two events are **mutually exclusive** if both the events cannot occur simultaneously.
A set of events is **collectively exhaustive** if one of the events must occur.

  - $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)$
    - Where $B_1, B_2, \ldots, B_k$ are k **mutually exclusive** and **collectively exhaustive** events
  - Expanding this gives:
  - E.g.:

P(willing to spend > $100)

$$= P(\text{Purch} \& > \$100) + P(\text{Not Purch.} \& > \$100) = \frac{4}{365} + \frac{48}{365} = \frac{52}{365}$$

	Purchase	Not Purchase	Total
>$100	4	48	52
< $100	27	286	313
Total	31	334	365

- The empirical probability of a joint event:
  - $P(A \text{ and } B) = \dfrac{\text{number of outcomes satisfying } A \text{ and } B}{\text{total number of outcomes}}$
  - E.g.:

P(Purchase and spend > $100)

$$= \frac{\text{number of customers that purchase AND spend} > \$100}{\text{total number of customers}} = \frac{4}{365}$$

	Purchase	Not Purchase	Total
>$100	4	48	52
< $100	27	286	313
Total	31	334	365

- Marginal and joint probabilities in a contingency table:

Event	Event B_1	Event B_2	Total
A_1	$P(A_1 \text{ and } B_1)$	$P(A_1 \text{ and } B_2)$	$P(A_1)$
A_2	$P(A_2 \text{ and } B_1)$	$P(A_2 \text{ and } B_2)$	$P(A_2)$
Total	$P(B_1)$	$P(B_2)$	1

Joint Probabilities

Marginal (Simple) Probabilities

Summary of probability concepts:

- Probability: numerical measure of how likely an event is to occur → it must satisfy:
  - $0 \le P(A) \le 1$ For any event A
- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1:
  - $P(A) + P(B) + P(C) = 1$
    If A, B, and C are *mutually exclusive* and *collectively exhaustive*

Basic probability rules:

- General addition rule:
  - $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
  - If A and B are mutually exclusive, then P(A and B) = 0, so the rule can be simplified:
    - $$P(A \text{ or } B) = P(A) + P(B)$$ For *mutually exclusive* events A and B
  - E.g.:

P(Purch. OR pay >$100) = P(Purch.) + P(>$100) – P(Purch.&>$100)

= 31/365 + 52/365 – 4/365 = 79/365

Don't count the four purchases with spend > $100 twice!

	Purchase	Not Purchase	Total
> $100	4	48	52
< $100	27	286	313
Total	31	334	365

- Multiplication rule:
  - $$P(A \text{ and } B) = P(A \mid B)P(B)$$
  - **Note:** If A and B are independent, then $P(A \mid B) = P(A)$ and the multiplication rule simplifies to $$P(A \text{ and } B) = P(A)P(B)$$
- Independence – two events are independent if and only if:
  - $$P(A \mid B) = P(A)$$
  - Events A and B are independent when the probability of one event is not affected or changed by the other event
  - Possibility of A given B, is just the possibility of A occurring itself – B has no impact

Conditional probability:

- A conditional probability is the probability of one event, given that another event has occurred:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$ → The conditional probability of A given that B has occurred

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$ → The conditional probability of B given that A has occurred

Where P(A and B) = joint probability of A and B
P(A) = marginal or simple probability of A
P(B) = marginal or simple probability of B

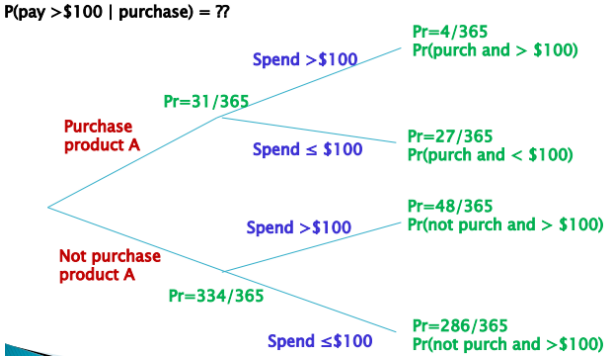
- Therefore: $$P(A \text{ and } B) = P(A \mid B) \times P(B)$$ (i.e. multiplication rule)

- E.g.1:

P(pay >$100 | purchase) = P(pay >$100 and purchase) = 4/365
P(purchase) 31/365
= 4/31 = 0.129

	Purchase	Not Purchase	Total
> $100	4	48	52
< $100	27	286	313
Total	31	334	365

P(pay >$100 | purchase) = ??



- E.g.2:

Of Amazon customers, 90% spent more than 5 minutes on the website and 40% made a purchase. 35% of customers did both.

	Purchase	No purchase	Total
> 5 mins	0.35	0.55	0.90
< 5 mins	0.05	0.05	0.10
Total	0.40	0.60	1.00

$$P(\text{purchase} \mid > 5\text{ mins}) = \frac{P(\text{purchase and } > 5\text{ mins})}{P(> 5\text{ mins})} = \frac{0.35}{0.90} = 0.3889$$

Bayes' Theorem:

- Used to revise prior or existing probabilities based on new information:
  - It allows us to usefully reverse the conditioning between two events
  - We can find probability of B given A if we know the probability A given B
- It is an extension of conditional probability
- Short version:

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

When would this be useful

where:

B = event of interest

A = new event that might impact P(B)

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

- Long version:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)}$$

  - where:
    B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

    A = new event that might impact $P(B_i)$ NB $P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$

  - To solve the denominator:
    - Expand out the marginal probability P(A), and then use multiplication rule with each B_i

- Example:

event S = successful television event F = favorable report
event S' = unsuccessful television event F' = unfavorable report

and

$$P(S) = 0.40 \quad P(F|S) = 0.80$$
$$P(S') = 0.60 \quad P(F|S') = 0.30$$

Then, using Equation (9),

$$P(S|F) = \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|S')P(S')} = \frac{(0.80)(0.40)}{(0.80)(0.40) + (0.30)(0.60)} = \frac{0.32}{0.32 + 0.18} = \frac{0.32}{0.50} = 0.64$$

The probability of a successful television, given that a favorable report was received, is 0.64. Thus, the probability of an unsuccessful television, given that a favorable report was received, is $1 - 0.64 = 0.36$.

Counting rules:

- Often discrete probabilities rely on being able to count all the events possible
  - When there are many, many possible events this can be tricky and time consuming, so counting rules can help

Counting Rule 1 Counting rule 1 determines the number of possible outcomes for a set of mutually exclusive and collectively exhaustive events.

COUNTING RULE 1
If any one of k different mutually exclusive and collectively exhaustive events can occur on each of n trials, the number of possible outcomes is equal to

$$k^n \quad (10)$$

- For example, using Equation (10), the number of different possible outcomes from tossing a two-sided coin five times is $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$.

Counting Rule 2 The second counting rule is a more general version of the first counting rule and allows the number of possible events to differ from trial to trial.

COUNTING RULE 2

If there are k_1 events on the first trial, k_2 events on the second trial, . . . , and k_n events on the nth trial, then the number of possible outcomes is

$$(k_1)(k_2) \ldots (k_n) \quad (11)$$

For example, a state motor vehicle department would like to know how many license plate numbers are available if a license plate number consists of three letters followed by three numbers (0 through 9). Using Equation (11), if a license plate number consists of three letters followed by three numbers, the total number of possible outcomes is $(26)(26)(26)(10)(10)(10) = 17{,}576{,}000$.

- **Counting Rule 3** The third counting rule involves computing the number of ways that a set of items can be arranged in order.

COUNTING RULE 3

The number of ways that all n items can be arranged in order is

$$n! = (n)(n-1) \ldots (1) \quad (12)$$

where $n!$ is called n factorial, and 0! is defined as 1.

- **Counting Rule 4** In many instances you need to know the number of ways in which a subset of an entire group of items can be arranged in *order*. Each possible arrangement is called a **permutation**.

COUNTING RULE 4: PERMUTATIONS

The number of ways of arranging x objects selected from n objects in order is

$$_nP_x = \frac{n!}{(n-x)!} \quad (13)$$

where

n = total number of objects
x = number of objects to be arranged
$n!$ = n factorial = $n(n-1) \ldots (1)$
P = symbol for permutations[1]

- Order matters for permutations

Counting Rule 5 In many situations, you are not interested in the *order* of the outcomes but only in the number of ways that x items can be selected from n items, *irrespective of order*. Each possible selection is called a **combination**.

COUNTING RULE 5: COMBINATIONS

The number of ways of selecting x objects from n objects, irrespective of order, is equal to

$$_nC_x = \frac{n!}{x!(n-x)!} \quad (14)$$

where

n = total number of objects
x = number of objects to be arranged
$n!$ = n factorial = $n(n-1) \ldots (1)$
C = symbol for combinations[2]

If you compare this rule to counting rule 4, you see that it differs only in the inclusion of a term $x!$ in the denominator. When permutations were used, all of the arrangements of the x objects are distinguishable. With combinations, the $x!$ possible arrangements of objects are irrelevant.

- Order does not matter e.g. 456 and 654 are treated the same

Week 5 – Discrete probability distributions

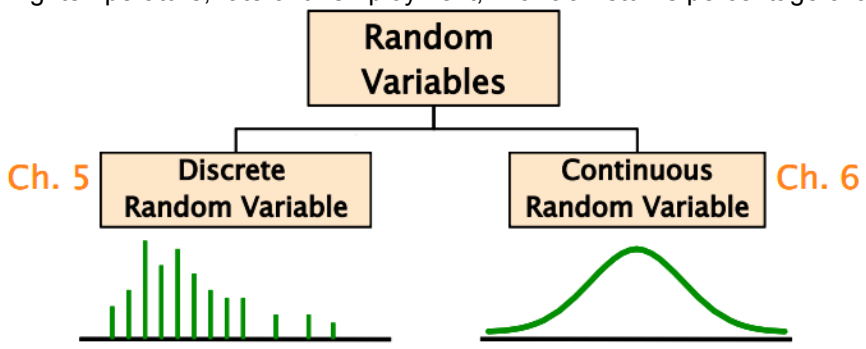
Week 5 LO's:

- Discrete random variables and probability distributions
- Common Discrete Probability Distributions
  - Binomial distribution
  - Poisson distribution
  - Hypergeometric distribution
- Covariance

Discrete random variables and probability distributions:

Definitions:

- Random variable (r.v.) – represents the possible outcomes from an uncertain event
- Numerical r.v.s can be discrete or continuous
- Discrete r.v.: set of all possible outcomes is a finite, or "countably infinite", number of values
  - E.g. prices and price changes, number of absent employees, number of new subscribers
- Continuous r.v.: takes values at every point in a given interval
  - E.g. temperature, rate of unemployment, financial returns/percentage changes

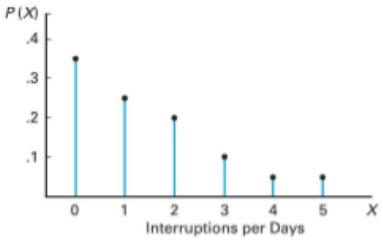


Probability distribution for a discrete random variable:

- A mutually exclusive and collectively exhaustive list of all possible outcomes for that r.v., and the associated probabilities for each outcome

Outcomes	Probability
X_1	p_1
X_2	p_2
:	:
X_M	p_M

Interruptions per Day	Probability
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05



E.g.:

Expected value (mean) of a discrete variable:

Expected Value (or mean) of a discrete random variable

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{M} x_i P(X = x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This does not violate the previous formula given

$$\begin{aligned}\mu = E(X) &= \sum_{i=1}^{N} x_i P(X = x_i)\\ &= (0)(0.35) + (1)(0.25) + (2)(0.20) + (3)(0.10) + (4)(0.05) + (5)(0.05)\\ &= 0 + 0.25 + 0.40 + 0.30 + 0.20 + 0.25\\ &= 1.40\end{aligned}$$

  - E.g.

Interruptions per Day (x_i)	$P(X = x_i)$	$x_i P(X = x_i)$
0	0.35	(0)(0.35) = 0.00
1	0.25	(1)(0.25) = 0.25
2	0.20	(2)(0.20) = 0.40
3	0.10	(3)(0.10) = 0.30
4	0.05	(4)(0.05) = 0.20
5	0.05	(5)(0.05) = 0.25
	1.00	$\mu = E(X) = 1.40$

  - Note: we use $\frac{1}{N}\sum_{i=1}^{N} x_i$ when the probability of each outcome is the same
- Exam note: Expected value is a population measure → reflects population values and their associated probabilities

Variance and S.D. of a discrete variable:

- Variance of a discrete variable:

$$\begin{aligned}\sigma^2 &= E(X^2) - [E(X)]^2\\ &= E[X - E(X)]^2\\ &= \sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i)\end{aligned}$$

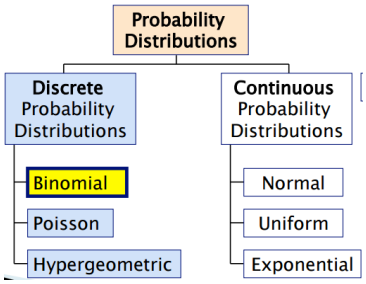
  - $\sigma^2 = \mathrm{E}\left[(\mathrm{X} - \mathrm{E}(\mathrm{X}))^2\right]$
  - Notation: E( ) = expected value
  - Thus: $\sigma^2 = \sum_{i=1}^{M} (x_i - \mathrm{E}(\mathrm{X}))^2 \mathrm{P}(\mathrm{X} = x_i)$
- Standard deviation of a discrete variable:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{M} [x_i - \mathrm{E}(\mathrm{X})]^2 \mathrm{P}(\mathrm{X} = x_i)}$$

  where:
  E(X) = Expected value of the discrete random variable X
  x_i = the i^{th} outcome of X
  $P(X=x_i)$ = Probability of the i^{th} occurrence of X
- Exam note: Variance and standard deviation of discrete variables are also population measures → reflects population values and their associated probabilities
- E.g. continued:

Interruptions per Day (x_i)	$P(X = x_i)$	$x_i P(X = x_i)$	$[x_i - E(X)]^2$	$[x_i - E(X)]^2 P(X = x_i)$
0	0.35	0.00	$(0 - 1.4)^2 = 1.96$	(1.96)(0.35) = 0.686
1	0.25	0.25	$(1 - 1.4)^2 = 0.16$	(0.16)(0.25) = 0.040
2	0.20	0.40	$(2 - 1.4)^2 = 0.36$	(0.36)(0.20) = 0.072
3	0.10	0.30	$(3 - 1.4)^2 = 2.56$	(2.56)(0.10) = 0.256
4	0.05	0.20	$(4 - 1.4)^2 = 6.76$	(6.76)(0.05) = 0.338
5	0.05	0.25	$(5 - 1.4)^2 = 12.96$	(12.96)(0.05) = 0.648
	1.00	$\mu = E(X) = 1.40$		$\sigma^2 = 2.04$
				$\sigma = \sqrt{\sigma^2} = 1.4283$

All probability distributions:



Discrete probability distributions:

Binomial probability distribution:

- R.v. X counts number of "events of interest" occurring from a fixed number of observations of trials n
- The sample consists of a fixed number of observations, n
- There is an upper limit to the number of "events of interest" that could occur
- Each observation is classified as to whether or not the "event of interest" occurred
  - The two categories are mutually exclusive and collectively exhaustive e.g. head and tails
  - P(event of interest occurs) = π
  - P(event doesn't occur) = 1 – π
  - Each observation has a constant probability for the event of interest occurring (denoted as π)
- Value of any observation is independent of the value of any other observation
  - The outcome of one observation does not affect any other observation or trial
  - Two random sampling methods can deliver independence, sampling from an:
    - Infinite population without replacement
    - Finite population with replacement
- Finding probability for event of interest:
  - When: X = n or X = 0:
    - $P(X=n)=\pi^n$
    - $P(X=0)=(1-\pi)^n$ *(by independence)*

    E.g.1:

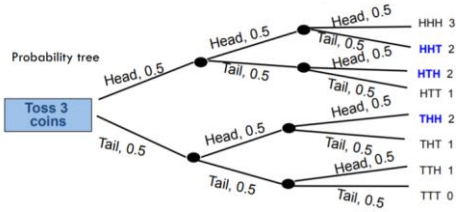
    E.g. The probability of getting 2 heads in 2 coin flips is 0.5*0.5 = 0.25, i.e.

    - $P(X=2\,|\,n=2,\pi=0.5)=0.5^2$
  - E.g.2 (when X ≠ n):

    Event of interest: obtaining heads on a coin toss.
    n = 3 tosses. In how many ways can you get exactly X = 2 heads?

    Possible ways: HHT, HTH, THH; i.e. three ways of getting exactly X = 2, each have probability: 0.5^3 = 0.125.

    - Thus P(X=2|n=3) is 3*0.125 = 0.375

    - This is fairly simple, but we need to be able to count the number of ways for more complicated and general situations, including when n is large, so we use counting rules in our formula

- Formula:

$$P(X = x|n,\pi) = {}^{n}C_{x}\pi^{x}(1-\pi)^{n-x}$$

$$= \frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$

P(X=x|n,π) = probability of **x** events of interest in **n** trials, with the probability of an "event of interest" being π for each trial

x = number of "events of interest" in sample, (x = 0, 1, 2, ..., n)

n = sample size (number of trials or observations)

π = probability of "event of interest"

Example: Flip a coin four times, let x = # heads:

n = 4

π = 0.5

1 - π = (1 - 0.5) = 0.5

X = 0, 1, 2, 3, 4

- E.g.1: What is the probability of one success in five observations if the probability of an event of observation is 0.1?

x = 1, n = 5, and π = 0.1

$$P(X=1|5,0.1) = \frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$

$$= \frac{5!}{1!(5-1)!}(0.1)^{1}(1-0.1)^{5-1}$$

$$= (5)(0.1)(0.9)^{4}$$

$$= 0.32805$$

- E.g.2: Suppose the probability of purchasing a defective computer is 0.02. What is the probability of purchasing 2 or fewer defective computers in a group of 10?

$$P(X \le 2|10,0.02) = P(X=0) + P(X=1) + P(X=2)$$

$$= \sum_{x=0}^{2} \frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$

$$= \sum_{x=0}^{2} \frac{10!}{x!(10-x)!}(.02)^{x}(1-.02)^{10-x}$$

$$= 0.999$$

- Characteristics:
  - Mean

$$\mu = E(X) = n\pi$$

  - Variance and Standard Deviation

$$\sigma^2 = n\pi(1-\pi)$$

$$\sigma = \sqrt{n\pi(1-\pi)}$$

Where n = sample size

π = probability of the event of interest for any trial

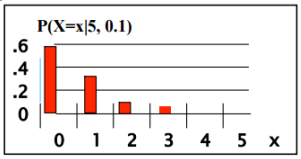
(1 – π) = probability of no event of interest for any trial

- Shape of binomial distribution depends on value of π and n:

Examples

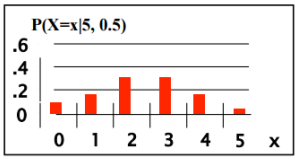
$$\mu = n\pi = (5)(.1) = 0.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.1)(1-.1)} = 0.6708$$



$$\mu = n\pi = (5)(.5) = 2.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.5)(1-.5)} = 1.118$$



Poisson probability distribution:

Definitions:

- Applies when counting the number of times an event occurs in a given "area/time of opportunity"
- An area of opportunity is a continuous interval of time, area, volume, etc., in which at least one occurrence of an event can occur e.g. the number of paint scratches on a car in a year
- Apply this distribution when:
  - Interest is in counting the number of times an event occurs in a given area/time/window of opportunity
  - There is no clear upper limit to the event of interest occurring
  - The probability that an event occurs in one area of opportunity is constant for all areas of opportunity
  - Events occur independently of each other
  - As the area of opportunity becomes smaller, probability of event occurring approaches zero
- The average number of events per area of opportunity is λ (lambda)
- Formula:

$$P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where:
x = number of events observed in an area of opportunity
λ = expected number of events
e = base of the natural logarithm system (2.71828...)

- Characteristics:
  - Mean

$$E(X) = \lambda$$

  - Variance and Standard Deviation

$$V(X) = \sigma^2 = \lambda$$

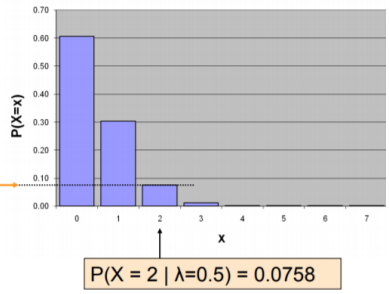
$$\sigma = \sqrt{\lambda}$$

  - where λ = expected number of events
  - Exam note: Variance and mean both equal lambda
- E.g.: On average is receive one letter every 2 days. What is the probability I receive 2 letters in one day?
  - The relevant window of opportunity here is one day
  - Lambda (average letters received) for 1 day is 0.5, so we have λ = 0.5

Graphically:

λ = 0.5

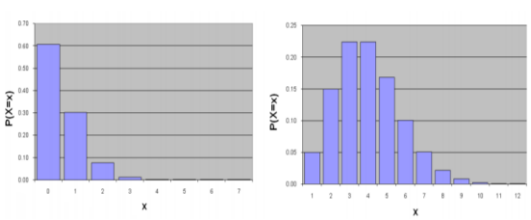
X	P = 0.5
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
6	0.0000
7	0.0000



P(X = 2 | λ=0.5) = 0.0758

- The shape of the Poisson distribution depends on the **parameter lambda**:

λ = 0.5 λ = 3



Hypergeometric probability distribution:

Definitions:

- The binomial distribution is applicable when selecting from a finite population with replacement, OR, selecting a finite sample of n from an infinite population without replacement
- The hypergeometric distribution is applicable when selecting from a finite population without replacement
- Apply this distribution when:
  - "n" trials in a sample taken from a finite population of size N
  - Sample taken without replacement
  - Outcomes of trials are dependent – each trial is now dependent on previous chosen outcomes as there is no replacement
  - Concerned with finding the probability of "X = x_i" items of interest in the sample where there are "A" items of interest in the population
  - E.g. drawing coloured balls from urns, or drawing cards from a deck
- Formula:

$$P(X = x \mid n, N, A) = \frac{[{}_A C_x][{}_{N-A} C_{n-x}]}{{}_N C_n} = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$$

Where

N = population size

A = number of *items of interest* in the population

N – A = number of items *not* of interest in the population

n = sample size taken

x = number of *items of interest* in the sample

n – x = number of events not of interest in the sample

- Characteristics:
  - The mean of the hypergeometric distribution is

$$\mu = E(X) = \frac{nA}{N}$$

  - The standard deviation is

$$\sigma = \sqrt{\frac{nA(N-A)}{N^2}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Where $\sqrt{\frac{N-n}{N-1}}$ is called the "Finite Population Correction Factor" used when sampling without replacement from a finite population

- E.g.:

Example: 3 different computers are selected from 10 in the department. 4 of the 10 computers have illegal software loaded. What is the probability that 2 of the 3 selected computers have illegal software loaded?

N = 10	n = 3
A = 4	x = 2

$$P(X = 2 \mid 3, 10, 4) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}} = \frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{(6)(6)}{120} = 0.3$$

The probability that 2 of the 3 selected computers have illegal software loaded is 0.30, or 30%.

Covariance and summing random variables:

Covariance:

$$\sigma_{XY} = \sum_{i=1}^{N} [x_i - E(X)][(y_i - E(Y)]\, P(x_i, y_i)$$

where: X = discrete random variable X
X_i = the i^{th} outcome of X
Y = discrete random variable Y
Y_i = the i^{th} outcome of Y
$P(X_i,Y_i)$ = probability of occurrence of the i^{th} outcome of X and the i^{th} outcome of Y

- $$\sigma_{XY} = E([X - E(X)][Y - E(Y)])$$
- The covariance measures the strength of the linear relationship and correlation between two numerical random variables X and Y
- A positive covariance indicates a positive relationship of the variables
- A negative covariance indicates a negative relationship of the variables
- E.g. – consider the return per $1000 for two types of investments:

Investment Returns

Consider the return per $1000 for two types of investments.

Prob.	Economic Condition	Investment	
		Value Fund X	Momentum Fund Y
0.2	Recession	- $25	- $200
0.5	Stable Economy	+ $50	+ $60
0.3	Expanding Economy	+ $100	+ $350

$E(X) = \mu_X = (-25)(.2) + (50)(.5) + (100)(.3) = 50$

$E(Y) = \mu_Y = (-200)(.2) + (60)(.5) + (350)(.3) = 95$

Interpretation: The Value fund X is averaging a $50.00 return, while the Momentum fund Y is averaging a $95.00 return, per $1000 invested.

- Mean:

$$\sigma_X = \sqrt{(-25-50)^2(.2) + (50-50)^2(.5) + (100-50)^2(.3)} = 43.30$$

$$\sigma_Y = \sqrt{(-200-95)^2(.2) + (60-95)^2(.5) + (350-95)^2(.3)} = 193.71$$

Interpretation: Even though Momentum fund Y has a higher average return, it is subject to much more variability and the amount of loss/gain is higher, compared to the Value fund X.

- Standard deviation:

$$\sigma_{XY} = (-25-50)(-200-95)(.2) + (50-50)(60-95)(.5) + (100-50)(350-95)(.3) = 8{,}250$$

- Covariance:
- Covariance interpretation: covariance is positive, so there is a positive relationship between the two investment funds, meaning that they tend to rise and fall together

The sum of two random variables:

- Expected value of the sum of two random variables:
  - $$E(X+Y)=E(X)+E(Y)$$
- Variance of the sum of two random variables:
  - $$Var(X+Y)=\sigma^2_{X+Y}=\sigma^2_X+\sigma^2_Y+2\sigma_{XY}$$
- Standard deviation of the sum of two random variables:
  - $$\sigma_{X+Y}=\sqrt{\sigma^2_{X+Y}}$$

The weighted sum of two random variables:

- Expected Value of the weighted sum of two random variables:
  - $$\begin{aligned}E(aX+bY)&=E(aX)+E(bY)\\&=aE(X)+bE(Y)\end{aligned}$$
- Variance of the weighted sum of two random variables:
  - $$\begin{aligned}Var(aX+bY)=\sigma^2_{aX+bY}&=\sigma^2_{aX}+\sigma^2_{bY}+2\sigma_{aXbY}\\&=a^2\sigma^2_X+b^2\sigma^2_Y+2ab\sigma_{XY}\end{aligned}$$
  - If it was Var(aX-bY): still add the squared variances together but subtract 2ab covariance

Portfolio returns: wX + (1-w)Y

- Expected Value of the weighted sum of two asset returns X, Y:
  - $$E(wX+(1-w)Y)=wE(X)+(1-w)E(Y)$$
- Variance of the weighted sum of two asset returns X, Y:
  - $$Var(wX+(1-w)Y)=w^2\sigma^2_X+(1-w)^2\sigma^2_Y+2w(1-w)\sigma_{XY}$$
- Choice of w will depend on whether the investors wants to to maximise returns (higher expected value) or minimising risk (lower variance)

Tips for quiz 1 and 2:

- Remember to bring your student ID card, plus pens, pencils, eraser, non-programmable calculator
- You will be given access to Excel (only) for calculations but all final answers must be given on paper
- Note that answers may need to be in pen
- Important: use plenty of words (not just symbols).
  - Demonstrate your understanding by showing all your working out (all intermediate steps)

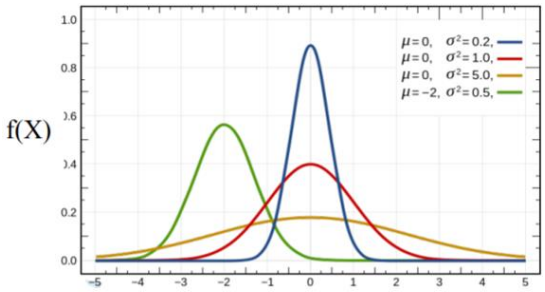
Week 6 – Continuous probability distributions

Week 6 LO's:

- Continuous Probability Distributions Introduction
- Normal Distribution
- Uniform Distribution
- Exponential Distribution

Continuous probability distributions:

- A continuous random variable can assume any value on a continuum/range, e.g.:
  - Thickness, height, weight, volume of an item
  - Time required to complete a task, time between events
  - Temperature
  - Financial return, percentage change
- A continuous random variable can potentially take on any value, depending on the ability to accurately and finely measure it
- Continuous random variables relevant to business:
  - Weight of cans, packets, boxes
  - Download times, web query times
  - Financial return, volume of trades, time between trades
- Continuous probability density:
  - Instead of probabilities for each value of X, a continuous r.v. has a probability density function

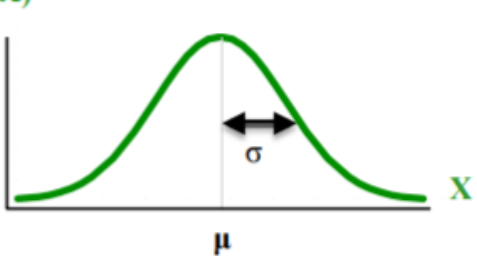


  -
  - Exam note: f(X) it is the relative probability of one point in that area compared to another point in another area vs. discrete distributions where the height of the function represents the true probability
  - Each density curve represents the relative likelihood of each X value
    - The area of each shaded region is the probability that X is in that region
  - We only consider the probability of X being in a certain range/region of values)
    - $$P(a < X < b) = \int_a^b f(x)dx$$
  - This is because there are essentially infinite possible outcomes for a continuous r.v., so the probability of any individual outcome is 0 → area under a single value for X is 0 i.e.:
    - $$P(X = a) = P(a \le X \le a) = \int_a^a f(x)dx = 0$$
  - Total area under the entire density curve is always equal to 1 i.e. if P(a < X < b) = 1, then (a, b) covers all possible values of X
    - $$1 \approx \int_{-\infty}^{\infty} f(x)dx$$

Normal/Gaussian probability distribution:

Definitions:

- Normal probability distribution is a bell-shaped density curve
  - Distribution is therefore symmetric (skewness = 0)
  - Mean = median = mode
  - The random variable has an infinite theoretical range
- Characteristics:
  - The mean parameter is mu (μ)
  - The standard deviation parameter is sigma (σ)
- Shape of the normal distribution curve:
  - Location of distribution is given by μ i.e. changing μ shifts the distribution left or right
  - Spread of distribution is given by σ i.e. changing σ increases or decreases the spread



- Note that most real data are NOT normally distributed
  - Some exceptions of normal distributions:
    - I.Q (constructed to be normal, but very slight right skew)
    - Positions of particles in fluid
    - Sum of many random variables (central limit theorem)
  - A normal probability distribution is mostly assumed as an approximation, but this is sometimes disastrous e.g. CDO pricing → Global Financial Crisis
- Normal density function (not examinable):

The formula for the normal **probability density function** is

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2\right]$$

Where $\exp(1) = e$ = mathematical constant ≈ 2.71828

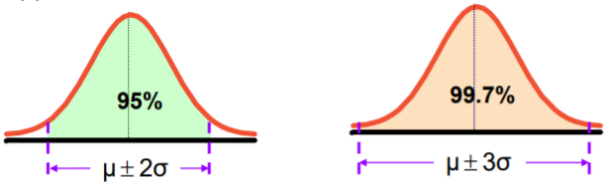
π = mathematical constant ≈ 3.14159

μ = the population mean $E(X)$

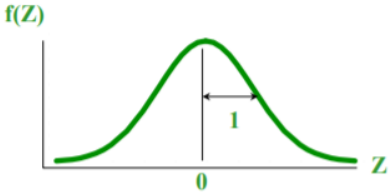
σ = the population standard deviation $\text{Var}(X) = \sigma^2$

X = a value of the continuous variable

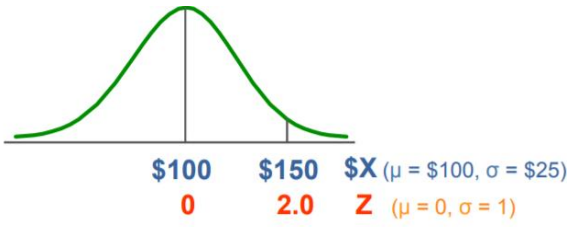
  - If a value of X is further from the mean, f(X) moves closer to zero
- The 'normal' rules → the normal distribution is exactly bell-shaped, so it follows the "empirical rule":
  - Approx. 68% of the data in a normal distribution is within ± one standard deviation of the mean, i.e. $\mu \pm 1\sigma$
  - Approx. 95% of the data in a normal distribution lies within ± two standard deviations of the mean, or $\mu \pm 2\sigma$
  - Approx. 99.7% of the data in a normal distribution lies within $\mu \pm 3\sigma$



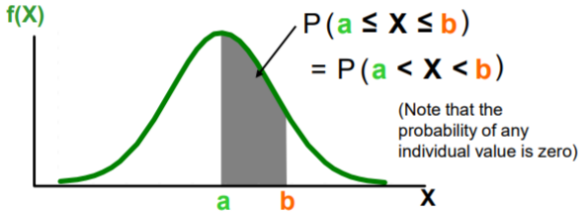
- The standardized normal:
  - Any normal distribution X (with any μ and σ combination) can be transformed into the standardized normal distribution (Z)
  - X units are translated into Z units by subtracting the mean of X and dividing by the standard deviation of X, i.e.: $$Z = \frac{X - \mu}{\sigma}$$
  - The standardized normal distribution (Z) has mean $\mu = 0$ and standard deviation $\sigma = 1$
  - This is also known as the "Z" distribution:



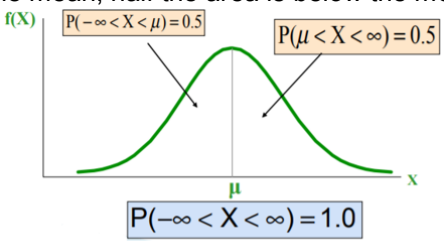
    - X values above the mean have positive Z-values, and X values below the mean have negative Z-values
- Standardized normal distribution e.g. – if X (e.g. weekly sales) is distributed normally with mean of $100 and standard deviation of $25, find the Z value for X = $150:
  - $$Z = \frac{X - \mu}{\sigma} = \frac{\$150 - \$100}{\$25} = 2.0$$
    - This says that X = $150 is two standard deviations (2 increments of $25 units) above the mean of $100
  - Comparing X and X units:



    - Note that the shape of the distribution is the same, only the mean and scale have changed. We can express the problem in the original units (X, e.g. dollars) or in standardized units (Z)
  - Recall that finding the probability for a continuous random variable's outcome in any continuous probability distribution involves measuring the area under the curve (probability density function)

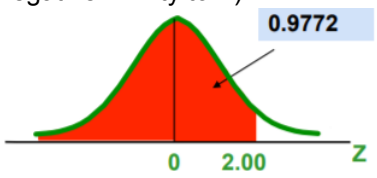


    - Note: P(X < a) is same as P(X ≤ a) because P(X = a) is zero on a continuous probability distribution
  - The total area under the curve is 1.0, and the curve is symmetric, so half the area is above the mean, half the area is below the mean (thus mean = median)



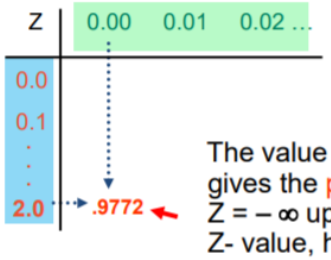
- Standardized normal table:
  - The (Cumulative) Standardized Normal Table will usually give the probability less than a desired value of Z (i.e., from negative infinity to Z):

Example:
P(Z < 2.00) = 0.9772



  - Rules for using Standardized Normal Table:

The column gives the second decimal place for the value of Z



The row shows the integer value and first decimal place of the value of Z

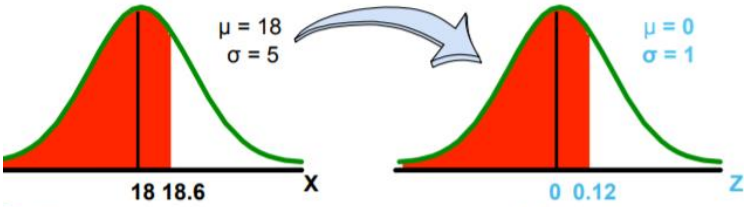
The value within the table gives the probability from Z = – ∞ up to the desired Z- value, here up to Z = 2.00

P(Z < 2.00) = 0.9772

- General procedure for finding normal distribution probabilities:
  - Draw the normal curve for the problem in terms of X
  - Translate X-values to Z-values
  - Use the Standardized Normal Table
- Example of computing probability in a normal distribution:

Let X represent the time taken to download the video
Suppose X is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds. Find P(X < 18.6)

$$Z = \frac{X - \mu}{\sigma} = \frac{18.6 - 18.0}{5.0} = 0.12$$



  - P(X < 18.6)

Standardized Normal Probability Table (Portion)

Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

P(Z < 0.12)

P(X < 18.6)
= P(Z < 0.12)
0.5478

  - P(X > 18.6) would just be 1 – P(X < 18.6) = 0.0.4522

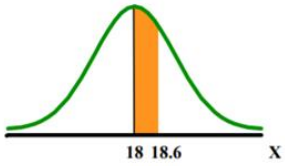
P(X > 18.6) = P(Z > 0.12)
= 1 - P(Z ≤ 0.12)
= 1 - 0.5478 = 0.4522

- Another example with finding probability between two values:
  - Suppose X is normal with mean 18.0 and standard deviation 5.0. Find P(18 < X < 18.6)

Calculate Z-values:

$$Z = \frac{X-\mu}{\sigma} = \frac{18-18}{5} = 0$$

$$Z = \frac{X-\mu}{\sigma} = \frac{18.6-18}{5} = 0.12$$

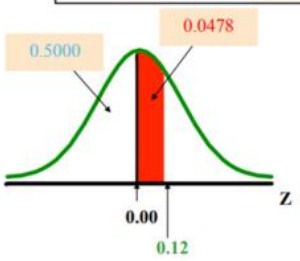


P(18 < X < 18.6)
= P(0 < Z < 0.12)

Standardized Normal Probability Table (Portion)

Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

P(18 < X < 18.6)
= P(0 < Z < 0.12)
= P(Z < 0.12) – P(Z ≤ 0)
= 0.5478 - 0.5000 = 0.0478



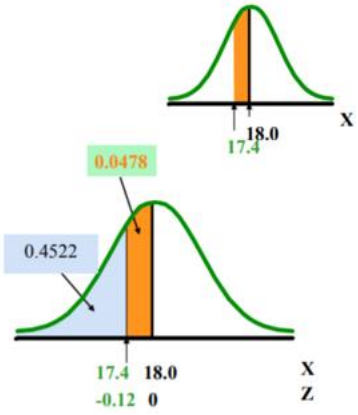
- Another example of finding probability in the lower tail:

Suppose X is normal with mean 18.0 and standard deviation 5.0.

Now Find P(17.4 < X < 18)...

P(17.4 < X < 18)
= P(-0.12 < Z < 0)
= P(Z < 0) – P(Z ≤ -0.12)
= 0.5000 - 0.4522 = 0.0478

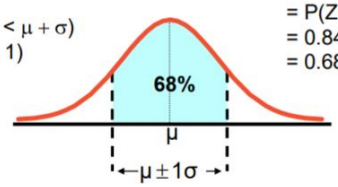
The Normal distribution is symmetric, so this probability is the same as P(0 < Z < 0.12)



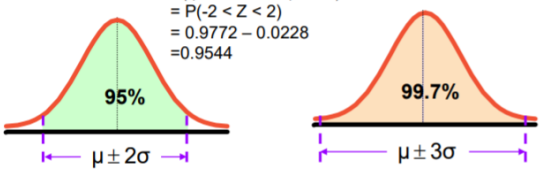
- Proving the empirical rule:
  - Recall: The empirical rule approximates the variation of data in a normal distribution

P(μ – σ < X < μ + σ)
= P(-1 < Z < 1)

P(-1 < Z < 1)
= P(Z<1) – P(Z< -1)
= 0.8413 – 0.1587
= 0.6827

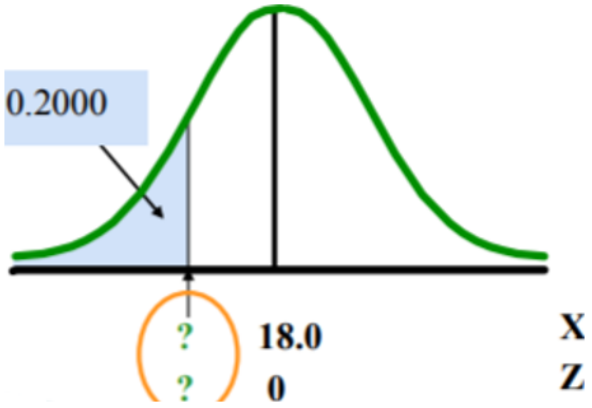


P(μ – 2σ < X < μ + 2σ)
= P(-2 < Z < 2)
= 0.9772 – 0.0228
=0.9544



P(μ – 3σ < X < μ + 3σ) =0.9973

- Working backwards – finding the X value when given a normal probability:
  - 1. Find the Z-value for the known probability
  - 2. Convert to X units using the formula: $X = \mu + Z\sigma$
  - E.g.: Let X represent the time it takes (in seconds) to download a video file from the internet. Suppose X is normal with mean 18.0 and standard deviation 5.0. Find X such that 20% of download times are less than X.

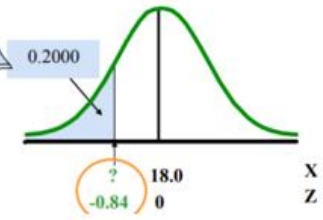


1. Find the Z-value for the known probability

Standardized Normal Probability Table (Portion)

Z	...	.03	.04	.05
-0.9	...	.1762	.1736	.1711
-0.8	...	.2033	.2005	.1977
-0.7	...	.2327	.2296	.2266

- 20% area in the lower tail is consistent with a Z-value of -0.84



2. Convert to X units using the formula:

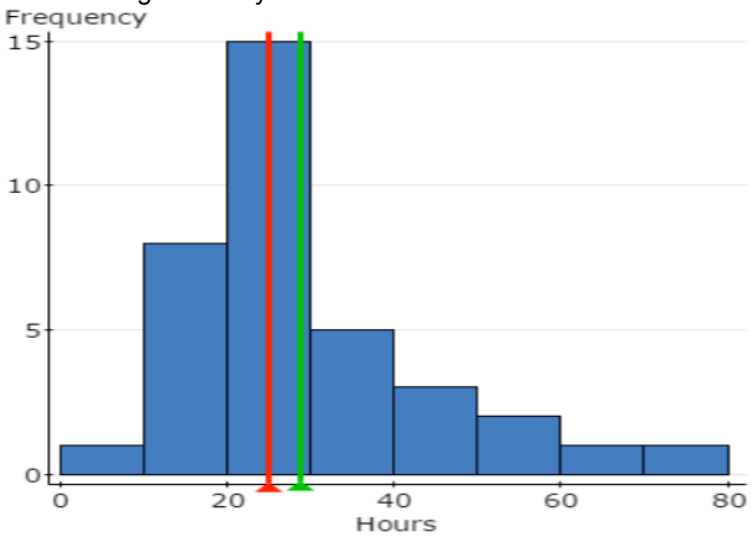
$$X = \mu + Z\sigma$$
$$= 18.0 + (-0.84)5.0$$
$$= 13.8$$

So approximately 20% of the values from a normal distribution with mean 18.0 and standard deviation 5.0 are less than 13.80 seconds

Assessing normality (informally):

- Is the sample mean ≈ sample median?
- Is the empirical rule approximately satisfied? (using $\bar{X}$ and s)
- Is the IQR ≈ 1.33 standard deviations?
- Are the boxplot (smaller n) and histogram (larger n) close to symmetric
- Is the histogram roughly bell-shaped?
- Is there an absence of clear extreme, outlying observations (or absence of "fat tails")?
- Are the sample skewness (between $\pm$ 0.5) and kurtosis (between $\pm$ 1) statistics ≈ 0?

- Example of assessing normality:



Summary statistics:

Column	n	Mean	Std. dev.	Median	Q1	Q3	IQR	Skewness	Kurtosis
Hours	36	28.96	14.90	25	19.75	35.5	15.75	1.525	2.911

Normal rules:

	μ+−σ	μ+−2σ	μ+−3σ
Compensation	75	94.4	97.2
Normal	68.3	95.5	99.7

  - Is the sample mean ≈ sample median?
    - The mean = 28.96 which is above the median of 25 → these do not seem very close to each other, caused by the positive skewness (this is also clear from the histogram)
  - Is the empirical rule approximately satisfied? (using $\bar{X}$ and s)
    - There are too many data points within 1 standard deviation from the mean
    - There are not enough data points within 3 standard deviations from the mean.
    - The data is thus peaked and have fatter tails, as compared to a normal distribution
  - Is the IQR ≈ 1.33 standard deviations?
    - 1.33 * σ = 1.33 * 14.90 = 19.81, which is not too close to the IQR = 15.75
  - Are the boxplot (smaller n) and the histogram (larger n) close to symmetric?
    - Histogram is clearly right-skewed
  - Is the histogram roughly bell-shaped?
    - Histogram is clearly right-skewed
  - Is there an absence of clear extreme, outlying observations (or absence of “fat tails”)?
    - Clearly the data points are right-skewed, though they have no potential outliers and are thus not fat-tailed
  - Are the sample skewness (between $\pm$ 0.5) and the kurtosis (between $\pm$ 1) statistics ≈ 0?
    - The sample skewness is well above 0 at 1.53
    - The sample kurtosis statistic is also high at 2.91, also well above 0
  - In summary:
    - We conclude that the data is NOT normally distributed, with high confidence, since they are not consistent with any of the properties of the normal distribution

Uniform/rectangular probability distribution:

- The continuous uniform distribution has equal density for all possible outcomes of the random variable, where the area under rectangle = 1
- The probability density function is:

$$f(X) = \begin{cases} \frac{1}{b-a} & \text{if } a \le X \le b \\ 0 & \text{otherwise} \end{cases}$$

where
f(X) = value of the density function at any X value
a = minimum possible value of X
b = maximum possible value of X

- Properties of the uniform distribution:
  The mean of a uniform distribution is

$$\mu = \frac{a+b}{2}$$

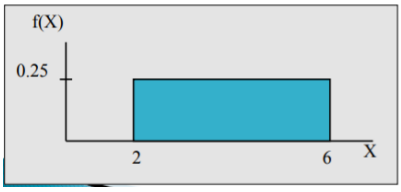
  The standard deviation is

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

- Example – assume that when a bus is late (> 2 minutes past scheduled time), it follows a uniform distribution in arrival times past scheduled time, of between 2 and 6 minutes:

  i.e. $X \sim \text{Uniform, over the range } 2 \le X \le 6$:

$$f(X) = \frac{1}{6-2} = 0.25 \quad \text{for } 2 \le X \le 6$$

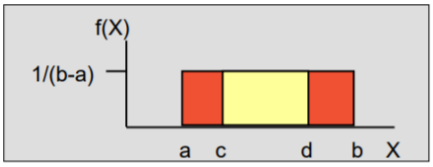


$$\mu = \frac{a+b}{2} = \frac{2+6}{2} = 4$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(6-2)^2}{12}} = 1.1547$$

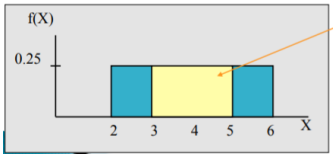
- General procedure for finding normal distribution probabilities:

$$P(c \le X \le d) = (\text{Base})(\text{Height}) = (d-c)(1/(b-a)) = \frac{d-c}{b-a}$$



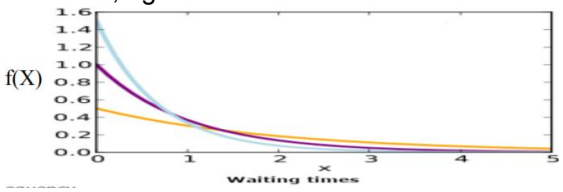
- E.g. continued – use the Uniform (2,6) distribution to find the probability that the bus is between 3 and 5 minutes late, i.e. P(3 ≤ X ≤ 5):

$$P(3 \le X \le 5) = (\text{Base})(\text{Height}) = (5-3)(0.25) = 0.5$$



Exponential probability distribution:

- A positive-valued, right skewed distribution



  - "Waiting time" is often on the x-axis
  - Often used to model the length of time between two occurrences of an event (i.e. time between events) e.g. waiting time to be seated in a restaurant, or time between trucks arriving at an unloading dock
- Related to the Poisson distribution:

  If Y ~ Poisson (λ), counted in a fixed period of time, then the time between each event counted (say X) has an Exponential distribution with mean $1/\lambda$

  - Exponential probability distribution only has a single parameter λ (lambda)
  - Exam note: The mean and standard deviation are both equal to $^1/_\lambda$
- The probability density function is:

$$f(x) = \lambda e^{-\lambda x} \quad ; \quad x > 0$$

  where e = constant, ≈ 2.71828
  $1/\lambda$ = the population mean
  x = any value of the continuous variable, $0 < x < \infty$
- Properties of the exponential distribution:

  Always right or positively skewed

  Mode < median < mean (always)

$$\text{Mean} = \frac{1}{\lambda} = \text{standard deviation}$$

  c.f. with Poisson where mean = variance
- The probability that an exponential r.v .is less than some specified X=x is:

$$P(X < x) = 1 - e^{-\lambda x}$$

  where e = constant, ≈ 2.71828
  $1/\lambda$ = the population mean = population standard deviation
  x = any value of the continuous variable where $0 < x < \infty$

  NB $P(X < x) = \int_{-\infty}^{x} \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$
- E.g. – customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?
  - The mean number of arrivals per hour is 15, so λ = 15
  - Three minutes is 0.05 hours

  $P(\text{arrival time} < .05) = 1 - e^{-\lambda X} = 1 - e^{-(15)(0.05)} = 0.5276$

  So there is a 52.76% chance that the arrival time between successive customers is less than three minutes

  NB the mean arrival time is 1/15 hours (4 mins)

Week 7 – Sampling distributions

Week 7 LO's:

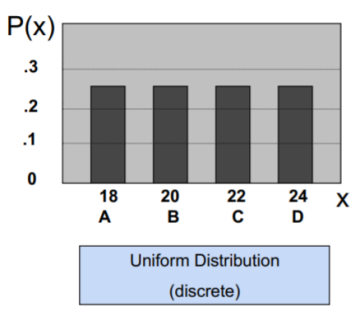
- Sampling Distributions – Chapter 7
- Sampling Distribution of the Mean
  - Standard Error of the Mean
  - For Normal Populations
  - For Non-Normal Populations
- Sampling Distributions of the Proportion
- Sampling Distributions from Finite Populations
- Confidence Intervals – Chapter 8
  - For the Sample Mean, for known Population Variance
  - … continued next week

Sampling distributions:

- Considers the distributions of all possible values of a sample statistic for a given sample size (n) selected from a population
- E.g.:
  - Coca-Cola samples 50 people regarding whether they would buy a new type of cola, and then calculates the sample proportion who would
  - Amazon samples 25 customers and calculates the sample mean of their annual purchase amounts
  - Sanitarium weigh 35 packets of "500g" breakfast cereal and calculates the sample mean weight
  - If Coke/Amazon/Sanitarium repeats this sampling many times, all with n = 50, there will be a different proportion/mean for each sample. These form the sampling distribution for the sample proportion/mean
- In statistics, we are interested in the sampling distribution of the sample proportion/mean, e.g. the distribution of all possible samples of 50 people
  - Coca-Cola estimates that it needs 10% of its customers to buy their new cola to be profitable. They sample 50 people regarding whether they would buy a new type of cola, then calculate the sample proportion who would, which is 4/50. Should they proceed with this new cola?
  - Amazon calculates it needs an annual average purchase amount of $50 to be profitable. They sample 25 customers and calculate a sample mean of $49.50. What should they do?
  - Sanitarium specifies that the weight of a particular cereal packet should be 500g. They weigh 35 packets of cereal and calculate the sample mean weight as 510g. What should they do?
- They each need to decide to consider a sampling distribution to help their decisions

Developing and visualising a sampling means distribution:

- Assume there is a population:
  - Population size N = 4
  - Random variable, X, is the number of business meetings this month
  - Values of X: 18, 20, 22, 24
- Summary measures for the population distribution:
  - $\mu = \frac{\sum X_i}{N} = \frac{18+20+22+24}{4} = 21$
  - $\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$
  - 



- Now considering all possible samples with a size n = 2:

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (SRS sampling with replacement)

16 Sample Means

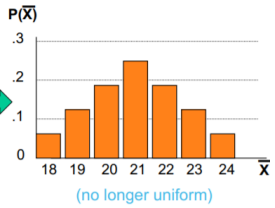
1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

- Sampling distribution of all sample means:

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Distribution of Sample Means



(no longer uniform)

- Summary measures of this sampling distribution:

$$\mu_{\bar{X}} = \frac{18+19+19+\cdots+24}{16} = 21$$

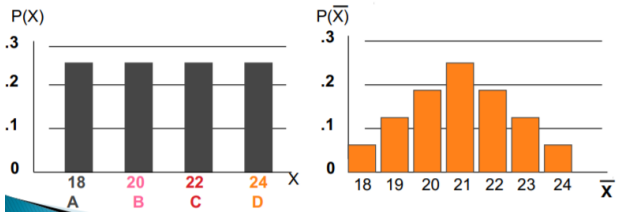
$$\sigma_{\bar{X}} = \sqrt{\frac{(18-21)^2+(19-21)^2+\cdots+(24-21)^2}{16}} = 1.58$$

Note: Here we divide by 16 because there are 16 different samples of size 2.

- Comparing the population distribution to the sample means distribution:

Population
N = 4
$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution
n = 2
$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$



Dividing 2.236 by square root of 2 will give 1.58

Standard error of the mean:

- Different samples of the same size (n) from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample compared to the true population mean is given by the standard error of the mean
  - This assumes that SRS is either with replacement or without replacement from an infinite population

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

    - Note that the standard error of the mean decreases as the sample size (n) increases → we want the sample means distribution standard deviation (standard error of the mean) to be smaller, for any sample mean to be more likely to be closer to the population mean

Sampling distribution of the sample mean (for a normal population):

- General case: for any population with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ has:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

(i.e. Sample mean is an **unbiased** estimator of μ)

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{Var(\bar{X})}$$

Recall means and variances for linear combinations of rvs

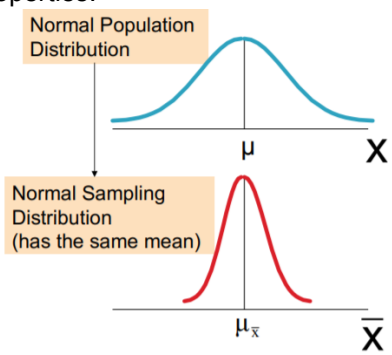
  - "Sample mean is an unbiased estimator of μ" means that the expected value of the sample mean is equal to the population parameter of μ
- BUT: if a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{Var(\bar{X})}$$

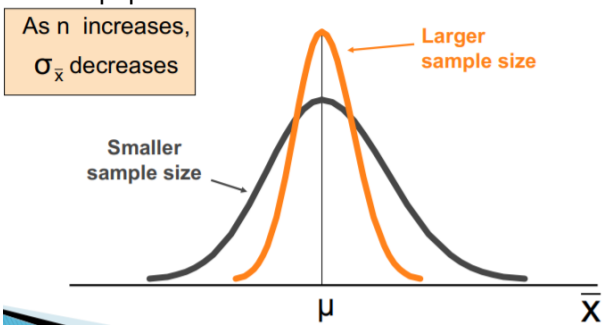
  - In this case, the sampling distribution maintains the same formulas as applied to the general case, but is also normally distributed
- Sample means distribution properties:

$$\mu_{\bar{X}} = \mu$$

(i.e. $\bar{X}$ is an unbiased estimator for μ)



  - Diagrams show that if the population is normal, the sampling means distribution will also be normally distributed
  - Diagram also shows that the mean of the sample means is equal to the population mean



  - As the sample size increases, the standard deviation of sample means decreases, so the probability of each sample mean being the true population mean is higher and standard deviation of each sample will be closer to the standard deviation of the population → samples more representative of population

Z value for sampling distribution of the mean:

- Z-value for the sampling distribution of $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where: $\bar{X}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size

- Exam note: For all questions where we convert the sample means distribution to the Z-distribution, remember to use the correct standard error of $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (rather than just σ) in the Z-value formula

Determining an interval including a fixed percentage of the sample means:

- E.g. – find a symmetrically distributed interval around μ that will include 95% of the sample means when μ = 368, σ = 15, and n = 25 and population is normal:
  - Since the interval contains 95% of the sample means, 5% of the sample means will be outside the interval
  - Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit
  - From the Standardized Normal Table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96

Calculating the lower limit

$$\bar{X}_L = \mu + Z_L \frac{\sigma}{\sqrt{n}} = 368 + (-1.96)\frac{15}{\sqrt{25}} = 362.12$$

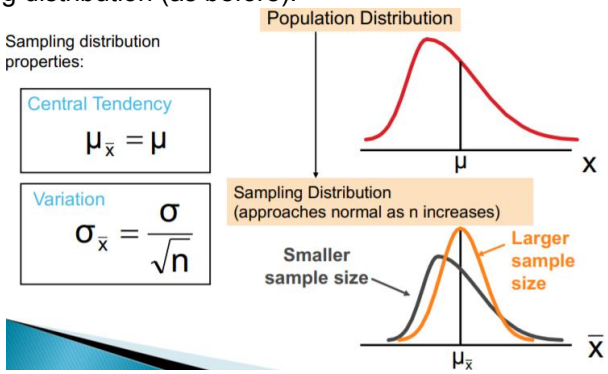
Calculating the upper limit

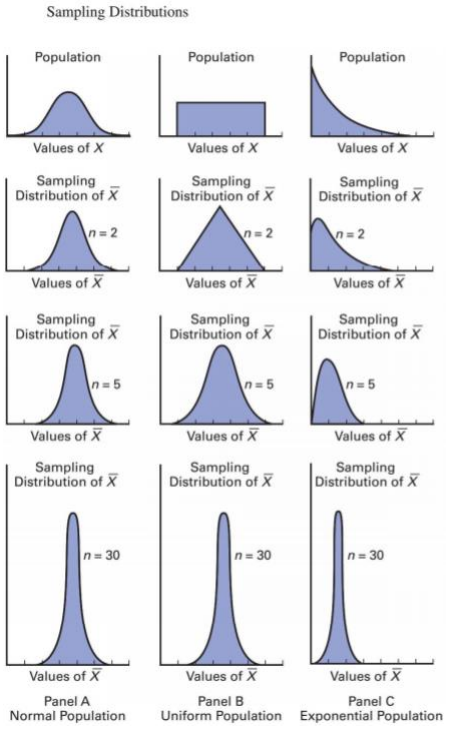
$$\bar{X}_U = \mu + Z_U \frac{\sigma}{\sqrt{n}} = 368 + (1.96)\frac{15}{\sqrt{25}} = 373.88$$

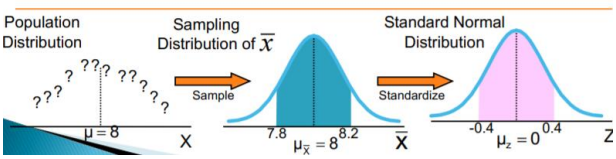
95% of all sample means, will lie between 362.12 and 373.88, when n=25

Sampling distribution of the sample mean (for a non-normal population):

- We may apply the Central Limit Theorem:
  - If the population is not normally distributed, the sample means of random samples from the population will be approximately normally distributed, as long as the sample size n is large enough
  - "Large enough" applies when n ≥ 30
- Assuming the population has a finite mean and finite standard deviation, the properties of the sampling distribution (as before):



- Required sample size for CLT to apply:
  - For most population distributions, n ≥ 30 will give a sampling distribution for the sample mean that is approximately (very close to) normality
  - For fairly symmetric population distributions, n ≥ 15 is enough for approximate normality of the sample means distribution
  - Exam note: For normal population distributions, the sampling distribution of the mean is always, exactly normally distributed e.g. even if n = 1, we would still have a normal sample means distribution
  - 

- Central Limit Theorem summary – for sufficiently large sample sizes (n ≥ 30):
  - 1) The distribution of sample means $\bar{X}$, is approximately normal
  - 2) The mean of this distribution is equal to μ, the population mean
  - 3) Its standard deviation is $\frac{\sigma}{\sqrt{n}}$
  - 4) Regardless of the shape of the original population distribution
- Therefore: the approximation to the normal distribution gets better as n gets larger → the sampling distribution of the sample mean is more and more like a normal distribution the larger the sample size n is.
- E.g. – suppose a population has mean μ = 8 and standard deviation σ = 3. Also suppose that a random sample of size n = 36 is selected. What is the probability that the sample mean is between 7.8 and 8.2?
  - Even if the population is not normally distributed, the sample means distribution is assumed as approximately normal because the central limit theorem can be used (n ≥ 30)
  - So, the sampling distribution of $\bar{X}$ is approximately normal
  - $\mu_{\bar{X}}$ = to μ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
  - $$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8-8}{3/\sqrt{36}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{8.2-8}{3/\sqrt{36}}\right)$$
    $$= P(-0.4 < Z < 0.4) = 0.6554 - 0.3446 = \boxed{0.3108}$$
  - 

- Sanitarium e.g. – Sanitarium specifies the weight of a particular cereal packet as 500g. The standard deviation of weights is 20g. They weigh 50 packets of cereal and calculate the sample mean weight as 510g. Assuming the mean really was 500, what is the chance of getting a sample mean as high as 510 or higher?

$$\mu = 500; \sigma = 20$$

$$\sigma_{\bar{X}} = \frac{20}{\sqrt{50}} = 2.828$$

$$Z = \frac{510 - 500}{2.828} = 3.5355$$

$$P(\bar{X} > 510 \mid \mu = 500) = P(Z > 3.5355)$$
$$= 1 - P(Z < 3.5355)$$
$$= 1 - 0.99979$$
$$= 0.0002$$

  - The chance of selecting a sample with mean weight of 510g or higher is extremely low (so this suggests that the true population average is higher than 500g in reality)

Sampling distributions of the proportion:

Population proportions:

- π = the population proportion
- Sample proportion (p) estimates π:

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

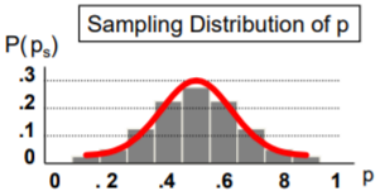
  - 0 ≤ p ≤ 1
- CLT applied to proportions → sample proportions distribution is also approximately distributed as a normal distribution when n is large:
  - "Large enough" applies when $n\pi > 5$ and $n(1 - \pi) > 5$
  - The sample proportions distribution is also assuming SRS sampling with replacement from a finite population, or without replacement from an infinite population
  - Assume that sample is randomly collected from the population
- Sample proportions distribution properties:

$$\mu_p = \pi \qquad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

  - (where π = population proportion)
- CLT – sampling distribution of proportions is approximated by a normal distribution if:

Approximated by a normal distribution if:

$$n\pi \geq 5 \text{ and } n(1-\pi) \geq 5$$



Where $\mu_p = \pi$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

  - (and π = population proportion)

- Z-value for sample proportions:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

- Exam note: For all questions where we convert the sample means distribution to the Z-distribution, remember to use the correct standard error of $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ (rather than just σ) in the Z-value formula
- E.g. – if the true proportion of voters who support a GST increase is π = 0.4, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?

i.e.: **if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \le p \le 0.45)$?**

Find :

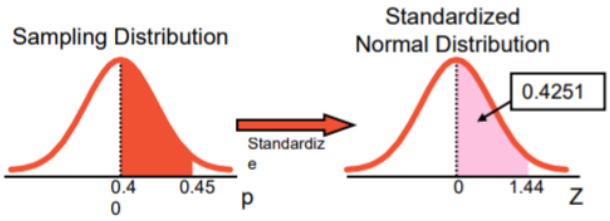
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

Convert to standardized normal:

$$P(0.40 \le p \le 0.45) = P\left(\frac{0.40 - 0.40}{0.03464} \le Z \le \frac{0.45 - 0.40}{0.03464}\right)$$
$$= P(0 \le Z \le 1.44)$$

Utilize the cumulative normal table:

$P(0 \le Z \le 1.44) = 0.9251 - 0.5000 =$ **0.4251**



- Coca-Cola e.g. – Coca-Cola needs 10% of its customers to buy their new cola to be profitable. They sample 50 people regarding whether they would buy a new type of cola, then calculate the sample proportion who would, which is 4/50. Should they proceed? Assuming the true proportion was 0.1, the chance of getting a sample proportion as low as 0.08, or lower, is?

First find σ_p

$$\pi = 0.1; n = 50;$$
$$p = 4/50 = 0.08$$
$$\sigma_p = \frac{\sqrt{0.1(0.9)}}{\sqrt{50}} = 0.0424$$

Second standardize *p* to a Z value.

$$Z = \frac{0.08 - 0.1}{0.0424} = -0.471$$

If the proportion really was 0.1, the chance of getting a sample p as low as 0.08, or lower, is

$$P(p < 0.08 \mid \pi = 0.1) = P(Z < -0.471)$$
$$= 0.3187$$

Therefore, there is a 32% (very large/significant) chance of getting a sample proportion less than 0.08, given that the population proportion is 0.1, so Coca-Cola might not worry too much based on this sample's result

Sampling distributions from finite populations:

- The Finite Population Correction (FPC) factor is:
  - $$\text{fpc} = \sqrt{\frac{N-n}{N-1}}$$
  - We need to apply the FPC when the sample size (n) is equal to or more than 5% of population size (N) i.e. $\frac{n}{N} > 0.05$, and sampling is from a finite population without replacement
- The FPC is used to adjust the standard error of both the sample mean and the sample proportion:

Standard Error of the Mean for Finite Populations

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Standard Error of the Proportion for Finite Populations

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

  - The FPC is always less than 1 so it always reduces the standard error
  - Therefore, this FPC adjustment results in more precise estimates of population parameters
- E.g.: A random sample of size 100 is drawn without replacement, from a population of size 1000 and standard deviation 40

  Here n=100, N=1000 and 100/1000 = 0.10 > 0.05.

  So using the fpc we get:

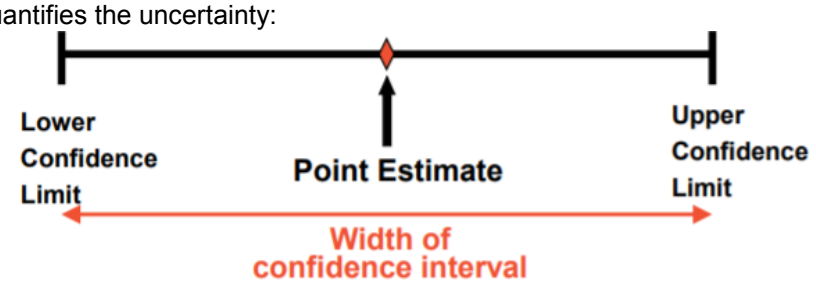
  $$\sigma_{\bar{X}} = \frac{40}{\sqrt{100}}\sqrt{\frac{1000-100}{1000-1}} = 3.8$$

Confidence intervals:

- A point estimate is a single number:

We can estimate a Population Parameter ...		with a Sample Statistic (a Point Estimate)
Mean	μ	$\overline{X}$
Proportion	π	p

- A confidence interval (CI) provides additional information about the variability of the estimate → it quantifies the uncertainty:



- A confidence interval gives a range of values:
  - Based on observations from only 1 sample
  - Takes into consideration variation in sample statistics, i.e. from sample to sample
  - Gives information about possible values of the unknown population parameters
  - Stated in terms of level of confidence e.g. 95% confident
- The general formula for a confidence interval is:

Point Estimate ± (Critical Value)(Standard Error)

Where:

- Point Estimate is the sample statistic estimating the population parameter of interest
- Critical Value is a value based on the sampling distribution of the point estimate and the desired confidence level
- Standard Error is the standard deviation of the point estimate

- The confidence level is our confidence that the interval around our point estimate will contain the unknown population parameter
  - It is a percentage (typically between 80% and 99.9%) → never 100% because we can never be 100% certain a population mean will fall within the confidence interval
  - The confidence level can also be written as $(1 - \alpha)$
    - α = significance level
    - E.g. if the confidence level is 95% then $\alpha = 0.05$
- E.g. – Suppose a population has $\mu = 368$ and $\sigma = 15$:
  - If you take a sample of size n = 25 you know (by CLT)
  - $368 \pm 1.96 * \frac{15}{\sqrt{25}}$ = (362.12, 373.88) contains 95% of the possible sample means when n = 25
  - We will usually not know μ, so we use $\bar{X}$ to estimate μ
    - If $\bar{X} = 362.3$ the 95% interval is $362.3 \pm 1.96 * \frac{15}{\sqrt{25}}$ = (356.42, 368.18)
    - We are 95% confident in this inequality of $356.42 \leq \mu \leq 368.18$, based on this sample
    - Supposing $\mu = 368$, we know that this CI based on $\bar{X} = 362.3$ will contain the true population parameter of μ
  - However, this may not be true about the CI's from other samples of size 25:

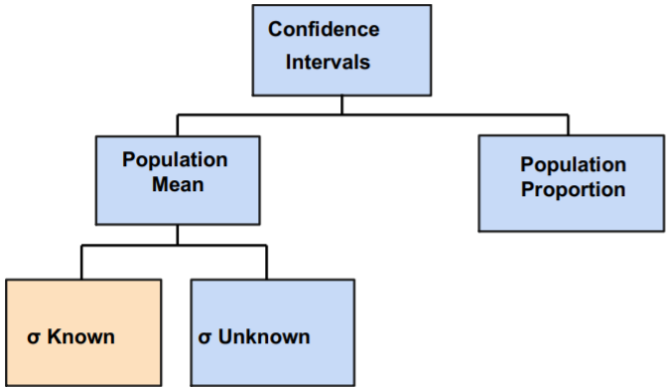
95% intervals

Sample #	$\bar{X}$	Lower Limit	Upper Limit	Contain μ?
1	362.30	356.42	368.18	Yes
2	369.50	363.62	375.38	Yes
3	360.00	354.12	365.88	No
4	362.12	356.24	368.00	Yes
5	373.88	368.00	379.76	Yes

  - In practice, we only take one sample of size n, and in practice we do not know μ, so we do not know if the interval calculated actually contains μ BUT we have a confidence level for how confident we are that our CI contains μ

- CI interpretation:
  - Suppose we determine a 95% confidence interval for a parameter (say μ) → this means that 95% of all the confidence intervals that can be constructed in that manner will contain the true value of μ
  - A relative frequency interpretation → without knowing μ, we can be 95% certain that our confidence interval based on our sample will contain μ
- Exam note: As the sample size increases, the CI decreases
  - A higher n allows us to maintain confidence levels with a smaller confidence interval → we can be just as confident that the population mean is captured by a smaller CI

Confidence interval for the population mean when population variance is known:



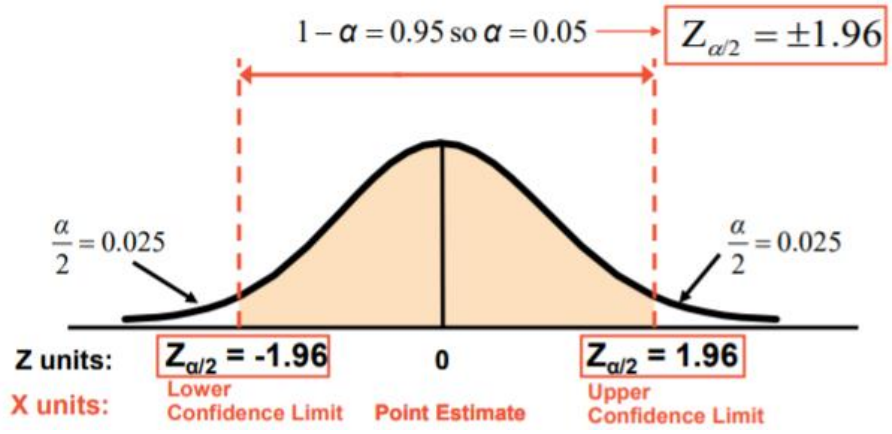
- Assumptions:
  - 1. Population standard deviation (σ) is known
  - 2. Population is normally distributed, or if the population is not normal→ the sample size for our sample means distribution must be sufficiently "large" (n ≥ 30 for CLT)
  - 3. Also assume that all samples are randomly selected (SRS)
- Confidence interval estimate:

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $\overline{X}$ is the point estimate
$Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail
$\sigma/\sqrt{n}$ is the standard error

- Finding the critical value ($Z_{\alpha/2}$):

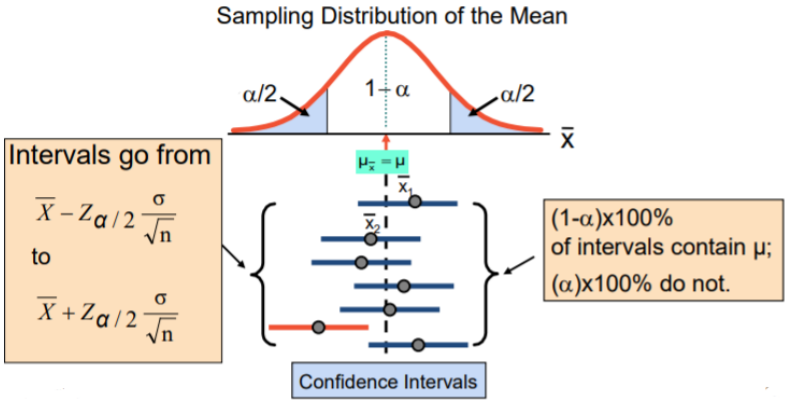
Consider a 95% confidence interval:



- Common confidence levels:
  - Most commonly used confidence levels are 90%, 95%, and 99% (others are OK too!)

Confidence Level	Confidence Coefficient, $1-\alpha$	$Z_{\alpha/2}$ value
80%	0.80	1.28
90%	0.90	1.645
95%	0.95	1.96
98%	0.98	2.33
99%	0.99	2.58
99.8%	0.998	3.08
99.9%	0.999	3.27

- Confidence intervals and confidence levels:



- E.g. – circuits produced in a factory have a target specified resistance of 2 ohms. A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We "know" from past testing that the population standard deviation is 0.35 ohms. Determine a 95% confidence interval for the true mean resistance of the population. Is this sample consistent with the specified target resistance of 2 ohms?
  - We have a normal population, n = 11 circuits, sample mean resistance = 2.20 ohms, and population standard deviation = 0.35 ohms
  - We need to determine a 95% confidence interval for the population mean resistance
  - Solution:

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(0.35/\sqrt{11})$$

$$= 2.20 \pm 0.2068$$

$$1.9932 \le \mu \le 2.4068$$

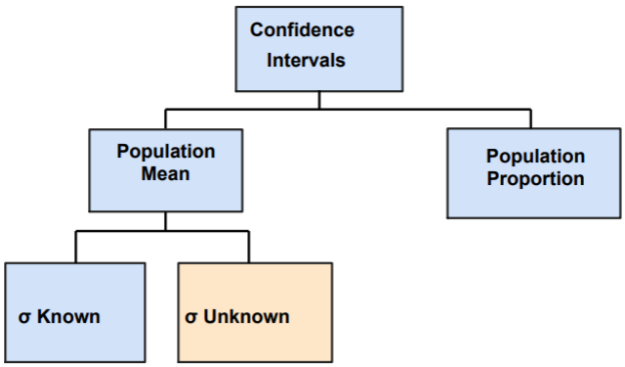
  - Interpretation:
    - Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean
    - Thus, we are 95% confident that the true mean resistance is between 1.9932 ohms and 2.4068 ohms
  - Conclusion:
    - The sample is consistent with the specified target (as 2 is contained in this CI), but 2 is close to the lower bound of the CI so we should take another mean from a larger sample to check our results

Week 8 – Confidence intervals:

Week 8 LO's:

- Confidence Intervals – Chapter 8
  - For mean, Population Variance known (last week)
  - For mean, Population Variance unknown
  - For proportion
  - Sample Size determination
- Hypothesis Testing – One Sample – Chapter 9

Confidence interval for the population mean when population variance is unknown:

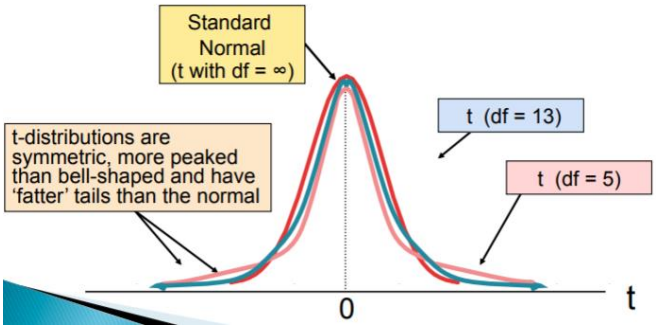


- For a vast majority of time in real world business situations, we do not truly know the population standard deviation → sometimes 'yes' (a very good approximation)
- If there is a situation where σ is known, then μ also may be known (since to calculate σ, you need to know μ).
- If we truly know μ, there would be no need to gather a (further) sample to estimate it

- So, if the population standard deviation (σ) is unknown, we substitute the sample standard deviation, S
  - This introduces extra uncertainty, since S varies from sample to sample
  - To properly account for that, we use the Student-t distribution instead of the normal distribution
- Student's t distribution:
  - If X is normal, then (from prior weeks), we have a (standardized) normal distribution with mean 0 and variance 1, where $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
  - However, if we use S to estimate σ, then we have a different, specific (standardized) "Student-t" distribution, where $$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
    - These $t_{\alpha/2}$ values depend on the number of degrees of freedom (d.f)
- Degrees of freedom (DF)
  - d.f. = n - 1
  - This represents the number of observations that are free to vary after the sample mean has been calculated
  - E.g. – suppose the mean of 3 numbers is 8.0. Let X_1 = 7, X_2 = 8, What is X_3?
    - X_3 must be 9 (i.e. X_3 is not free to vary). Here, n = 3, so the DF are 3 – 1 = 2 (2 values can be any numbers, but the third is not free to vary for a given mean, when n = 3)

- Student's t distribution:

Note: t ⟶ Z as n increases

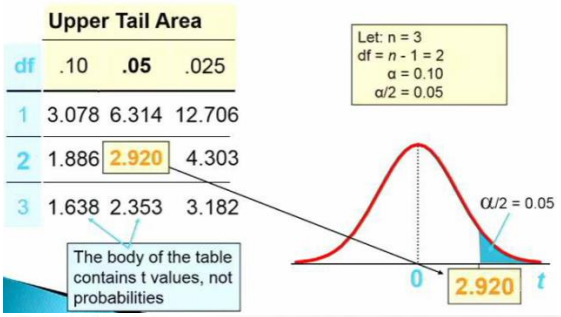


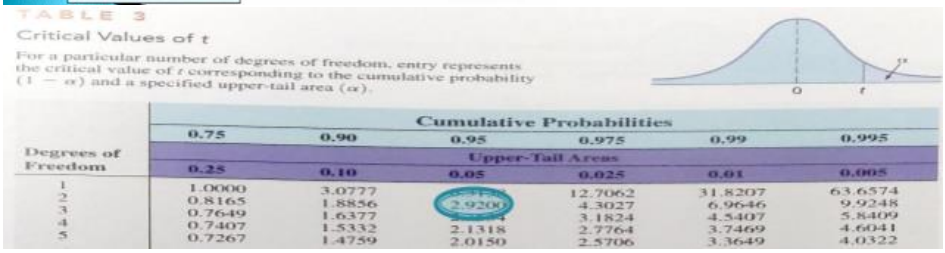
  - "More peaked than bell-shaped" → there is more probability of values being further away from the mean → the t distribution allows for more uncertainty and more extreme values
  - As the DF increase, we get closer to a normal distribution i.e. if DF = ∞, the t distribution is the standard normal distribution
  - For low values of degrees of freedom, there is substantially more probability of a value in the tails (as S is a less accurate estimate of σ in this case)
- Assumptions:
  - 1. Population standard deviation (σ) is unknown
  - 2. Population is normally distributed, or if the population is not normal→ the sample size for our sample means distribution must be sufficiently "large" (n ≥ 30 for CLT)
  - 3. Also assume that all samples are randomly selected (SRS)
  - How I think about the distributions:
    - With assumption 2, the population distribution and sample means distribution are the same, regardless of whether σ is known
    - The only difference is that if σ is unknown, we transition from the sample means distribution into a t distribution (instead of a Z distribution)
- Confidence interval estimate:

$$\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with n -1 "degrees of freedom" and an area of $\alpha/2$ in each tail)

- Using a Student's t table:





  - Make sure to use the relevant DF and α values when finding the t value

- Comparison of t values to z values:
Cumulative Probability	t (10 d.f.)	t (20 d.f.)	t (50 d.f.)	t (100 d.f.)	Z (∞ d.f.)
0.9	1.3722	1.3253	1.2987	1.2901	1.2816
0.95	1.8125	1.7247	1.6759	1.6602	1.6449
0.975	2.2281	2.0860	2.0086	1.9840	1.9600
0.995	3.1693	2.8453	2.6778	2.6259	2.5758

    - In general, t values are larger than z values as there is more uncertainty, so it is more likely for more extreme values which lie further out into the tails
    - t values converge to z values as n (and thus degrees of freedom) increases because as the sample size grows, our estimate of S becomes closer to the true σ
- Example of a confidence interval in a t distribution:
  - A random sample of n = 25 has $\bar{X}$ = 50 and S = 8. Form a 95% confidence interval for μ
  - Degrees of freedom = 24 and $t_{\alpha/2}$ = $t_{0.025}$ = 2.0639

    The 95% confidence interval is

    $$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

    $46.698 \le \mu \le 53.302$
- Recall – assumptions for a valid Student-t distribution confidence interval interpretation:
  - Interpreting the t-distribution-based interval requires that the population is normal OR that CLT holds i.e. n is large enough … (as before)
  - Normality of the population can be assessed by:
    - Histogram and/or boxplot
    - Assessing whether: mean ≈ median, IQR ≈ 1.33S, skewness ≈ 0, kurtosis ≈ 0, empirical rule holds, etc.
- E.g.: Sanitarium specifies the weight of a particular cereal packet as 500g. They weigh 50 packets of cereal and calculate the sample mean weight as 510g, the sample standard deviation of weights is S = 21g. What should they do?
  - We're given the sample standard deviation (don't know population standard deviation), so we should use a t distribution
  - n = 50 is a large sample, so the sample mean distribution should be approximately normal because CLT applies (don't need to assume normality)
  - We need to determine a 99% confidence interval for the population mean weight
  - Solution:
    - $\bar{X} = 510; s = 21$
    - $\sigma_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{21}{\sqrt{50}} = 2.9698$
    - $df = n - 1 = 49$; $(1-\alpha) = 0.99 \rightarrow \alpha/2 = 0.005$

      $t_{49;0.995} = 2.6800$ Table value = T.INV.2T(0.01,49) =T.INV(0.005,49) In Excel

      The confidence interval is

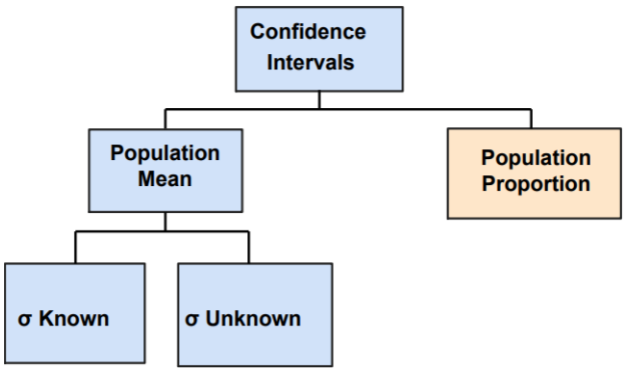
      $\bar{X} - 2.68 * 2.9698, \bar{X} + 2.68 * 2.9698$

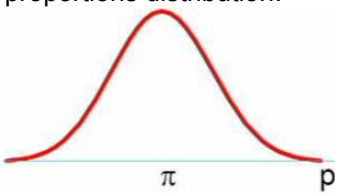
      $= (510 - 7.959, 510 + 7.959)$

      $= (502.04, 517.96)$

- Interpretation:
  - Although the true mean may or may not be in this interval, 99% of intervals formed in this manner will contain the true mean
  - Thus, we are 99% confident that the true mean weight is between 502.04g and 517.98g
- Conclusion:
  - 500g is not contained in this interval, so Sanitarium should decrease cereal packet weights as these results suggest that they are producing more than 500g on average

Confidence interval for the population proportion (π):



- Sample proportions distribution:
  - 

  - An interval estimate for the population proportion (π) can be calculated via the estimate of the sampling distribution of the sample proportion (p)
- Recall that the distribution of the sample proportions is approximately normal, if:
  - $n\pi \geq 5 \quad and \quad n(1-\pi) \geq 5$ with standard deviation $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
    - We will estimate this standard error with $\sqrt{\frac{p(1-p)}{n}}$
  - Exam note: No Student-t type adjustment needed for the sample proportions distribution → can always use a Z statistic for our confidence intervals
- Confidence interval estimate:

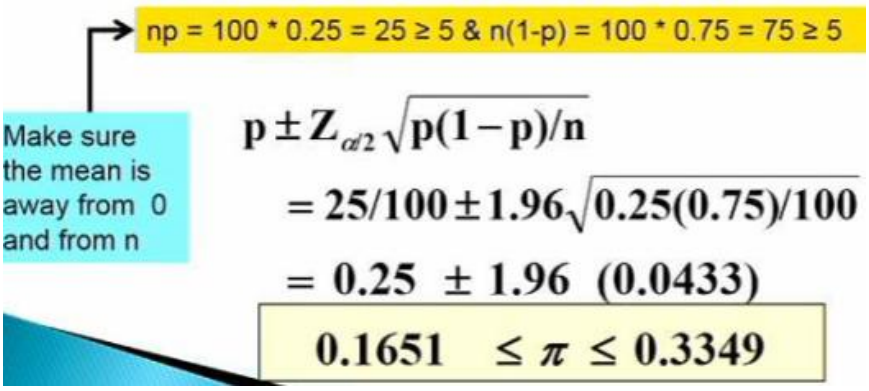
$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where

- $Z_{\alpha/2}$ is the standard normal value for the level of confidence desired
- p is the sample proportion
- n is the sample size
- π is the population proportion

  - Exam note: To be normal, $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Since we don't know π, we estimate these with p, i.e. $np \geq 5$ and $n(1-p) \geq 5$

- E.g.: A random sample of 100 employees shows that 25 have adaptable height office desks. Form a 95% CI for the true proportion of employees with these desks i.e. use $Z_{\alpha/2}$ = 1.96



  - We are 95% confident that the true percentage of employees with adaptable desks in the population is between 16.51% and 33.49%.
  - Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion π, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion of employees with adaptable height office desks
- E.g.: Coca-Cola needs at least 10% of its customers to buy their new cola to be profitable. They sample 50 people regarding whether they would buy a new type of cola, then calculate the sample proportion who would, which is 4/50. Should they proceed?
  - We need to determine a 95% confidence interval for the population proportion of customers willing to buy the new cola
  - Solution:
    - $n = 50;\ p = 4/50 = 0.08;\ \ np = 4 < 5!!$
    - $\sigma_p = \frac{\sqrt{0.08(0.92)}}{\sqrt{50}} = 0.0384$
    - Note that np < 5 so we are less confident about the accuracy of this interval, as the sample proportions distribution becomes less approximately normal in this case

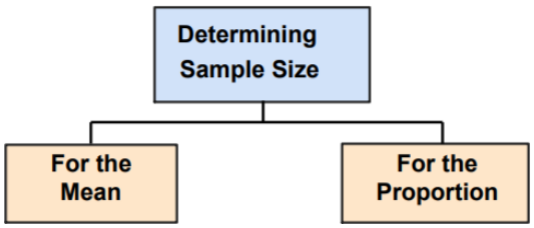
      95% confidence gives a = 0.05, a/2 = 0.025

      $$Z_{0.975} = 1.96$$

      The confidence interval is:

      $$p - Z^*\hat{\sigma}_p, p + Z^*\hat{\sigma}_p$$
      $$= (0.08 - 1.96*0.0384, 0.08 + 1.96*0.0384)$$
      $$= (0.08 - 0.075, 0.08 + 0.075)$$
      $$= (0.005, 0.155)^*$$
  - Interpretation:
    - Although the true proportion may or may not be in this interval, 95% of intervals formed in this manner will contain the true proportion
    - Thus, we are 95% confident that the true proportion of customers is between 5% and 15.5%
  - Conclusion:
    - 0.1 is well-contained within this interval, so Coca-Cola can proceed, but it should be cautious and use more data (especially since np = 4 < 5)

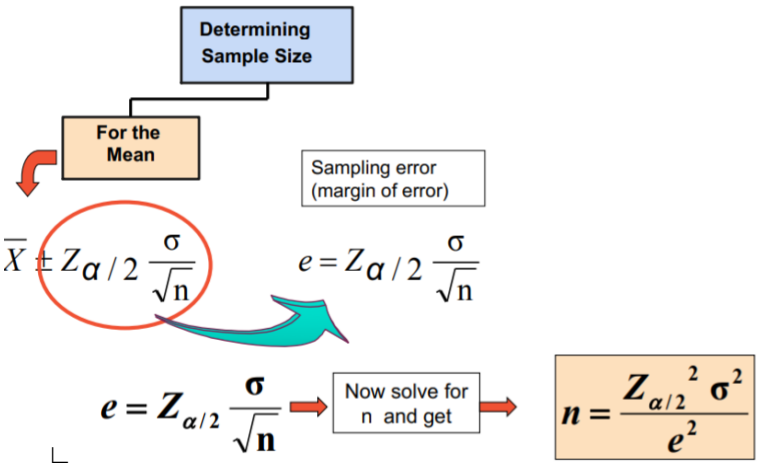
Sample size determination:

- 

Sample error:

- The required sample size can be found that obtains a desired margin of error (e), with a specified level of confidence (1 - α)
- The margin of error e is also called the sampling error:
  - The amount of imprecision in the estimate of the population parameter
  - The amount added and subtracted to the point estimate to form the confidence interval

Determining sample size for the population mean CI when population variance is known:

- 

- To determine the required sample size for the mean, we must know:
  - The desired level of confidence (1 – α), which determines the critical value, $Z_{\alpha/2}$
  - The acceptable sampling error, e
  - The standard deviation, σ
- Required sample size e.g.: If σ = 45, what sample size is needed to estimate the population mean to within ± 5 with 90% confidence?

$$n = \frac{Z_{\alpha/2}^2\,\sigma^2}{e^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19$$

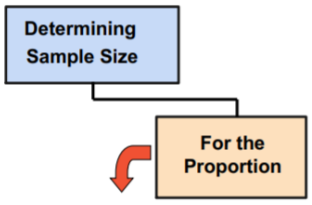
So the required sample size is n = 220

(Always round up)

  - Exam note: Always round the sample size up as this way guarantees that the sample is sufficient to meet the margin of error at the desired level of confidence
- If σ is unknown → σ can be estimated with S when determining the required sample size:
  - Use a value for σ that is expected to be at least as large as the true σ
  - Select a pilot sample and estimate σ with the sample standard deviation, S
  - The error in terms of the Student-t can be solved numerically if desired, i.e.
    - $e = t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}$ n appears twice here!

39

Determining sample size for the population proportion CI:

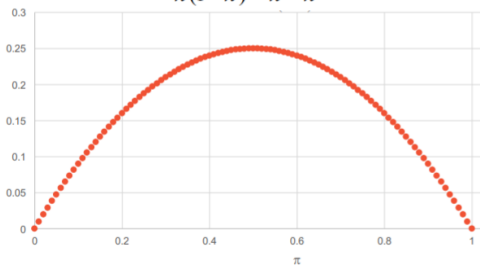


- $e = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$ → Now solve for n and get → $n = \frac{(Z_{\alpha/2})^2 \pi(1-\pi)}{e^2}$
- To determine the required sample size for the proportion, we must know:
  - The desired level of confidence (1 – α), which determines the critical value, $Z_{\alpha/2}$
  - The acceptable sampling error, e
  - The true proportion of events of interest, π
    - If we don't know π → π can be estimated (by p) with a pilot sample, or we can conservatively use π = 0.5

$$n = \frac{(Z_{\alpha/2})^2 \pi(1-\pi)}{e^2}$$

Why is π = 0.5 "conservative" here?

$$\pi(1-\pi) = \pi - \pi^2$$



  - π = 0.5 is 'conservative as π *(1- π) is at its maximum when π = 0.5 → this then gives largest possible value of n, our required sample size
- Required sample size e.g.: How large a sample would be necessary to estimate the true proportion of defectives in a population of light globes accurate to within ± 3%, with 95% confidence? (Assume a pilot sample yields p = 0.12).

For 95% confidence, use $Z_{\alpha/2}$ = 1.96

e = 0.03

p = 0.12, so use this to estimate π

$$n = \frac{Z_{\alpha/2}^{\ 2} \pi(1-\pi)}{e^2} = \frac{(1.96)^2(0.12)(1-0.12)}{(0.03)^2} = 450.74$$

So use n = 451

  - Note that e = 0.03 here since the question specifies 3%
- E.g. continued: How large a sample would be necessary if the true proportion were p = 0.01 or p = 0.5?

For 95% confidence, use $Z_{\alpha/2}$ = 1.96

$$n = \frac{Z_{\alpha/2}^{\ 2} \pi(1-\pi)}{e^2} = \frac{(1.96)^2(0.01)(1-0.01)}{(0.03)^2} = 42.26$$

So use n = 43

$$n = \frac{(1.96)^2(0.5)(1-0.5)}{(0.03)^2} = 1067.07$$

So use n = 1068

i.e. conservative

Application to auditing:

- Many auditors make extensive use of probability sampling and confidence intervals.
- The population of company "accounts" is often too large or too time consuming or too costly to keep up with → dealing with a probability sample of accounts is more feasible
- E.g. – an auditor has a population of 1000 vouchers and wants to estimate the mean and total value of that population. A sample of 50 vouchers is taken that has a mean of $1076.39 and a S = $273.62. A 95% confidence interval for the mean voucher amount is required:
  - The FPC needs to be applied with N = 1000 because the sample size is equal to 5% of the population, and sampling is from a finite population without replacement
  - A 95% confidence interval for the mean voucher amount is:
    - $$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 1076.39 \pm (2.0096) \frac{273.62}{\sqrt{50}} \sqrt{\frac{950}{999}}$$
    - = ($1000.56, $1152.22)
  - Thus, we are 95% confident that the population mean voucher value is between $1000.56 and $1152.22. We are also 95% confident that the population total voucher value is between 1000*$1000.56 and 1000*$1152.22 i.e. ($1000559, $1152221)

Ethical issues:

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate
  - The level of confidence should always be reported
  - The sample size should be reported
  - An interpretation of the confidence interval estimate should also be provided
- Otherwise, another measure to accurately quantify the level of uncertainty should be included

Introduction to hypothesis testing:

- Hypothesis testing is at the heart of all scientific enquiry
- The basis of the modern scientific method is that a theory should lead to questions or claims or predictions that can be tested
- The most common way to test such hypotheses is via empirical methods i.e. data

What is a hypothesis:

- A hypothesis is a claim, often about a population parameter

Population mean:	Population proportion:
Example: The mean monthly mobile phone bill is μ = $42	Example: Telstra's market share proportion in mobile phone customers, π, is greater than 0.5
Example: Amazon's mean annual customer purchase amount is $50	Example: At least 10% of Coca Cola's customers will purchase their new cola brand
Example: A cereal packet's mean weight is 500 grams	Example: Less than 50% of voters will have their vote count toward a particular political party (after preferences)

The null hypothesis (H_0):

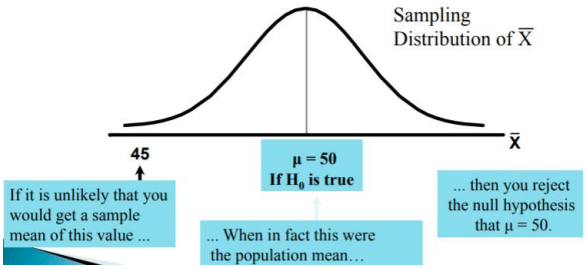
- States a default or status quo claim or assertion
  - E.g.: The average diameter of a manufactured bolt is 30mm (H_0: μ = 30)
- Is always about a population parameter, not about a sample statistic
  - $H_0 : \mu = 30$ $H_0 : \bar{X} = 30$
- Tests usually begin by assuming a "null" hypothesis is true (similar to the notion of innocent until "proven" guilty)
  - Can refer to the "status quo" or historical value or just a relevant value to the test
  - May or may not be rejected in the test.
  - Cannot be proven by the test (but can be disproved)

The alternative hypothesis (H_1):

- Opposes the null hypothesis in some way
  - E.g.1: The average diameter of a manufactured bolt is not equal to 30mm (H_1: $\mu \neq 30$)
  - E.g.2: The average diameter is less than 30mm (H_1: $\mu < 30$)
  - E.g.3: The average diameter is greater than 30mm (H_1: $\mu > 30$)
- Challenges the "status quo"
- Is generally the hypothesis that the researcher is trying to find evidence for (or against)
- Often formed first → we usually start with t

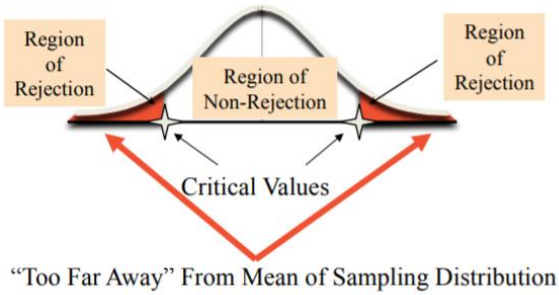
Hypothesis testing process example:

- Amazon wants a mean of at least $50 for annual customer purchase amount. Is the mean $50?
  - Claim – the population mean amount is 50:
    - H_0: $\mu = 50$
    - H_1: $\mu \neq 50$
  - To test this, we sample the population and find the sample mean
    - Suppose the sample mean amount was $\overline{X} = 45$.
    - This is lower than the claimed mean population age of 50
    - If the null hypothesis were true, the probability of getting such a different sample mean might be small, so then you would reject the null hypothesis
    - In other words, if getting a sample mean of 45 is very unlikely, when the population mean is 50, you conclude that the population mean is very likely not 50.



- The test statistic and critical values:
  - If the sample mean is "close" enough to the stated population mean, the null hypothesis is not rejected
  - If the sample mean is "far" enough from the stated population mean, the null hypothesis is rejected
  - How "far" is far enough to reject H_0?
    - The critical value of a test statistic creates a "line in the sand" for decision making → it helps to answer the question of how far is far enough

Sampling Distribution of the test statistic



"Too Far Away" From Mean of Sampling Distribution

  - This is for when "not equal to" is the alternative
    - This is called a two-tailed test because there is a rejection region in both tails

Possible errors in hypothesis test decision making:

- Type I error:
  - Rejecting a null hypothesis, given that it is true
  - Probability of Type I Error is α = P(reject null | null true)
  - Called "level of significance" or "size" of the test
  - Set by researcher in advance
- Type II error:
  - Failure to reject a null hypothesis, given that it is false
  - Probability of Type II Error is β = P(not reject null | null false)
  - Determined by the test, sample size, etc. and us|ally not able to be known precisely
- **Possible Hypothesis Test Outcomes** (study this table)

Decision	Actual Situation: H_0 True	Actual Situation: H_0 False
Do Not Reject H_0	Correct decision Probability 1 – α	Type II Error Probability β
Reject H_0	Type I Error Probability α	Correct decision Probability 1 - β

  - α and β are conditional probabilities
  - The confidence coefficient $(1 - \alpha)$ is the probability of not rejecting H_0 when it is true
  - The confidence level of a hypothesis test is $(1 - \alpha)$*100%
  - The power of a statistical test $(1 - \beta)$ is the probability of rejecting H_0 when it is false
  - P(reject H0 | H0 true) = a : the size of a test
    P(reject H0 | H0 false) = 1 - b : the power of a test
- Type I and II relationship:
  - Type I and Type II errors cannot happen at the same time (since they are conditioned upon different outcomes)
    - A Type I error can only occur given H_0 is true
    - A Type II error can only occur given H_0 is false
  - When the Type I error probability (α) increases, Type II error probability (β) decreases
  - We cannot reduce the probability of both Type I and II errors unless we use a bigger sample size
- Factors affecting type II error:
  - All else equal,
    - β ↑ when the difference between hypothesized parameter and its true value ↓
    - β ↑ when α ↓
    - β ↑ when σ ↑
    - β ↑ when n ↓

(study these relationships)

Week 9 – Hypothesis testing (one sample tests)

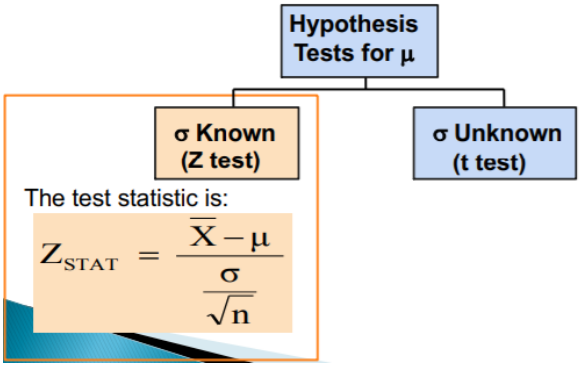
Week 9 LO's:

- Introduction to Hypothesis Testing (last week) – Chapter 9
- Hypothesis tests for the mean – One Sample Tests
  - σ Known, Two-Tail Tests
    - Critical value approach
    - p-value approach
    - Comparison to confidence intervals
  - σ unknown, Two-Tail Tests
    - Critical value and p-value approaches
    - Comparison to confidence intervals
  - One-Tail Tests
    - Critical value approach
    - p-value approach
- Hypothesis tests for the Proportion – One Sample Tests
  - Critical value approach
  - p-value approach
- Two Sample Tests - Chapter 10

DCOVA

Hypothesis testing for the mean when σ is known (two-tail test):

- Use Z tests
- Convert sample statistic ($\bar{X}$) to a Z_{STAT} test statistic

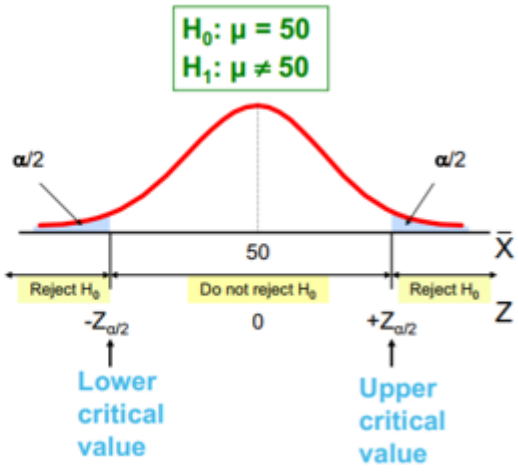


- Assumptions:
  - 1. Population standard deviation (σ) is known
  - 2. Population is normally distributed, or if the population is not normal→ the sample size for our sample means distribution must be sufficiently "large" ($n \geq 30$ for CLT)
  - 3. Also assume that all samples are randomly selected (SRS)

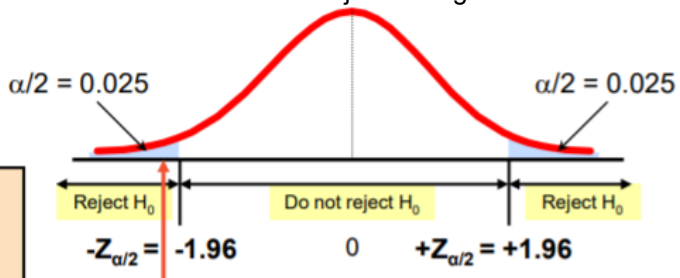
Critical value approach to testing:

- For a two-tail test:

There are two cutoff values (critical values), defining the regions of rejection



- 6 steps in critical value hypothesis testing:
  - 1. State the null and alternative hypotheses, H_0 and H_1
  - 2. Choose level of significance (α) and sample size (n)
  - 3. Determine appropriate test statistic and sampling distribution i.e. appropriate technique
  - 4. Determine critical values and identify rejection and non-rejection regions
    - This is determined by our chosen level of significance (α) by using the Standardized (use table or computer)
  - 5. Collect data and compute test statistic value
    - Determine the test statistic by converting (i.e. standardizing) the sample statistic ($\bar{X}$) to a test statistic (Z_{STAT})
  - 6. Make the statistical decision and state the managerial conclusion
    - If the test statistic falls into the non-rejection region → do not reject (NOT 'accept') the null hypothesis H_0
    - If the test statistic falls into the rejection region: reject the null hypothesis
    - Express the managerial conclusion in the context of the business problem
- E.g. – test the claim that the true mean diameter of a manufactured bolt is 30mm. (Given σ = 0.8):
  - 1. State the null and alternative hypotheses, H_0 and H_1
    - H_0: μ = 30
    - H_1: $\mu \neq 30$ → this is a two-tail test
  - 2. Specify the desired level of significance and the sample size
    - Suppose α = 0.05 and n = 100
  - 3. Determine the appropriate technique
    - σ is known so this is a Z test
  - 4. Determine the critical values
    - For α = 0.05, the critical Z values (Z_{CRIT}) are $\pm$ 1.96
  - 5. Collect the data and compute the test statistic
    - Suppose the sample results are n = 100, $\bar{X}$ = 29.84 (σ = 0.8 is known)
    - The test statistic would be $Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{0.08} = -2.0$
  - 6. Make the statistical decision and state the managerial conclusion
    - We need to determine if the test statistic is in the rejection region

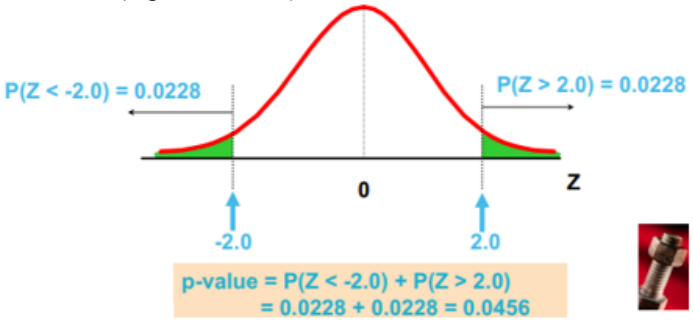


Reject H_0 if $Z_{STAT} < -1.96$ or $Z_{STAT} > 1.96$; otherwise do not reject H_0

Here, Z_{STAT} = -2.0 < -1.96, so the test statistic is in the rejection region

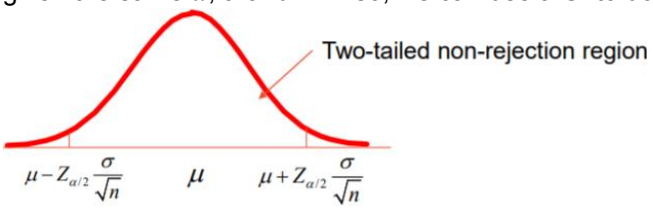
- Since Z_{STAT} = -2.0 < -1.96, reject the null hypothesis and conclude there is sufficient evidence, at the 5% significance level, that the mean diameter of the manufactured bolts is not equal to 30

P-value approach to testing:

- p-value – probability of obtaining a value for the test statistic equal to or more extreme (as defined by the alternative hypothesis) than the observed sample value, given H_0 is true
  - The p-value is a conditional probability (conditional upon null hypothesis being true)
  - The p-value is also called the observed level of significance
  - Decision rule:
    - If p-value < α, reject H_0
    - If p-value ≥ α, do not reject H_0
- 5 steps in p-value hypothesis testing:
  - 1. State the null and alternative hypotheses, H_0 and H_1
  - 2. Choose level of significance (α) and sample size (n)
  - 3. Determine appropriate test statistic and sampling distribution i.e. appropriate technique
  - 4. Collect data and compute the test statistic value and the associated p-value
  - 5. Make the statistical decision and state the managerial conclusion:
    - **If the p-value is < α then reject H_0, otherwise do not reject H_0.**
    - State the managerial conclusion in the context of the business problem
- E.g. – test the claim that the true mean diameter of a manufactured bolt is 30mm. (Given σ = 0.8):
  - 1. State the null and alternative hypotheses, H_0 and H_1
    - H_0: μ = 30
    - H_1: $\mu \neq 30$ → this is a two-tail test
  - 2. Specify the desired level of significance and the sample size
    - Suppose α = 0.05 and n = 100
  - 3. Determine the appropriate technique
    - σ is known so this is a Z test
  - 4a. Collect the data and compute the test statistic
    - Suppose the sample results are n = 100, $\bar{X}$ = 29.84 (σ = 0.8 is known)
    - The test statistic would be $Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{0.08} = -2.0$
  - 4b. Calculate the p-value
    - We then calculate the p-value → in general, the p-value is the probability of observing a test statistic more extreme than that observed, in the direction of the alternative hypothesis, given that the null hypothesis is true
    - So, in a two-tailed test, we sum up both upper and lower tail probabilities
    - We are looking at how likely is it to get a Z_{STAT} of $\pm$ 2 or further (from the mean of 0), given that H_0 is true
    - 

  - 5. Make the statistical decision and state the managerial conclusion
    - p-value = 0.0456 < α = 0.05
    - Reject H_0 at the 5% significance level.
    - There is sufficient evidence to conclude the average diameter of a manufactured bolt is statistically significantly different to 30mm at the 5% significance level.

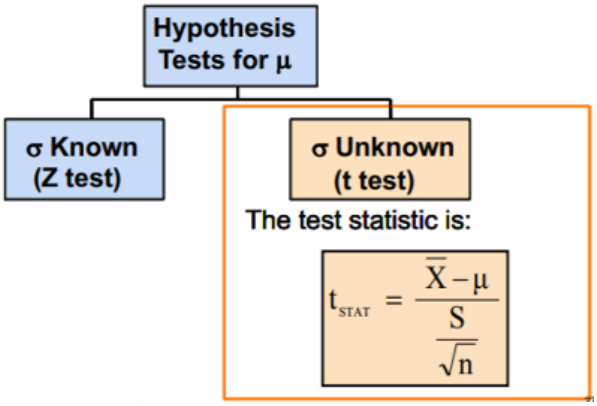
Connection between two-tail tests and confidence intervals:

- For $\bar{X}$ = 29.84, σ = 0.8 and n = 100, the 95% CI is:
  - $29.84 - (1.96)\frac{0.8}{\sqrt{100}}$ to $29.84 + (1.96)\frac{0.8}{\sqrt{100}}$ = $29.6832 \le \mu \le 29.9968$
  - Since this interval does not contain the hypothesized mean (30), we reject the null hypothesis at α = 0.05
  - If the hypothesized value is not in confidence interval → reject the null hypothesis
- These two procedures always give the same conclusion, because the width of the non-rejection region is equal to the width of the associated confidence interval, using the same α
  - Therefore, a two-sided test and confidence interval will give the same answer when given the same α, σ and n → so, we can use a CI to double-check our answer
  - 

    - Do not reject H0 if $\bar{X}$ is in the interval: $\mu \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (two-tailed test)
    - Is equivalent to
      Do not reject H0 if μ is in the interval: $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (confidence interval)
    - i.e., both methods ask the same question:
      Is $|\bar{X} - \mu| < Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$?

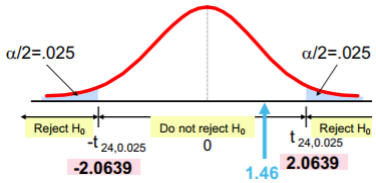
Hypothesis testing for the mean when σ is unknown (two-tail test):

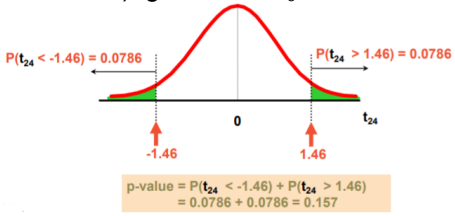
Convert sample statistic ($\bar{X}$) to a t_{STAT} test statistic



- If the population standard deviation is unknown, we instead use the sample standard deviation (S)
- Thus, we use the Student-t distribution, instead of the Z distribution, to test the null hypothesis about the mean
- Assumptions:
  - 1. Population standard deviation (σ) is unknown
  - 2. Population is normally distributed, or if the population is not normal→ the sample size for our sample means distribution must be sufficiently "large" (n ≥ 30 for CLT)
  - 3. Also assume that all samples are randomly selected (SRS)
- All other steps, concepts, and conclusions are the same
  - Exam note: The only difference is that since σ is unknown, we transition from the sample means distribution into a t distribution (instead of a Z distribution)

Critical value and p-value approach to testing example:

- E.g.: The average cost of a hotel room in New York is said to be $168 per night. To determine if this is accurate, a random sample of 25 hotels is taken and resulted in $\bar{X}$ = $172.50 and S = $15.40. Test the appropriate hypotheses at α = 0.05
  - 1. State the null and alternative hypotheses, H_0 and H_1
    - H_0: μ = 168
    - H_1: $\mu \neq 168$ → this is a two-tail test
  - 2. Specify the desired level of significance and the sample size (to find DF)
    - Suppose α = 0.05 and n = 25
    - DF = n – 1 = 25 – 1 = 24
  - 3. Determine the appropriate technique
    - σ is unknown so this is a t test
- Critical value approach:
  - 4. Determine the critical values
    - For α = 0.05 and n = 25, the critical t values are $\pm$ 2.0639
    - $t_{CRIT} = t_{24,0.025} = \pm$ 2.0639
  - 5. Collect the data and compute the test statistic
    - Sample results are n = 25, $\bar{X}$ = 172.5 (σ is unknown so we use S = 15.4)
    - The test statistic would be $t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$
    - 

  - 6. Make the statistical decision and state the managerial conclusion
    - t_{STAT} = 1.46 < t_{CRIT} = 2.0639
    - Do not reject H_0 at the 5% significance level
    - There is insufficient evidence to conclude that the average hotel room price in New York is statistically significantly different than $168, at the 5% significance level
- P-value approach:
  - 4a. Collect the data and compute the test statistic
    - Sample results are n = 25, $\bar{X}$ = 172.5 (σ is unknown so we use S = 15.4)
    - The test statistic would be $t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$
  - 4b. Calculate the p-value
    - We are looking at how likely is it to get a t_{STAT} of $\pm$1.46 or further (from the mean of 0), given that H_0 is true
    - 

  - 5. Make the statistical decision and state the managerial conclusion
    - p-value = 0.157 > α = 0.05
    - Do not reject H_0 at the 5% significance level
    - There is insufficient evidence to conclude that the average hotel room price in New York is statistically significantly different than $168, at the 5% significance level
- Note we should make an extra assumption here that the population is normally distributed since n is only 25

Connection between two-tail tests and confidence intervals:

- For $\bar{X}$ = 172.5, S = 15.40 and n = 25, the 95% confidence interval for μ is:

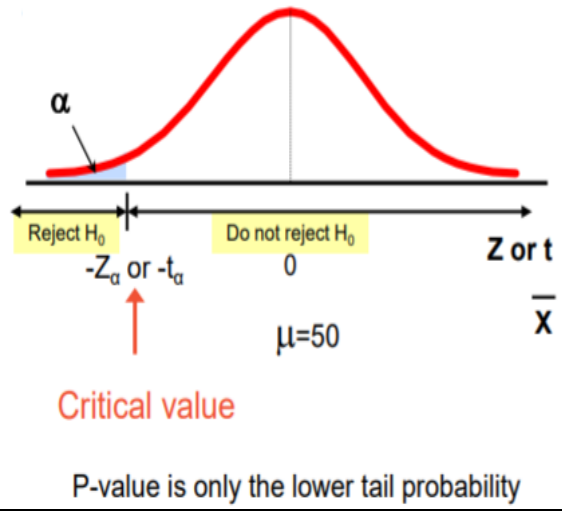
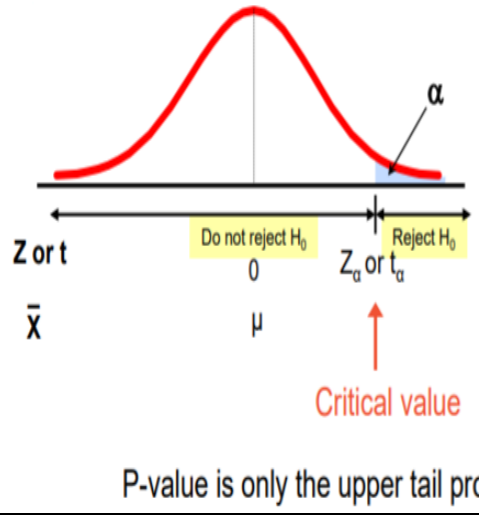
$$172.5 - (2.0639)\,15.4/\sqrt{25} \quad \text{to} \quad 172.5 + (2.0639)\,15.4/\sqrt{25}$$

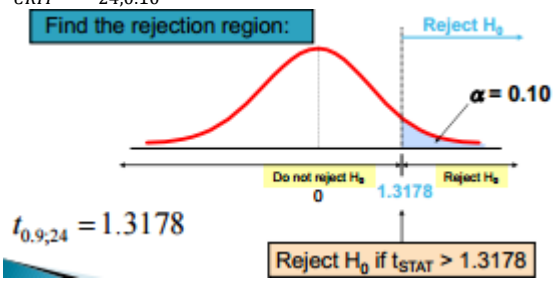
$$166.14 \le \mu \le 178.86$$

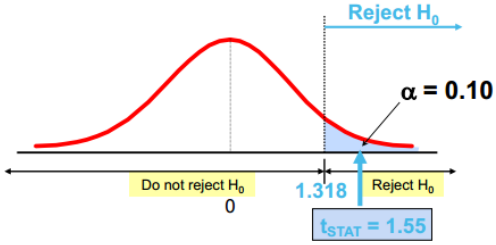
  - Since this interval contains the hypothesized mean (168), we do not reject the null hypothesis at $\alpha = 0.05$
- Recall that a two-tailed test and a confidence interval around the sample mean will always give the same answer

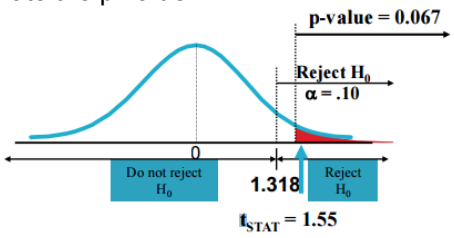
One-tail tests:

- In many cases, the alternative hypothesis focuses on a particular direction e.g.:
  - Amazon want an average purchase amount above $50 per customer i.e. $\mu > 50$
  - Coca Cola want at least 10% of their customers to purchase a new cola i.e. $\pi > 0.1$
  - ALP want the LNP vote to be less than 50% i.e. $\pi < 0.5$
- For one-tail tests, there is only one critical value, since the rejection area is in only one tail
- Exam note: Be careful to use the correct α in the Standardized Normal Table (z table) or t table when it is a one tail test, when given a certain α
  - No longer using $\alpha/2$, like we did in the two-tail tests

Lower-tail test:	Upper-tail test:
$H_0: \mu = 50$ (OR $\mu \ge 50$) $H_1: \mu < 50$	$H_0: \mu = 50$ OR $\mu \le 50$ $H_1: \mu > 50$
This is a lower-tail test since the alternative hypothesis is focused on the lower tail **below** the value of 0.5	This is an upper-tail test since the alternative hypothesis is focused on the upper tail **above** the mean of 50
 P-value is only the lower tail probability	 P-value is only the upper tail probability

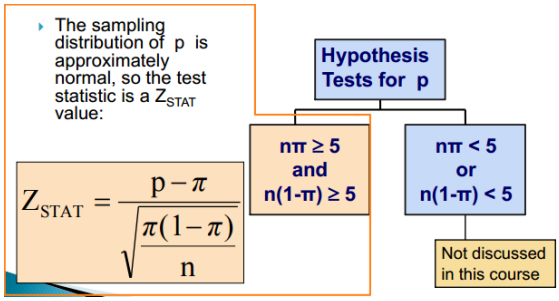
- E.g.: Amazon wants more than $50 purchase amount per customer. The company wishes to test this claim. σ is unknown.
  - 1. State the null and alternative hypotheses, H_0 and H_1
    - H_0: μ = 50 the average is equal to or less than $50
      H_1: μ > 50 the average is greater than $50
  - 2. Specify the desired level of significance and the sample size
    - Suppose α = 0.10 and n = 25
  - 3. Determine the appropriate technique
    - σ is unknown so this is a t test
  - 4. Determine the critical values
    - For α = 0.10 and n = 25, the upper critical t value is 1.3178
    - $t_{CRIT} = t_{24,0.10} = 1.3178$
    - 

  - 5. Collect the data and compute the test statistic
    - Suppose now a sample is taken with the following results of n = 25, $\bar{X}$ = 53.1, and S = 10
    - The test statistic would be $t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{53.1 - 50}{\frac{10}{\sqrt{25}}} = 1.55$
  - 6. Make the statistical decision and state the managerial conclusion:
    - t_{STAT} = 1.55 > t_{CRIT} = 1.3178
    - Reject H_0 at the 5% significance level
    - There is sufficient evidence to conclude that the true mean amount is statistically significantly greater than $50, at the 5% significance level
    - 

- E.g. continued – using the p-value for the test:
  - 4. Calculate the p-value
    - 

  - 5. Make the statistical decision and state the managerial conclusion:
    - **Reject H_0 since p-value = 0.067 < α = .10**
    - We conclude that there is sufficient evidence to support the claim that the mean purchase amount is statistically significantly above $50, at the 10% level
- Note we should make an extra assumption here that the population is normally distributed since n is only 25

Hypothesis tests for the proportion – one sample tests:



- Involves categorical variables e.g. vote yes or no in recent survey
- Two possible outcomes:
  - Possesses characteristic of interest (1)
  - Does not possess characteristic of interest (0)
- Fraction or proportion of the population in the category of interest is denoted by π
- Sample proportion in the category of interest is denoted by p
  - $p = \frac{X}{n} = \frac{\text{number in category of interest in sample}}{\text{sample size}}$
- Z test for proportion in terms of number in category of interest:

$$Z_{STAT} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$ Simply divide through by n

  - Both Z_{STAT}'s are equivalent forms
  - The top formula is expressed in the terms of the number in the category of interest (X)
  - We can convert between the two equations using X = np
- CLT applied to proportions → when both $n\pi$ and $n(1-\pi)$ are at least 5, p can be approximated by a normal distribution with mean and standard deviation:

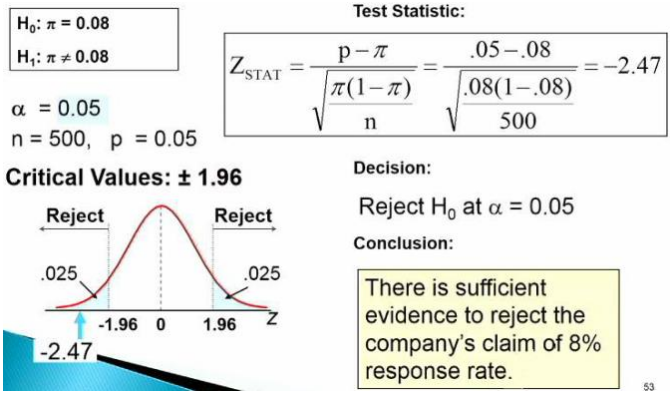
$$\mu_p = \pi \qquad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- E.g.: A marketing company claims that it receives responses from 8% of those surveyed. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the α = 0.05 significance level.

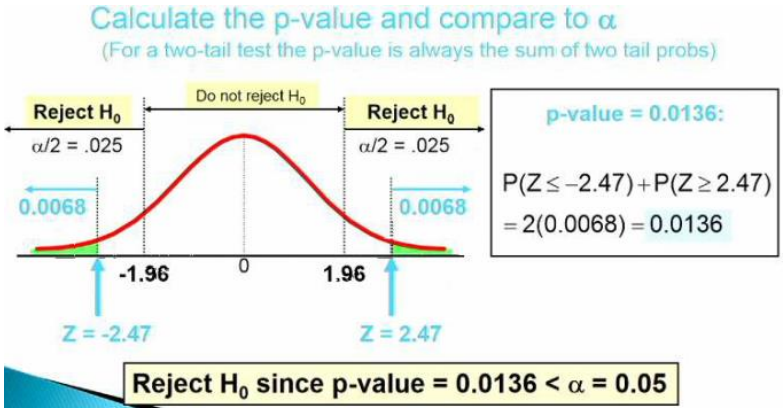
Check: ✓
nπ = 500*0.08 = 40
and
n(1-π) = 460

NB use the π specified in the null hypothesis

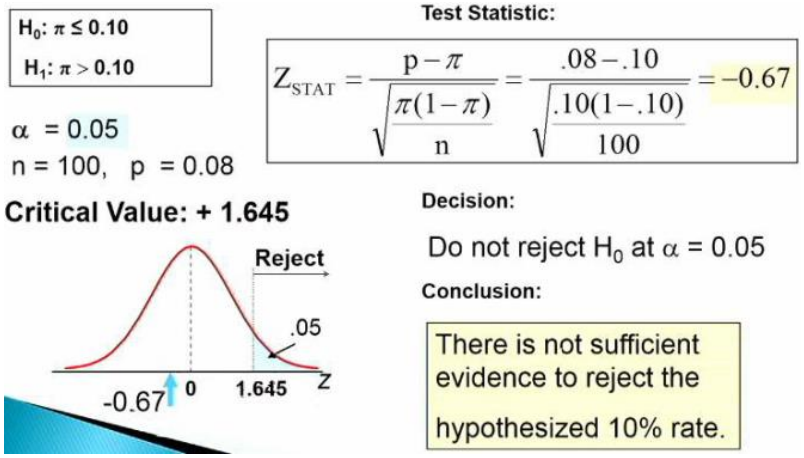
  - Critical value solution:



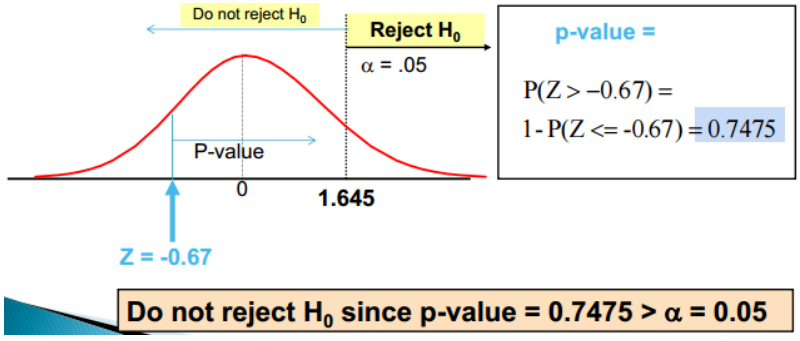
- P-value solution:



- Another e.g.: Coca Cola wants at least 10% of customers to purchase a new cola. To test this claim, a random sample of 100 were surveyed with 8 positive responses. Test at the α = 0.05 significance level.
  - Critical value solution:



  - P-value solution:



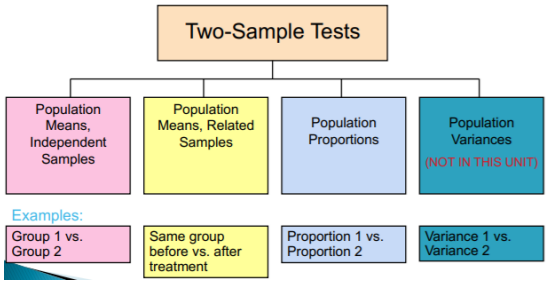
Potential pitfalls and ethical considerations:

- Use randomly collected data (probability samples) to reduce selection biases and non-sampling error and allow the sampling distribution theory to be used!
- Choose the level of significance, α, and the type of test (one-tail or two-tail) before data collection
- Do not employ "data snooping" to choose between one-tail and two-tail tests, or to determine α (switching from two-tail test to one-tail test can make a value significant when it previously wasn't)
- Do not practice "data cleansing" (deleting data set) to hide observations that do not support a stated hypothesis
- Report all pertinent findings including both statistical significance and practical importance

Week 10 – Hypothesis testing (two sample tests):

Week 10 LO's:

- Difference in Mean - Independent Samples
  - Pooled Variance T-Test using critical values, p-values, and confidence intervals
  - Separate Variance T-Test using critical values, p-values, and confidence intervals
- Mean Difference in related (paired) samples
  - Paired T-Test using critical values, p-values, and confidence intervals
- Difference in Population Proportions
  - Z-test using critical values and p-values
  - Confidence intervals
- Covariance and Correlation

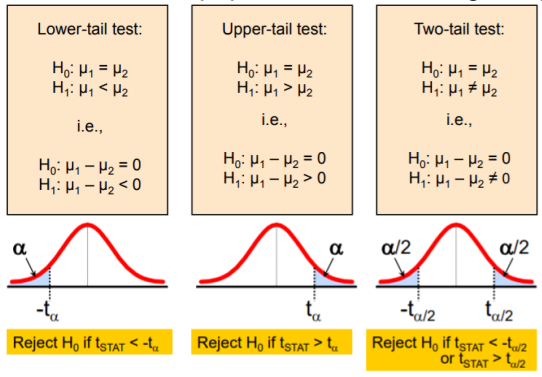


Difference between two population means:

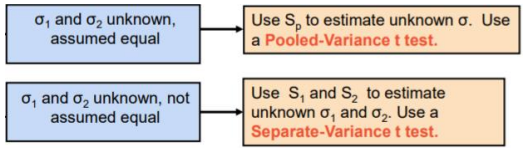
- Goal: Test a hypothesis or form a confidence interval for the difference between two population means, $\mu_1 - \mu_2$
  - The point estimate for the difference is $\bar{X}_1 - \bar{X}_2$
- The CLT says: A linear combinations of independent random variables are approximately normally distributed, as the sample size gets larger
  - Thus: CLT applies here, when the two groups are independent and randomly sampled, as long as both groups are either normally distributed or has large enough sample size

Difference between two population means – independent samples:

- Different data sources:
  - Unrelated
  - Independent → sample selected from one population has no effect on the sample from the other population
- Hypothesis tests for two population means using independent samples:



- 2 cases to consider:



Hypothesis tests for $\mu_1 - \mu_2$, with σ_1 and σ_2 unknown and assumed equal (case 1 – pooled variance):

- Assumptions:
  - 1. Samples are independently and randomly drawn
  - 2. Populations are normally distributed or both sample sizes are at least 30 (or at least 15 for populations that are symmetric)
  - 3. Population variances are unknown but assumed equal
- The pooled variance is:
  - $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 - 1 + n_2 - 1}$$
  - This is derived from:
    - $$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{n-2}$$ (not examinable)
- The test statistic is:
  - $$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
  - This is derived from:

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + (-1)^2 V(\bar{X}_2) + 2 \times 1 \times (-1) \times Cov(\bar{X}_1, \bar{X}_2)$$
$$= \frac{\sigma_p^2}{n_1} + \frac{\sigma_p^2}{n_2} - 2 \times 0 = \sigma_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

    - Assuming a common variance
  - The t_{STAT} DF:
    - $n_1 + n_2 - 2$

Pooled-variance t test example (case 1):

- You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	21	25
Sample mean	3.27%	2.53%
Sample std dev	1.30	1.16

  - Assume equal population variance because sample standard deviations are fairly close
  - Exam note: Both samples are below n = 30 but above n = 15 so we need to assume population distributions are both normal or (at least) symmetrically distributed for CLT to hold

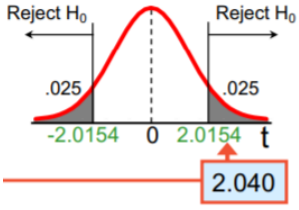
Assuming both populations have equal variances, is there a difference in mean yield ($\alpha = 0.05$)?

- Solution:
  - 1. State the null and alternative hypotheses
    - H_0: $\mu_1 - \mu_2 = 0$ (i.e. $\mu_1 = \mu_2$)
    - H_1: $\mu_1 - \mu_2 \neq 0$ (i.e. $\mu_1 \neq \mu_2$) → this is a two-tail test
  - 2. Specify the desired level of significance and the sample size
    - α = 0.05, n_1 = 21, and n_2 = 25
  - 3. Determine the appropriate technique
    - σ is unknown so this is a t test
    - Population variances are assumed equal, so this is a pooled-variance t test
  - 4. Determine the critical values
    - For α = 0.05 and DF = 21 +25 – 2 = 44, the critical t values (t_{CRIT}) are ± 2.0154
  - 5. Collect the data and compute the test statistic

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3.27 - 2.53) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = 2.040$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(21-1)1.30^2 + (25-1)1.16^2}{(21-1) + (25-1)} = 1.5021$$

  - 6. Make the statistical decision and state the managerial conclusion
    - t_{STAT} = 2.040 > t_{CRIT} = 2.0154
    - Reject H_0 at the 5% significance level
    - There is sufficient evidence to conclude that there is a statistically significant difference in mean dividend yields between the NYSE and NASDAQ, at the 5% significance level



- Using Excel:

Pooled-Variance t Test for the Difference Between Two Means
(assumes equal population variances)

	A	B	
4	Hypothesized Difference	0	
5	Level of Significance	0.05	
6	**Population 1 Sample**		
7	Sample Size	21	=COUNT(DATA!$A:$A)
8	Sample Mean	3.27	=AVERAGE(DATA!$A:$A)
9	Sample Standard Deviation	1.3	=STDEV.S(DATA!$A:$A)
10	**Population 2 Sample**		
11	Sample Size	25	=COUNT(DATA!$B:$B)
12	Sample Mean	2.53	=AVERAGE(DATA!$B:$B)
13	Sample Standard Deviation	1.16	=STDEV.S(DATA!$B:$B)
14			
15	**Intermediate Calculations**		
16	Population 1 Sample Degrees of Freedom	20	=B7 - 1
17	Population 2 Sample Degrees of Freedom	24	=B11 - 1
18	Total Degrees of Freedom	44	=B16 + B17
19	Pooled Variance	1.502	=((B16 * B9^2) + (B17 * B13^2)) / B18
20	Standard Error	0.363	=SQRT(B19 * (1/B7 + 1/B11))
21	Difference in Sample Means	0.74	=B8 - B12
22	t Test Statistic	2.040	=(B21 - B4) / B20
23			
24	**Two-Tail Test**		
25	Lower Critical Value	-2.015	=-T.INV.2T(B5, B18)
26	Upper Critical Value	2.015	=T.INV.2T(B5, B18)
27	p-value	0.047	=T.DIST.2T(ABS(B22),B18)
28	Reject the null hypothesis		=IF(B27<B5,"Reject the null hypothesis", "Do not reject the null hypothesis")

p-val = 0.047 < 0.05

Confidence interval for $\mu_1 - \mu_2$, with σ_1 and σ_2 unknown and assumed equal (case 1):

- The confidence interval for $\mu_1 - \mu_2$ is:
  - $$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$
  - Exam note: Remember that $t_{\alpha/2}$ has DF = $n_1 + n_2 - 2$
- E.g. continued: Since we rejected H_0 can we be 95% confident that $\mu_{NYSE} \neq \mu_{NASDAQ}$?
  - 95% confidence interval for $\mu_{NYSE} - \mu_{NASDAQ}$:
    - $$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 0.74 \pm 2.0154 \times 0.3628 = (0.009,\ 1.471)$$
  - Since 0 is less than the entire interval, we can be 95% confident that $\mu_{NYSE} \neq \mu_{NASDAQ}$

Hypothesis tests for $\mu_1 - \mu_2$, with σ_1 and σ_2 unknown and not assumed equal (case 2 – separate variance):

- Assumptions:
  - 1. Samples are independently and randomly drawn
  - 2. Populations are normally distributed or both sample sizes are at least 30 (or at least 15 for populations that are symmetric)
  - 3. Population variances are unknown and not assumed equal
- The test statistic is:
  - $$t_{STAT} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$
    - We use the sample standard deviations as estimates of the separate population variances
  - This is derived from:

$$E\left(\bar{X}_1 - \bar{X}_2\right) = E\left(\bar{X}_1\right) - E\left(\bar{X}_2\right) = \mu_1 - \mu_2$$

$$V\left(\bar{X}_1 - \bar{X}_2\right) = V\left(\bar{X}_1\right) + (-1)^2 V\left(\bar{X}_2\right) + 2 \times 1 \times (-1) \times Cov\left(\bar{X}_1, \bar{X}_2\right)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2 \times 0$$

$$\approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

    - **Not** assuming a common variance
  - The t_{STAT} DF:
    - $$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Separate variance t test example 1 (case 2):

- You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	21	25
Sample mean	3.27%	2.53%
Sample std dev	1.30	1.16

  - Now suppose that we chose to assume unequal population variance, using the same data set as before
  - Both samples are below n = 30 but above n = 15 so we need to assume population distributions are both normal or (at least) symmetrically distributed for CLT to hold

Is there a difference in mean yield ($\alpha = 0.05$)? $H_0 : \mu_{Nasdaq} = \mu_{NYSE}$

$H_1 : \mu_{Nasdaq} \neq \mu_{NYSE}$

- Sample results:

Hypothesis test results:

μ_1 : Mean of Population 1
μ_2 : Mean of Population 2
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(without pooled variances)

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	0.74	0.3664699	40.574393	2.0192654	0.0501

- Decision:
  - Do not reject H_0 (just!) at the 5% significance level
- Conclusion:
  - There is insufficient evidence to conclude that there is a statistically significant difference in mean dividend yields between the NYSE and NASDAQ at the 5% level
- Comparison between pooled variance test and separate variance test:

Equal variance test

$$t_{STAT} = \frac{3.27 - 2.53}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \frac{0.74}{0.3628} = 2.040 \quad \text{with 44 degrees of freedom}$$

$$p - val = 2 \times P(t_{44} > 2.040) = 0.047$$

Separate variance test

$$t_{STAT} = \frac{3.27 - 2.53}{\sqrt{\left(\frac{1.30^2}{21} + \frac{1.16^2}{25}\right)}} = \frac{0.74}{0.3665} = 2.019 \quad \text{with 40.6 degrees of freedom}$$

$$p\text{-}val = 2 \times P(t_{40.6} > 2.019) = 0.0501$$

Separate variance t test example 2 (case 2):

- WOW regular run "promotions" on Coca-Cola. Does this increase sales?

Summary statistics for RAW_SLS:
Group by: Promotion Flag

Promotion Flag	n	Mean	Std. dev.	Median	Skewness	Kurtosis	IQR
N	17	69	22.1	70	-0.06	-0.12	30
Y	26	580.4	379.2	462	1.55	1.90	354

  - Here, we would assume population variances are separate because sample standard deviations are extremely different
  - Both samples are below n = 30 but above n = 15 so we need to assume population distributions are both normal or (at least) symmetrically distributed for CLT to hold
- State the null and alternative hypotheses:

$H_0 : \mu_P = \mu_{NP}$ $H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_P > \mu_{NP}$ $H_A : \mu_1 - \mu_2 > 0$ (without pooled variances)

- Sample results:

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	511.42	74.56	25.26	6.859	<0.0001

NB Critical value is 2.4851 for df 25

- Decision:

$$p-value = P(t_{25.26} > 6.859) = 1 - \text{T.DIST}(6.859, 25.26)$$

p-value < 0.0001 < 0.01 = a. Thus reject the null hypothesis

- Conclusion:
  - There is sufficient evidence to conclude that Coca Cola sales when "promoted" are statistically significantly higher than when "non-promoted", at the 1% significance level

Separate variance t test example 3 (case 2):

Amazon trialled two different sales generation methods:

1. "Expert" raters & famous authors upload ratings and reviews to inform buyers.
2. Buyers rate and review books. Also, big data analytic methods used to recommend books to buyers, based on their purchases and those of other buyers.

Amazon conducted an online experiment, with users randomly allocated to each sales method.
For method 1, the average purchase amount per year was $23.10.
For method 2, the amount was $45.52.

What should Amazon do? NB:

$$n_1 = 120; S_1 = 5.34$$
$$n_2 = 125; S_2 = 8.46$$

- S_2 is significantly larger than S_1 so we assume separate population variances
- Both n > 30, so we can assume that CLT holds

- Solution:
  - 1. State the null and alternative hypotheses
    - H_0: $\mu_1 - \mu_2 = 0$ (i.e. $\mu_1 = \mu_2$)
    - H_1: $\mu_1 - \mu_2 \neq 0$ (i.e. $\mu_1 \neq \mu_2$) → this is a two-tail test (since no direction was assumed or mentioned)
  - 2. Specify the desired level of significance and the sample size
    - α = 0.05, n_1 = 120, and n_2 = 125
  - 3. Determine the appropriate technique
    - σ is unknown so this is a t test
    - Population variances are assumed unequal, so this is a separate-variance t test
  - 4. Determine the DF and critical values
    - $$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{5.34^2}{120} + \frac{8.46^2}{125}\right)^2}{\frac{\left(\frac{5.34^2}{120}\right)^2}{120 - 1} + \frac{\left(\frac{8.46^2}{125}\right)^2}{125 - 1}} = 210.5$$
    - For α = 0.05 and DF = 210.5, the critical t values (t_{CRIT}) are $\pm$ 1.9173
  - 5. Collect the data and compute the test statistic
    - $$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{23.1 - 45.52}{\sqrt{\frac{5.34^2}{120} + \frac{8.46^2}{125}}} = \frac{-22.42}{0.90} = -24.91$$
  - 6. Make the statistical decision and state the managerial conclusion
    - t_{STAT} = – 24.91 < t_{CRIT} = – 1.9173
    - Reject H_0 at the 5% significance level
    - There is sufficient evidence to conclude that there is a statistically significant difference between the two methods for Amazon, at the 5% significance level

Excel example for separate variance t-test (case 2):

- FoodPlace Supermarket wants to compare the sales of a new Cola – "Real Citrus Cola", based on two potential locations within each store:
  - 1. Beverage section
  - 2. Produce section
  - Sample results:

Sample	Mean	Std. dev.
Beverage	50.3	18.73
Produce	72	12.54

    - 20 stores are included in the trial, 10 are randomly allocated to each option
  - Required assumptions:
    - Assume that population variances are separate because sample standard deviations are significantly different
    - Assume that populations distributions are normally distributed because sample sizes are both < 15

- Solution:
  - 1. State the null and alternative hypotheses
    - H_0: $\mu_1 - \mu_2 = 0$ (i.e. $\mu_1 = \mu_2$)
    - H_1: $\mu_1 - \mu_2 \neq 0$ (i.e. $\mu_1 \neq \mu_2$) → this is a two-tail test (since no direction was assumed or mentioned)
  - 2. Specify the desired level of significance and the sample size
    - α = 0.05, n_1 = 10, and n_2 = 10
  - 3. Determine the appropriate technique
    - σ is unknown so this is a t test
    - Population variances are assumed unequal, so this is a separate-variance t test
  - 4. Determine the DF and critical values
    - DF = 16
    - For α = 0.05 and DF = 16, the critical t values (t_{CRIT}) are $\pm$ 2.1199
  - 5. Collect the data and compute the test statistic
    - t_{STAT} = – 3.04455
  - 6. Make the statistical decision and state the managerial conclusion
    - t_{STAT} = – 3.04455 < t_{CRIT} = – 2.1199
    - Reject H_0 at the 5% significance level
    - There is sufficient evidence to conclude that there is a statistically significant difference between the two locations for FoodPlace Supermarket, at the 5% significance level
- Excel output:

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	Variable 2
Mean	50.3	72
Variance	350.6778	157.3333
Observations	10	10
Hypothesized Mean Difference	0	
df	16	
t Stat	-3.04455	
P(T<=t) one-tail	0.003863	
t Critical one-tail	1.745884	
P(T<=t) two-tail	0.007726	
t Critical two-tail	2.119905	

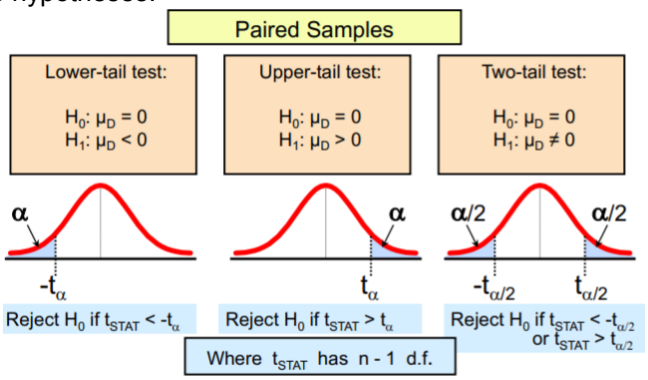
Confidence interval for $\mu_1 - \mu_2$, with σ_1 and σ_2 unknown and not assumed equal (case 2):

- E.g. continued: Since we rejected H_0 can we be 95% confident that $\mu_1 \neq \mu_2$?
  - 95% confidence interval for $\mu_1 - \mu_2$:
    - $$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{df,\alpha/2}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$
      $$= -21.7 \pm 2.1199 \times 7.1275 = (-36.81, -6.59)$$
  - Since 0 is less than the entire interval, we can be 95% confident that $\mu_1 \neq \mu_2$

Difference between two population means – related samples:

The paired difference test:

- Tests means of 2 related populations:
  - Paired or matched samples e.g. repeated measures (before vs. after)
  - Uses the difference between paired values:
    - $D_i = X_{1i} - X_{2i}$
    - D_i represents the i^{th} paired difference
- This test eliminates variation among subjects
  - E.g. quiz 1 and quiz 2 scores (for the same students)
- Assumptions:
  - 1. Both populations are normally distributed
  - 2. Or, if not normal, use large enough sample (n > 30 or n > 15 if symmetric)
    - Exam note: n is the number of pairs in the paired sample
- The point estimate ($\overline{D}$) for the paired difference population mean (μ_D) is:
  - $$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$
- The sample standard deviation is S_D:
  - $$S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}$$
- The test statistic for $\overline{D}$ is:
  - $$t_{STAT} = \frac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$
  - The t_{STAT} DF:
    - n – 1
- Possible hypotheses:



- The paired difference confidence interval for μ_D is:
  - $$\overline{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

  where $S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}$

Difference between population proportions:

- Goal: Test a hypothesis or form a confidence interval for the difference between two population proportions, $\pi_1 - \pi_2$
  - The point estimate for the difference is $p_1 - p_2$
- To compute the test statistic, we assume the null hypothesis is true, so we assume $\pi_1 = \pi_2$, and pool the two sample estimates
  - The pooled estimate for the overall proportion is:

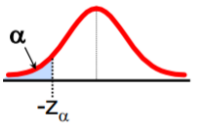
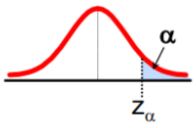
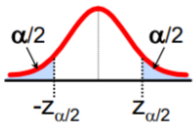
$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

    - where X_1 and X_2 are the number of items of interest in samples 1 and 2
- The test statistic for $\pi_1 - \pi_2$ is a Z statistic:

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  where $\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$, $p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$
- Hypothesis tests:

Population proportions

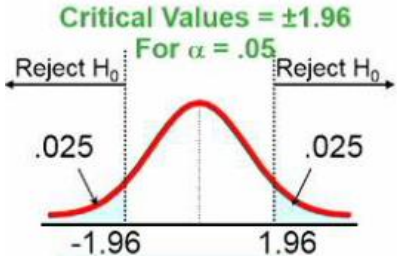
Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 < \pi_2$ i.e., $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 < 0$	$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 > \pi_2$ i.e., $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 > 0$	$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$ i.e., $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 \neq 0$
		
Reject H_0 if $Z_{STAT} < -Z_\alpha$	Reject H_0 if $Z_{STAT} > Z_\alpha$	Reject H_0 if $Z_{STAT} < -Z_{\alpha/2}$ or $Z_{STAT} > Z_{\alpha/2}$

Hypothesis test for two population proportions example:

- Is there a significant difference between the proportion of men and the proportion of women who will vote 'Yes' on a referendum question?
  - In a random sample, 36 of 72 men and 35 of 50 women indicated they would vote 'Yes'
  - Test at the 0.05 level of significance
- The hypothesis test is:
  - $H_0: \pi_1 - \pi_2 = 0$ (the two proportions are equal)
  - $H_1: \pi_1 - \pi_2 \neq 0$ (the two proportions are different)
- The sample proportions are:
  - Men: $p_1 = 36/72 = 0.50$
  - Women: $p_2 = 35/50 = 0.70$
- The pooled estimate for the overall proportion is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{36 + 35}{72 + 50} = \frac{71}{122} = 0.582$$

- Rejection region:



- The test statistic for $\pi_1 - \pi_2$ is:

$$z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(.50 - .70) - (0)}{\sqrt{.582(1-.582)\left(\frac{1}{72} + \frac{1}{50}\right)}} = -2.20$$

- Decision:
  - Reject the null hypothesis
- Conclusion:
  - There is sufficient evidence of a statistically significant difference in proportions who will vote 'Yes' between men and women, at the 5% level of significance
- On Excel:

Z Test for Differences in Two Proportions

	A	B	
4	Hypothesized Difference	0	
5	Level of Significance	0.05	
6	**Group 1**		
7	Number of items of interest	36	
8	Sample Size	72	
9	**Group 2**		
10	Number of items of interest	35	
11	Sample Size	50	
12			
13	**Intermediate Calculations**		
14	Group 1 Proportion	0.5	=B7/B8
15	Group 2 Proportion	0.7	=B10/B11
16	Difference in Two Proportions	-0.2	=B14 - B15
17	Average Proportion	0.582	=(B7 + B10)/(B8 + B11)
18	Z Test Statistic	-2.20	=(B16-B4)/SQRT((B17*(1-B1 B11))
19			
20	**Two-Tail Test**		
21	Lower Critical Value	-1.96	=NORM.S.INV(B5/2)
22	Upper Critical Value	1.96	=NORM.S.INV(1 - B5/2)
23	p-value	0.028	=2*(1 – NORM.S.DIST(ABS(B18)))
24	Reject the null hypothesis		=IF(B23 < B5,"Reject the null hypothesis", "Do not reject the null hypothesis")

Since -2.20 < -1.96

Or

Since p-value = 0.028 < 0.05

We reject the null hypothesis

Decision: Reject H_0

Conclusion: There is evidence of a difference in proportions who will vote yes between men and women.

64

Confidence interval for two population proportions:

- The confidence interval for $\pi_1 - \pi_2$ is:

$$(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  - Exam note: Remember to use the different standard error here
- CLT:

$$n_1 p_1 \geq 5;\ n_1(1-p_1) \geq 5$$
$$n_2 p_2 \geq 5;\ n_2(1-p_2) \geq 5$$

are needed for the CLT to hold here

Covariance and correlation:

The covariance:

- Measures the "strength" of the linear relationship between two numerical variables (X & Y)
- The sample covariance:
  - $$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
  - Only concerned with the strength of the linear relationship
  - No causal effect is implied i.e. "correlation not causation"

Interpreting covariance:

- Covariance between two variables:
  - cov(X,Y) > 0 ⟶ X and Y tend to move in the same direction
  - cov(X,Y) < 0 ⟶ X and Y tend to move in opposite directions
  - cov(X,Y) = 0 ⟵ X and Y are independent
- Covariance has a flaw:
  - It is not possible to determine the relative strength of the relationship from the size of the covariance

Coefficient of correlation:

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

where

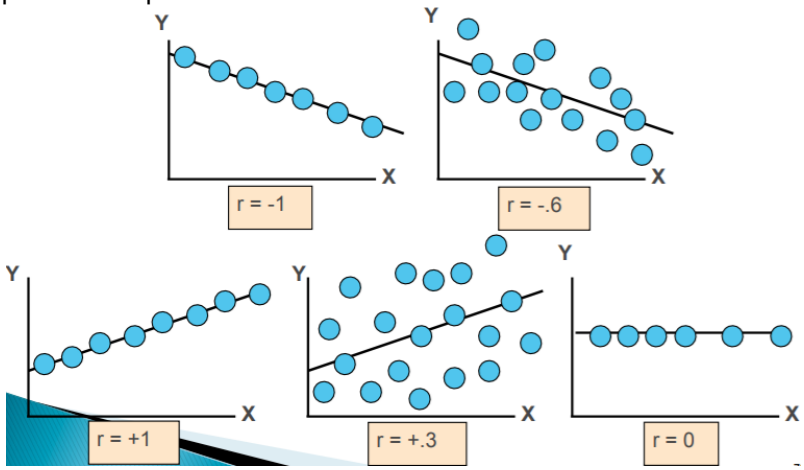
$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

  - Exam note: The sign of the sample coefficient of correlation only depends on the covariance because $S_X S_Y$ will always be positive
- Coefficient of correlation overcomes the covariance flaw, but it still has weaknesses such as:
  - It does not show the direction of the relationship
  - It does not show a causal relationship

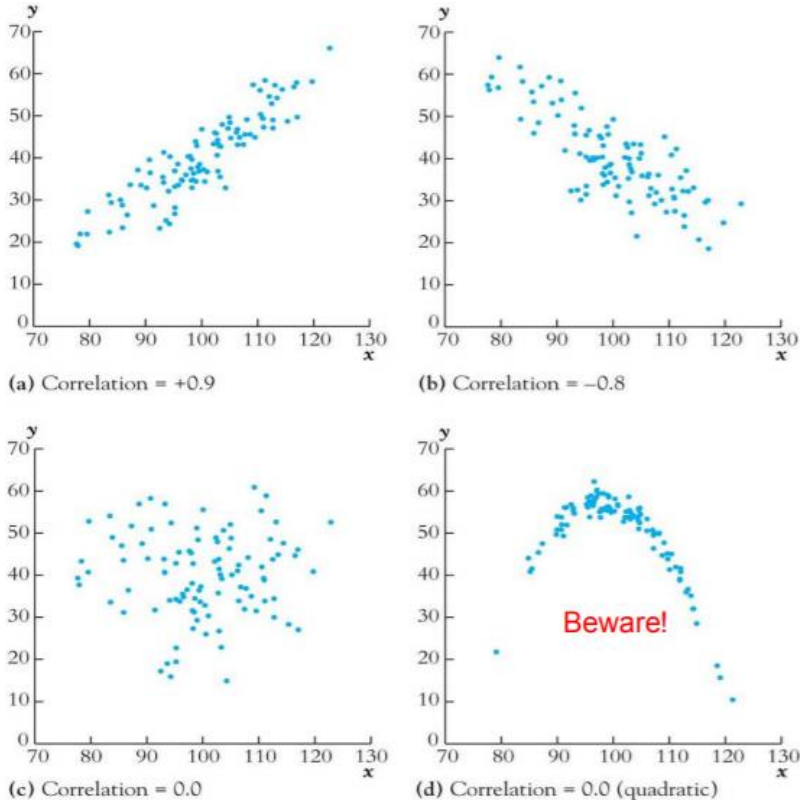
Features of the coefficient of correlation:

- The population correlation is denoted ρ
- The sample correlation is denoted r
- Both ρ and r have the following features:
  - Unit free
  - Range between – 1 and 1
    - The closer to – 1, the stronger the negative linear relationship
    - The closer to 1, the stronger the positive linear relationship
    - The closer to 0, the weaker the linear relationship
    - E.g.: A correlation coefficient of 1 shows that X and Y are perfectly positively correlated, but this does not show the direction of the relationship (i.e. X change leads to Y change), nor does it show causation (X changing is causing Y to change)

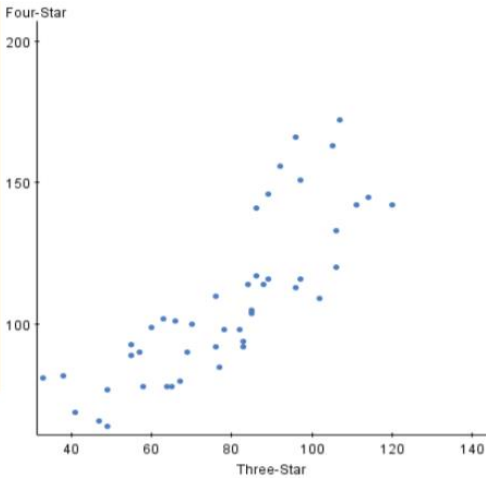
- Scatter plots of sample data with various coefficients of correlation:



- The correlation coefficient measures linear association:



- Interpreting the coefficient of correlation example:
  - r = .852
  - There is a strong positive linear relationship between room rates for 3 and 4 star hotels over 47 major cities across the world.
  - Cities with higher 3 star hotel rates tended to also have higher 4 star hotel room rates.



Week 11-12 – Simple Linear Regression:

Week 11-12 LO's:

Agenda

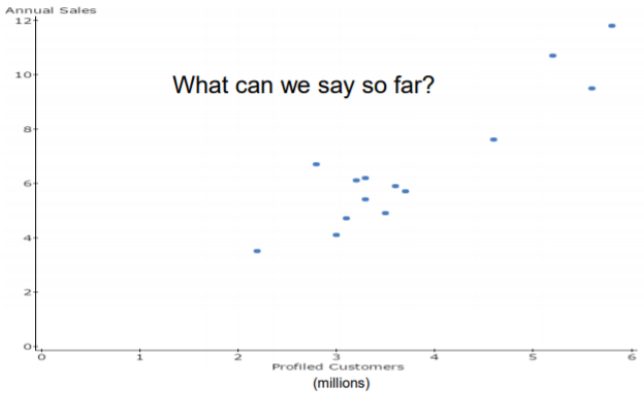
- Background
- Simple Linear Regression
- Measures of Variation
- Residual Analysis to verify assumptions
- Tests that the linear relationship is not due to chance
- Estimating Means and Individual Values
- Multiple Linear Regression
- Autocorrelation
- Pitfalls of Regression

Background:

Correlation vs. regression:

- A scatter plot can be used to show the relationship between two variables
- A correlation analysis is used to measure the strength of the linear relationship between two variables
  - Correlation is only concerned with strength and sign of the relationship, but doesn't tell us what the relationship is so we cannot predict one value from another
  - No causal effect nor direction is implied by correlation i.e. it does not suggest direction nor causation of the relationship (whether X is causing Y to change, or Y is causing X to change)
- A regression analysis is used to:
  - Predict the value of a "dependent" variable (Y) based on the value of at least one "independent" variable (X)
  - Explain the impact of changes in an independent variable on the dependent variable
  - Dependent variable (Y) – the variable we wish to predict or explain
  - Independent variable (X) – the variable used to predict or explain the dependent variable

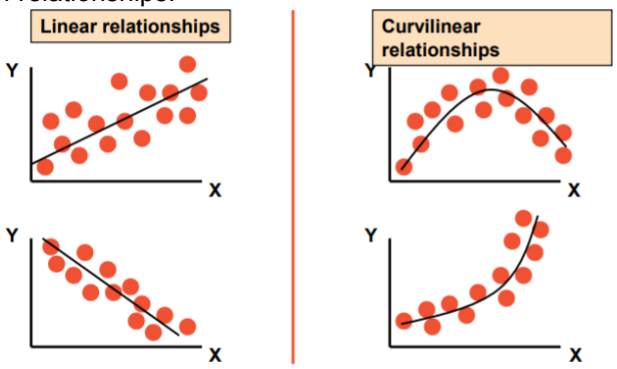
Sunflowers Apparel: wants to forecast annual sales,
based on the number of customers who live <30 mins from each store



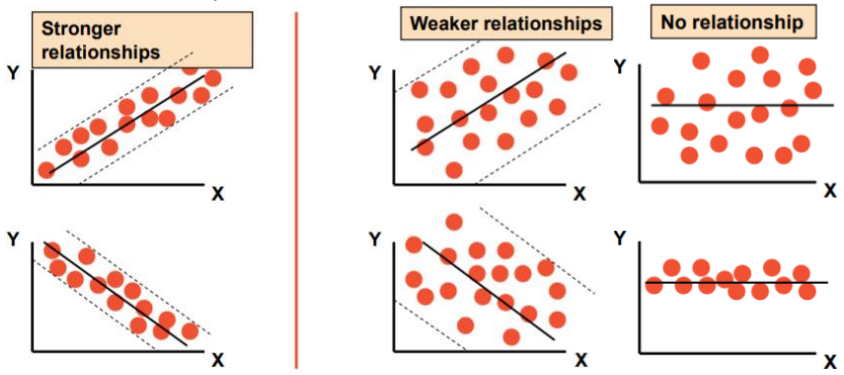
- Correlation between Annual Sales and Profiled Customers is: **0.921**
  - There is a very strong positive and linear relationship based off the correlation
- Regression to the mean:
  - If a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, and if it is extreme on its second measurement, it will tend to have been closer to the average on its first
  - E.g. shooting a basketball extremely well today implies you will be less likely to shoot as well the next day (i.e. will not be as lucky) → accuracy will move back to the mean

Simple Linear Regression (SLR):

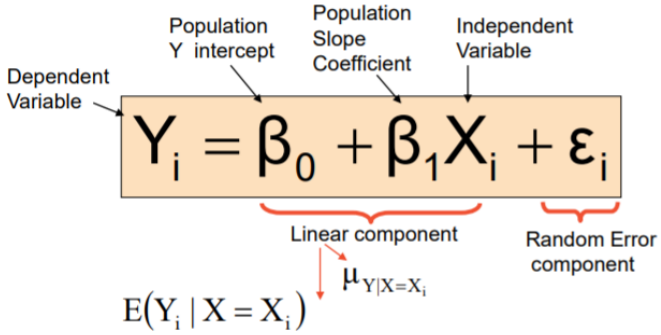
- Now looking at two different variables (rather than only one as seen in previous week's content)
  - There is always on dependent variable (Y) in regression analysis
  - For SLR, there is only one independent variable (X)
  - Relationship between X and Y is described by a linear function (assume X and Y are linear)
  - Unlike correlation, we now assume a direction (the direction X → Y is implicit in SLR)
    - Changes in Y are assumed to be related to changes in X
  - Note: Regression still does NOT prove causation, but it (implicitly) assumes it
- Terminology:
  - The independent variable (X) is also called: an explanatory variable, a predictor variable, a regressor, a feature, a factor
  - The dependent variable (Y) is also called: the response variable
- Types of relationships:



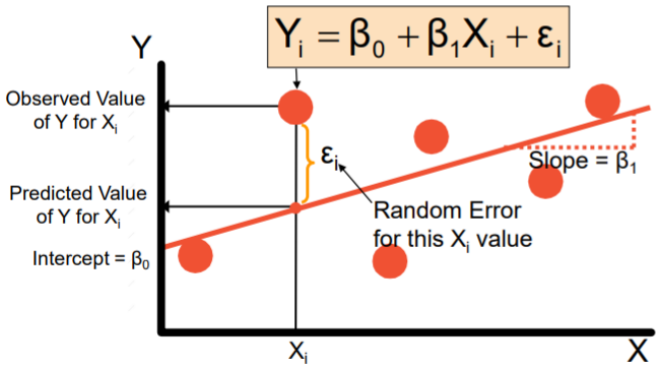
  - We will only be analysing linear relationships
- Types of linear relationships:



  - Stronger relationships will have higher correlation values (close to – 1 or 1)
  - No relationship: changing X has no effect on the value of Y
- SLR model – population regression function (PRF):

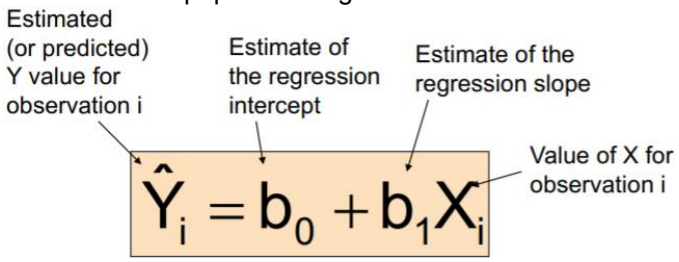


  - Same idea as Y = mX+C
  - β_0 and β_1 represents our population parameters
  - The PRF can also be called the population regression line

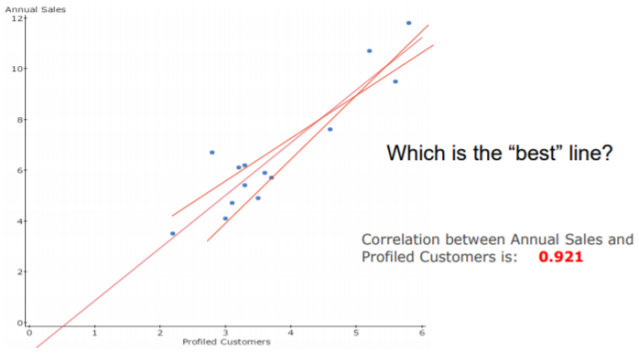


- β_0 is the Y-intercept of our SLR model, representing the mean value of Y when X = 0
- β_1 is the slope of the PRF
- ε represents a random error component → we don't expect every single point to lie perfectly on the PRF, so we allow for the vertical deviations from the line with the error component
- The linear component is what we expect on average i.e. for any given value of X, on average, Y should fall onto the linear component (error term is zero on average for all values of X)

- Estimated SLR equation – sample regression function (SRF):
  - Since we won't have all the population data points, we cannot construct a population regression function
  - We estimate the PRF with the SRF i.e. the estimated SLR equation provides an estimate of the population regression line



  - b_0 is the sample estimate of β_0
  - b_1 is the sample estimate of β_1
  - Note here that we are calculating the estimated/predicted Y value → this is denoted by Y hat ($\hat{Y}$)
  - $\hat{Y}_i$ lies on the SRF for any value of X_i
  - Exam note: Errors are assumed to be zero when we are making a prediction because we are predicting for the average, and on average the errors cancel out to be zero
  - The SRF can also be called the sample prediction line



- How best to estimate the line?
  - We need to be able to compute the b_0 and b_1 values to be able to draw this estimated SLR line

One approach for estimating the SLR equation – the Least Squares method:

- Given some data pairs (Y_i, X_i), where i = 1, …, n, b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences (residuals) between Y and $\hat{Y}$:
  - $$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$
  - We can now determine the b_0 and b_1 values that minimize the function above
- This method is based on calculus techniques that produce a minimum of the sum of squared errors:
  - The partial derivatives:
    - $$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$
    - $$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$
    - 2 equations in 2 unknowns (Derivation not examinable)
  - Solve the 2 equations to reach our general solutions:
    - $$b_0 = \bar{y} - b_1 \bar{x}$$
    - $$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$
  - Note the link between b_1 formula and the covariance formula
    - $$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
    - b_1 formula is the same as covariance divided by variance of X
    - Exam note: Recall that the sign of correlation only depends on the covariance. The sign of the slope of the SRF (b_1) also only depends on the covariance, since the variance in the denominator is always positive. Therefore, $b_1$and the correlation coefficient (and covariance) will always have the same sign.
  - After finding b_1, we can easily find b_0 by substituting into the b_0 equation
- SS_{xy} and SS_{xx}:

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = (n-1)S_x^2$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

  - SS_{xy} is equal to the numerator term of the covariance formula → it is called the corrected sum of the products of X and Y
  - SS_{xx} is equal to the numerator term of the variance → it is called the sum of the squares of X
  - b_1 is covariance divided by variance as the denominator term (n – 1) cancels out

- Interpretation of the intercept (b_0) and slope (b_1):
  - b_0 is the estimated intercept:
    - $b_0 = \hat{E}(Y \mid X = 0) \longleftrightarrow \beta_0 = E(Y \mid X = 0)$
    - b_0 is the estimated mean value of Y when the value of X is zero (X = 0)
    - b_0 is an estimator of the expected Y value when X is zero whereas β_0 is the true expected Y value when X is zero
  - b1 is the estimated slope:
    - $b_1 = \hat{E}(Y \mid X+1) - \hat{E}(Y \mid X)$

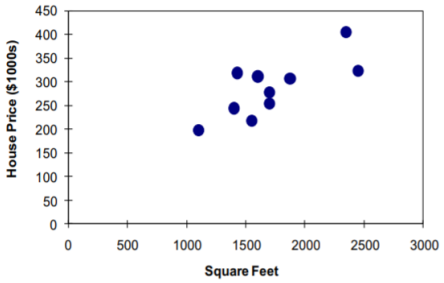
      $\updownarrow$

      $\beta_1 = E(Y \mid X+1) - E(Y \mid X)$
    - b_1 is the estimated change in the mean value of Y as a result of a one-unit increase in X (i.e. X → X + 1)
    - b_1 is an estimator of the expected change in Y when X increases from X to X + 1 where β_1 is the true expected change in Y when X increases from X to X + 1

SLR example:

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
  - A random sample of 10 houses is selected
    - Dependent variable (Y) = house price in $1000s
    - Independent variable (X) = square feet
  - Data points:
House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

      House price model: Scatter Plot

  - Regression output:

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

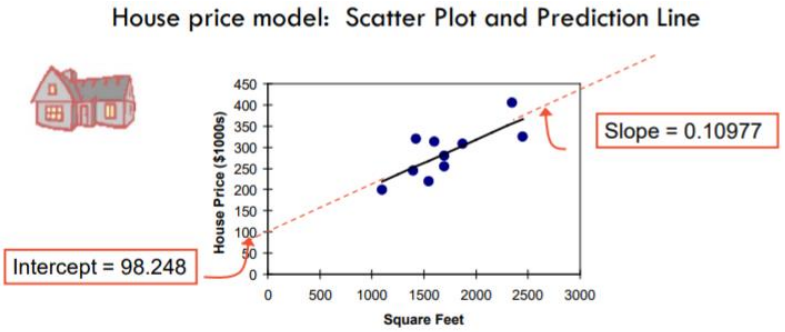
    The regression equation is:

    $\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

- Graphical representations of SRF:



$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$

- Interpretation of b_0:
  - $\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$
  - b_0 is the estimated mean value of Y when the value of X is zero
  - Because a house cannot have a square footage of 0, b_0 has no practical application. We choose not to interpret it explicitly here.
  - Interpret b_0 only if X = 0 is a possible value and if 0 is in the range of observed X values
- Interpretation of b_1:
  - $\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$
  - b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
  - Here, b_1 = 0.10977 tells us that the mean value (on average) of a house increases by 0.10977($1000) = $109.77, on average, for each additional one square foot of size
- Using SLR equation to make predictions:

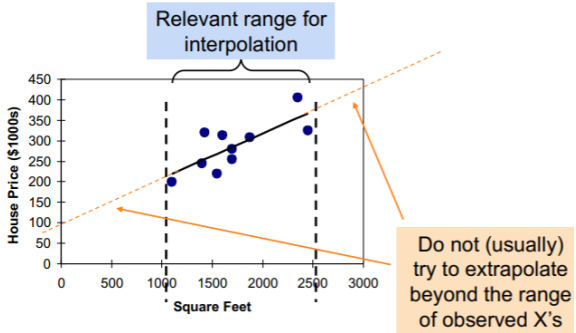
Predict the price for a house with an area of 2000 square feet:

NB Is 2000 in the range of observed X values?

$$\begin{aligned}\widehat{\text{house price}} &= 98.24833 + 0.10977\,(\text{sq.ft.}) \\ &= 98.24833 + 0.10977(2000) \\ &= 317.78\end{aligned}$$

The predicted price for a house with a total area of 2000 square feet is 317.78($1,000s) = $317,780

  - When interpreting b_1 → Y's change is on average
  - Interpolation range:



  - 2000 is clearly in the range of observed X values
  - When using a regression model for prediction, we (usually) only predict within the relevant range of data
  - The validity of a point is lacking if there are not many other data points around that value

Measures of variation:

Total variation:

- When using the least squares method, we need to calculate 3 types of variation
- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum(Y_i - \bar{Y})^2$	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$SSE = \sum(Y_i - \hat{Y}_i)^2$

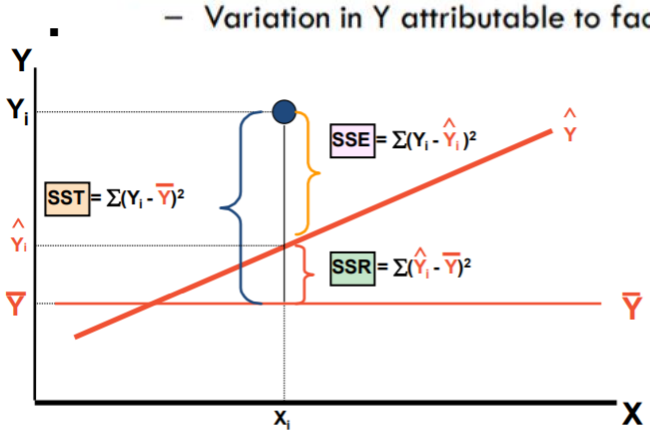
where:

$\bar{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

- SST = total sum of squares (Total Variation)
  - Measures the variation of the Y_i values around their mean $\bar{Y}$
- SSR = regression sum of squares (Explained Variation)
  - Variation attributable to the relationship between X and Y
- SSE = error sum of squares (Unexplained Variation)
  - Variation in Y attributable to factors other than X



Coefficient of determination (r^2):

- The coefficient of determination is the proportion of the total variation in Y that can be explained by the variation in X
  - The coefficient of determination is also called r-squared and is denoted as r^2

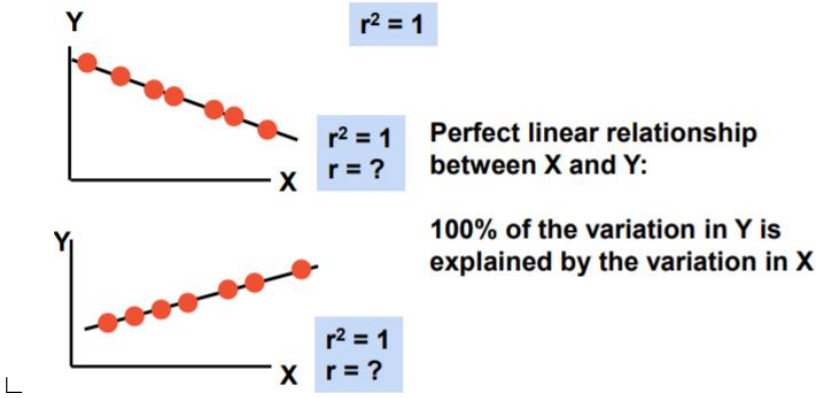
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note: $0 \le r^2 \le 1$

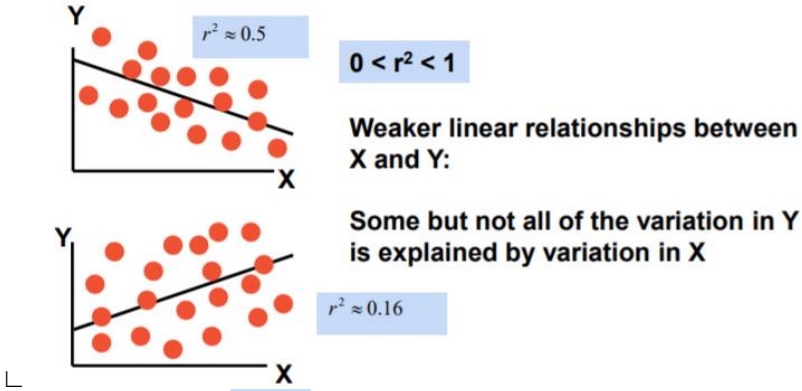
Here r IS the coefficient of correlation!

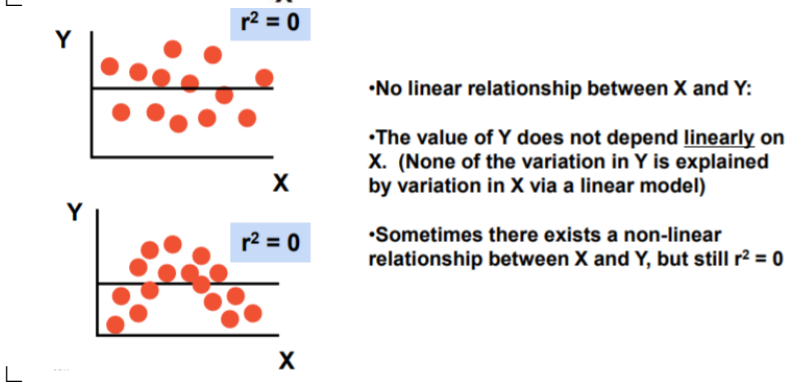
  - Value of 0 – the model explains none of the total variation in Y
  - Value of 1 – the model explains all of the total variation in Y
  - The higher the r^2 value, the better → more of the variation of Y is expected/can be explained)

- Examples of approximate r^2 values:



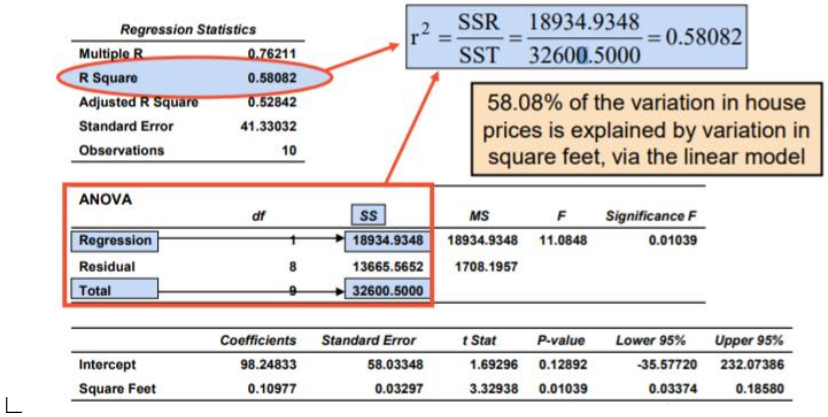
  - r is - 1 for top model and 1 for bottom model
  - Exam note: Be cautious that a r^2 = 1 value can come from a regression where r is 1 or - 1





  - A $r^2 = 0$ may mean there is no relationship at all; however, there could be a quadratic (or any other non-linear) relationship
  - r^2 is only strictly looking at linear relationships (just like r)

- On Excel:



Standard error of the estimate:

- The standard error of the observations around the regression line is estimated by:

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

SSE = error sum of squares

n = sample size

  - S_{YX} is measured in the same units as Y
  - We divide by n – 2 because we've lost 2 degrees of freedom by needing to estimate the intercept and the slope term

- On Excel:

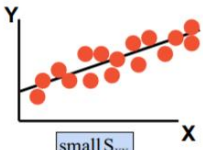
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$S_{YX} = 41.33032$

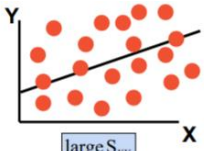
ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

- Comparing standard errors:
  - S_{YX} is a measure of the variation of observed Y values from the regression line



small S_{YX}
large r^2



large S_{YX}
small r^2

The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., S_{YX} = \$41.33K is moderately small relative to house prices in the \$200K - \$400K range

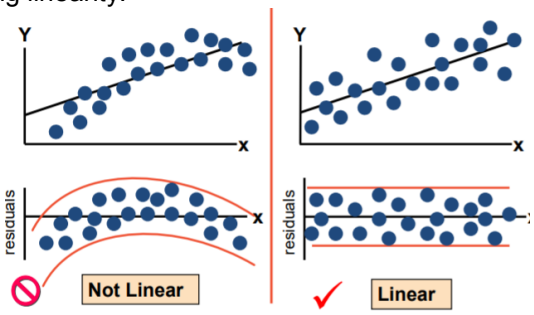
  - Models with a better fit have smaller standard errors (and larger r^2)

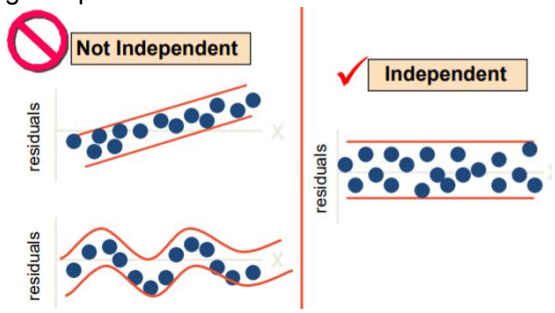
Residual analysis to verify the assumptions for a simple linear regression model:

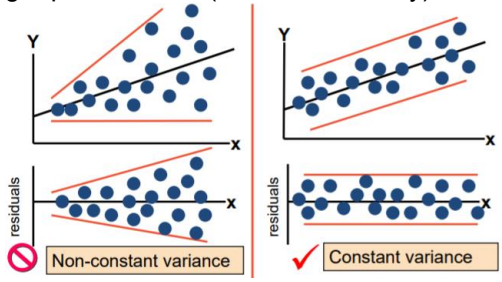
Assumptions of regression (LINE) Need to fulfil these for a valid SLR equation

- We need to fulfil these LINE assumptions for a valid SLR equation (for valid inferences)
- Determines whether prediction model (a linear model) selected is appropriate as an estimator for the data set
- Linearity:
  - The relationship between X and Y is linear
- Independence of errors:
  - Error values are statistically independent (this becomes more difficult to satisfy with a data set given over a period of time i.e. time series data) → our data values need to be independent
- Normality of error:
  - Error values are normally distributed for any given X value (this assumption is only required for small samples, or for prediction intervals)
- Equal variance (also called homoscedasticity)
  - The probability distribution of the errors has a constant variance for any given X value

Residual analysis (using residuals as estimates of errors to check assumptions about the errors):

- The residual (e_i) for observation i is the difference between the i^{th} observed and predicted values
  - $e_i = Y_i - \hat{Y}_i$
- Check the assumptions of regression errors by examining the residuals (plot the residuals against the X variable)
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption (ONLY in small samples OR when doing prediction intervals)
  - Examine for equal variance for all levels of X (homoscedasticity)
- Checking linearity:
  - 

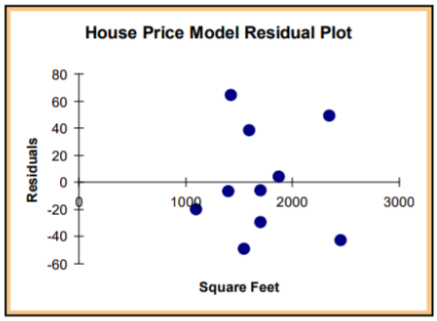
  - Distance of residuals from both sides of the line should be constant (and not follow a non-linear pattern)
- Checking independence:
  - 

  - Residuals should be randomly positive and negative (rather than follow a pattern)
- Checking normality (checking that the probability distribution of the residuals is normal):
  - Exam note: This is necessary only in small samples (n < 30), or when doing individual prediction intervals (see later)
  - Examine the boxplot of the residuals
  - Examine the histogram of the residuals
  - Assess whether empirical rule (and other normal-based rules) are followed closely.
  - Use rules of thumb for checking normality of the residuals
- Checking equal variance (homoscedasticity):
  - 

  - Non-constant variance is called heteroskedasticity
  - Want distance of residuals from both sides of the line to be constant (and not increase along X)

- Example:

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Hard to tell in small sample!

Tests that the linear relationship is not due to chance:

- With all assumptions satisfied and knowing SLR equation is appropriate, we can now make inferences about the linear relationship between the variables in the population (t-tests, F-tests, confidence intervals and prediction intervals)
- Inferences about the slope:
  - The standard error of the regression slope coefficient (b_1) is estimated by:

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \overline{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

- Central Limit Theorem:
  - The Least Squares (LS) estimates of the intercept and the slope parameters are linear combinations of the observations Y
    - Thus, the CLT applies to the intercept and slope parameters
    - If Y is normally distributed, then so are the intercept and slope estimators
  - In large samples, the intercept and slope estimators are approximately normally distributed
- Inferences about the slope – t-test:

t test for a population slope
- Is there a linear relationship between X and Y?

Null and alternative hypotheses
- H_0: $\beta_1 = 0$ (no linear relationship)
- H_1: $\beta_1 \neq 0$ (linear relationship does exist)

Test statistic

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

  - If null hypothesis is rejected, we conclude there is a signification linear relationship
  - The t_{STAT} test statistic has degrees of freedom n – 2

- Inferences about the slope t test example:

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098\ (\text{sq.ft.})$$

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

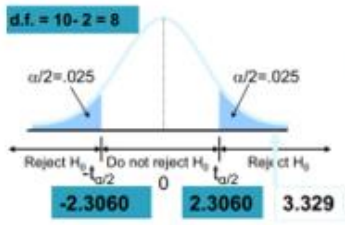
b_1 S_{b_1}

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

- Critical value approach:

Test Statistic: $t_{STAT} = 3.329$

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage significantly affects house price

- P-value approach:

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

p-value

Decision: Reject H_0, since p-value $< \alpha = 0.05$

There is sufficient evidence that square footage affects house price, at the 5% significance level.

- Confidence interval estimate for the slope:

$$b_1 \pm t_{\alpha/2} S_{b_1}$$ d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

  - The 95% confidence interval for the slope is (0.0337, 0.1858)
  - Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.74 and $185.80 per square foot of house size
  - This 95% confidence interval does not include 0
  - Conclusion: There is sufficient evidence to conclude a statistically significant relationship between house price and square feet at the 0.05 level of significance (not just by chance)
  - Exam note: The $t_{\alpha/2}$ value is calculated assuming $\beta_1 = 0$ (which will be the case for most regression software outputs), so remember to adjust this based on the question accordingly
- t-test for a correlation coefficient:

Hypotheses

H_0: $\rho = 0$ (no correlation between X and Y)

H_1: $\rho \neq 0$ (correlation exists)

Test statistic

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$ (with n – 2 degrees of freedom)

  - We are now testing that correlation is not due to chance
- Correlation coefficient t-test example:

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

H_0: $\rho = 0$ (No correlation)

H_1: $\rho \neq 0$ (correlation exists)

$\alpha = .05$, df = 10 - 2 = 8

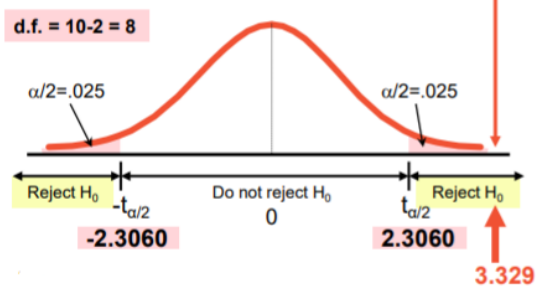
$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762 - 0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$

  - Slope t statistic (for testing whether $\beta_1 = 0$) is same as the correlation coefficient t statistic

$$t_{STAT} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762-0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$

Decision: Reject H_0

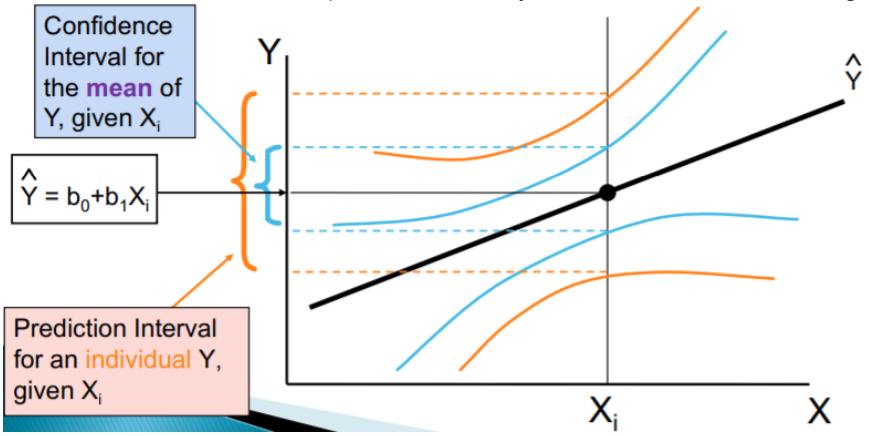
Conclusion:
There **is evidence** of a linear association at the 5% level of significance



- Both the t test for the slope coefficient and correlation coefficient show that there is a significant relationship between X and Y at the 5% level of significance

Estimating means and individual values:

- Goal: Form intervals around Y to express uncertainty about the value of Y for a given X



  - Confidence and prediction intervals are narrower at the mean of X and wider for X values further away from the mean
  - Blue lines represent the confidence interval for the mean of Y
  - Orange lines represent the prediction interval for an individual Y value
- Confidence interval for the average Y, given X:

Confidence interval estimate for the mean value of Y given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

This is a central limit theorem result

Size of interval varies according to distance away from mean, $\bar{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

  - The confidence interval is wider for X_i values further from the mean → this is shown in the formula by h_i increasing when X_i is further from the mean

- Prediction interval for an individual Y, given X:

Confidence interval estimate for an Individual value of Y given a particular X_i

Confidence interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1+h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

This is ONLY valid if the population for Y|X is conditionally normally distributed, around the regression line

  - The prediction interval is wider for X_i values further from the mean → this is again shown in the formula by h_i increasing when X_i is further from the mean
- Estimation of mean values example:

Confidence Interval Estimate for $\mu_{Y|X=X}$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i$ = 317.78 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.78 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90: we are 95% confident that the true mean price for 2000 foot houses is between $280,660 and $354,900, assuming the central limit theorem holds (or the data are normally distributed around the regression line)

  - Note the "assuming CLT holds (or the data are normally distributed around the regression line)"
- Estimation of individual values example:

Prediction Interval Estimate for $Y_{X=X}$

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price Y_i = 317.78 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.78 \pm 102.28$$

The prediction interval endpoints are 215.50 and 420.07: we are 95% confident that the house price for a 200 foot home is between $215,500 and $420,070, assuming that house prices are conditionally normally distributed around the regression line.

  - Note the "assuming the house prices are conditionally normally distributed around the regression line"
- Degrees of freedom for confidence and prediction intervals are n – k – 1 (for calculating the test statistic)

Week 13 – Multiple Linear Regression:

Week 13 LO's:

Agenda

- Background
- Simple Linear Regression
- Measures of Variation
- Residual Analysis to verify assumptions
- Tests that the linear relationship is not due to chance
- Estimating Means and Individual Values
- Multiple Linear Regression
- Autocorrelation
- Pitfalls of Regression

Multiple Linear Regression (MLR):

- Now looking at more than two different variables
  - Again, there is always on dependent variable (Y) in regression analysis
  - For MLR, there is more than one independent variable (X)
- MLR model – population regression function (PRF):

Idea: Examine the linear relationship between dependent (Y)

& several independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

  - The only difference now is that there are k independent X variables rather than just 1 in SLR
- Estimated MLR equation – sample regression function (SRF):

The coefficients of the multiple regression model are again estimated using sample data

Multiple regression equation with k independent variables:

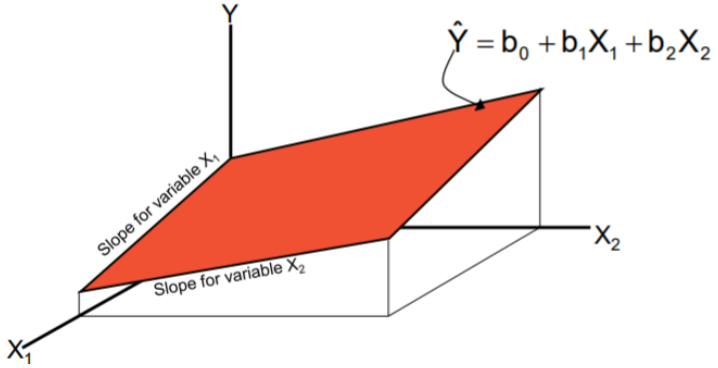
Estimated (or predicted) value of Y | Estimated intercept | Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

In this chapter we will use Excel to obtain the regression slope coefficients and other regression summary measures.

  - Again, note that there is no error term for the predicted value of Y

Two variable model



Example of MLR equation with 2 independent variables:

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand. Data points are collected for 15 weeks:
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $), Advertising ($100's)

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{Sales} = b_0 + b_1 (Price) + b_2 (Advertising)$$

- Excel output:

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{Sales} = 306.526 - 24.975(Price) + 74.131(Advertising)$$

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

  - The t test looks at an individual slope coefficient to see if they statistically significantly different to the hypothesized value (generally zero), or whether it is insignificant
  - The f-test looks at all the slope coefficients of all the X variables at the same time to see if any of the slope coefficients are statistically significantly different to the hypothesized value (generally zero), or whether they are jointly insignificant i.e. if the F value is zero, it suggests none of the X variables are explaining any movements in Y (sales) → measures the overall usefulness of this model
  - The p-value for the F-test here is 1.2% < 5%, so we would reject the null hypothesis that all of the slopes are zero

- The MLR equation interpretation:

$$\widehat{Sales} = 306.526 - 24.975(Price) + 74.131(Advertising)$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100's.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, **if the level of advertising is held constant**

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, **if the price is held constant**

  - When interpreting b_k → Y's change is on average
  - Exam note: We must always assume that we are holding the other X variables as constants
- Using MLR equation to make predictions:

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\begin{aligned}\widehat{Sales} &= 306.526 - 24.975(Price) + 74.131(Advertising)\\ &= 306.526 - 24.975(5.50) + 74.131(3.5)\\ &= 428.62\end{aligned}$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $X_2 = 3.5$

Coefficient of multiple determination (r^2):

- The coefficient of determination is the proportion of the total variation in Y that can be explained by the variation in all X variables together

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

  - Recall: r^2 value is between 0 and 1
    - Value of 0 – the model explains none of the total variation in Y
    - Value of 1 – the model explains all of the total variation in Y with all the independent X variables
- On Excel:

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted r^2 (r^2_{adj}):

- r^2 never decreases when a new X variable is added to the model (any new X variables added, regardless of how insignificant, will never decrease the amount of total variation in Y explained by existing X variables, but they may increase the r^2 value by random chance, which can distort the model)
  - This can be a disadvantage when comparing two or more models because the higher r^2 value may be distorted to suggest that one model is better than the other when in reality, it isn't truly better
- Net effect of adding a new variable:
  - We lose a degree of freedom when a new X variable is added
  - We thus need to determine whether the new X variable adds enough explanatory power to offset the loss of one degree of freedom
- r^2_{adj} shows the proportion of the total variation in Y that can be explained by the variation in all X variables together, adjusted for the number of X variables used

$$r^2_{adj} = 1 - \left[(1 - r^2)\left(\frac{n-1}{n-k-1}\right)\right]$$

  - (where n = sample size, k = number of independent variables)
  - This penalizes excessive use of "unimportant" independent variables
    - Increasing k will decrease the adjusted r^2
  - This is useful when comparing among competing models that use a different number of regressors
    - Now can choose the model with the highest adjusted r^2 value as the most relevant model
  - r^2_{adj} is always smaller than r^2
  - r^2 only between 0 and 1 but adjusted r^2 can be negative
  - Another r^2_{adj} formula:

$$r^2 = 1 - \frac{SSE}{SST}$$

$$r^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSE}{SST}$$

$$= 1 - \frac{\hat{\sigma}^2_\varepsilon}{s^2}$$

Smaller than r^2, since n-1 > n-k-1

Compares two unbiased estimates of variation and asks by how much is the estimate of Var(Y | X) < Var (Y)

    - NB adjusted r^2 can be negative!
- Excel output:
  -

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$r^2_{adj} = .44172$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F-test for overall significance of the model:

- Use F tests for looking at all X variable slopes together
  - This is used to test if any of the X variables are related to Y, or if none are
  - Uses a F statistic in a F test
- F distribution:
  - A F distribution is an extension of the Student-t distribution.
  - It has 2 types of degrees of freedom
    - 1. A numerator degrees of freedom (k)
    - 2. A denominator degrees of freedom (n – k – 1)
  - In the simplest case we have:
    - $F_{1,n-2} \equiv t^2_{n-2}$
  - In general, an F statistic can be defined as a quadratic function of a set of t statistics
  - Defined only for a positive domain (for values > 0)
    - Exam note: Rejection region only for the upper-end
- Hypotheses:
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)
    H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)
- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

where F_{STAT} has numerator d.f. = k and denominator d.f. = (n – k - 1)

Can you see a connection to r^2
MSR = Regression Mean Square
MSE = Mean Square Error

  - Connection to r^2:
    - When r^2 increases, the F statistic will also increase → F stat will more likely to be in the rejection region, meaning we are more likely to reject the null hypothesis (shows how our model will be more useful when r^2 increases)
- Excel output:

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

- Results:

H_0: $\beta_1 = \beta_2 = 0$
H_1: β_1 and β_2 not both zero

$\alpha = .05$
$df_1 = 2$ $df_2 = 12$

Critical Value:
$F_{0.05} = 3.885$

$\alpha = .05$

0 F
Do not reject H_0 | Reject H_0
$F_{0.05} = 3.885$

Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 6.5386$$

Decision:

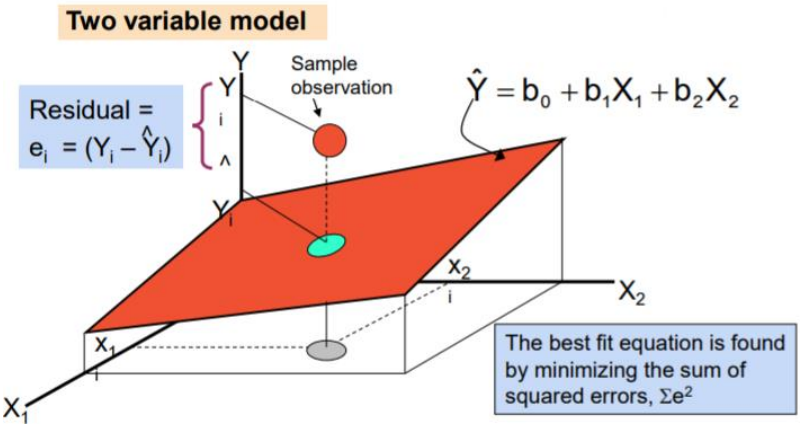
Since the observed p-value is 0.01201 (smaller than 0.05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

Residual analysis to verify the assumptions for a multiple linear regression model:

- Residuals in MLR equation:



  - Recall the formula for residuals from the estimated regression model:
    - $e_i = (Y_i - \hat{Y}_i)$
- Required assumptions about ε in a MLR model:
  - The relationship between X_k and Y is linear
  - Error values are statistically independent
  - Errors are normally distributed for any given X_k value for all k (this assumption is only required for small samples, or for prediction intervals)
  - The probability distribution of the errors has a constant variance for any given X_k value for all k
- Residual plots used in multiple regression:
  - Residuals vs. $\hat{Y}_i$ (predicted)
  - Residuals vs. X_{1i}
  - Residuals vs. X_{2i}
  - Residuals vs. time (if time series data)

  Use the residual plots to check for violations of regression assumptions

t-test for individual X variable's significance:

- We use a t test to test for the significance of an individual slope
  - This tests if there is a linear relationship between the variable X_k and Y, holding constant the effects of other X variables
- Hypotheses:
  - $H_0: \beta_j = 0$ (no linear relationship)
  - $H_1: \beta_j \neq 0$ (linear relationship does exist between X_j and Y)
- Test statistic:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

(df = n – k – 1)

  - Exam note: t_{STAT} has the DF = n – k – 1 for MLR (rather than just n – 2 for SLR), where k represents the number of slope parameters we need to estimate in our sample regression function

- Excel output:

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

- Results:

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

d.f. = 15-2-1 = 12

α = .05

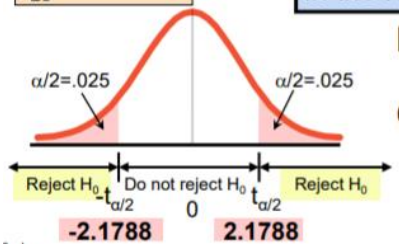
$t_{\alpha/2}$ = 2.1788

From the Excel output:

For Price t_{STAT} = -2.306, with p-value .0398

For Advertising t_{STAT} = 2.855, with p-value .0145

The test statistic for each variable falls in the rejection region (p-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at α = .05

Exam note: We must do the t tests separately for each variable, whilst holding all other variables constant when testing for our variable

- Confidence interval estimate for the population slope, β_j:

$$b_j \pm t_{\alpha/2} S_{b_j}$$ where t has (n – k – 1) d.f.

	Coefficients	Standard Error
Intercept	306.52619	114.25389
Price	-24.97509	10.83213
Advertising	74.13096	25.96732

Here, t has (15 – 2 – 1) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales:

-24.975 ± (2.1788)(10.832)

So the interval is (-48.576 , -1.374)

(This interval does not contain zero, so price has a significant effect on sales)

	Coefficients	Standard Error	...	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	...	57.58835	555.46404
Price	-24.97509	10.83213	...	-48.57626	-1.37392
Advertising	74.13096	25.96732	...	17.55303	130.70888

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price, holding the level of advertising constant

Using “dummy” variables:

- A dummy variable is a categorical independent variable with two or more levels:
  - E.g. yes or no, on or off, male or female → coded as 0 or 1
  - If more than two levels, the number of dummy variables needed is (number of levels - 1)
- Dummy-variable model example (with 2 levels):

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

Let:

Y = pie sales

X_1 = price

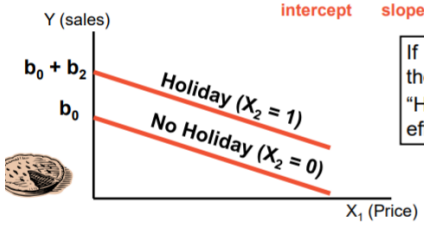
X_2 = holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

  - X_2 is the dummy variable in this example

$\hat{Y} = b_0 + b_1X_1 + b_2(1) = (b_0 + b_2) + b_1X_1$ **Holiday**

$\hat{Y} = b_0 + b_1X_1 + b_2(0) = b_0 + b_1X_1$ **No Holiday**

Different intercept; Same slope



If H_0: $\beta_2 = 0$ is rejected, then “Holiday” has a significant effect on pie sales

  - Exam note: A categorical variable does not affect the slope of our SRF, but will create a difference intercept when it the variable is ‘switched on’ i.e. X_2 = 1, meaning a holiday occurred during the week
- Interpreting the dummy variable coefficient (with 2 levels):

Example: $\widehat{Sales} = 300 - 30(Price) + 15(Holiday)$

Sales: number of pies sold per week
Price: pie price in $

Holiday: 1 If a holiday occurred during the week
0 If no holiday occurred

b_2 = 15: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

- Dummy-variable model example (with more than 2 levels):

The number of dummy variables is one less than the number of levels

Example:

Y = house price ; X_1 = square feet

If style of the house is also thought to matter:

Style = ranch, split level, colonial

Three levels, so two dummy variables are needed

- Let "colonial" be the default (baseline) category, and let X_2 and X_3 be used for the other two categories:

  Y = house price
  X_1 = square feet
  X_2 = 1 if ranch, 0 otherwise
  X_3 = 1 if split level, 0 otherwise

  The multiple regression equation is:

  - $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$
- We can only have 2 dummy variables here so let colonial be the default category

  Consider the regression equation:

  $\hat{Y} = 20.43 + 0.045X_1 + 23.53X_2 + 18.84X_3$

  For a colonial: $X_2 = X_3 = 0$

  $\hat{Y} = 20.43 + 0.045X_1$

  For a ranch: $X_2 = 1; X_3 = 0$

  $\hat{Y} = 20.43 + 0.045X_1 + 23.53$

  With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a colonial.

  For a split level: $X_2 = 0; X_3 = 1$

  $\hat{Y} = 20.43 + 0.045X_1 + 18.84$

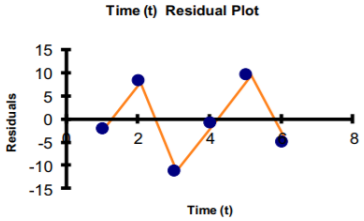
  With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a colonial.

  - Since colonial is the baseline category, b_0 represents the estimated intercept value (the estimated house price a colonial house that theoretically has 0 square feet)

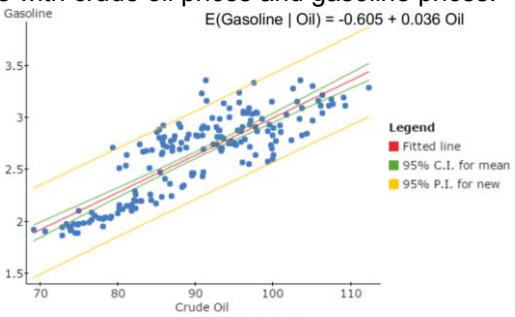
Autocorrelation:

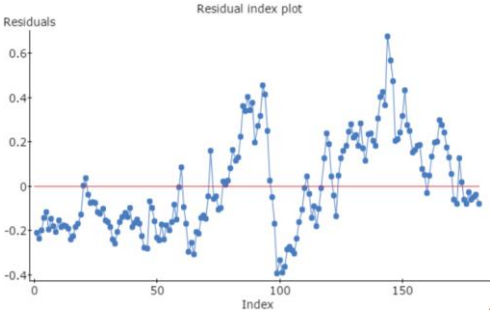
- Autocorrelation is correlation of the errors over time:

  Here, residuals show a repeating pattern (not random.) Such patterns are a sign of positive autocorrelation

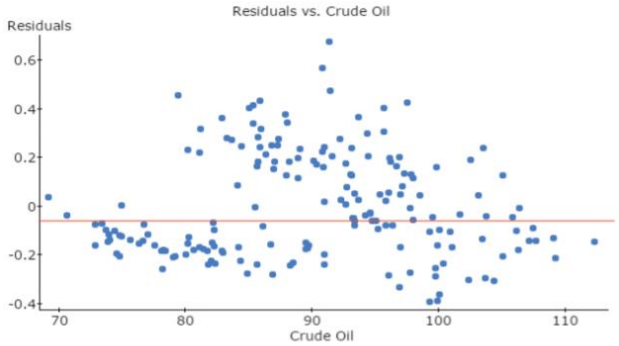
  

  - The residuals suggest that our model violates the regression assumption that errors are random and independent → as a consequence, least squares method as a suitable estimation procedure
- Example with crude oil prices and gasoline prices:

- There is a run of many negative residuals in a row followed by many successive positive residuals in a row, which suggests strong autocorrelation (that the errors are not independent)
- Residuals vs X variable (crude oil):



  - NB variance also seems to be non-constant
  - We can also see there is a non-constant hows a change in the variance of errors (more variance around the middle of crude oil)

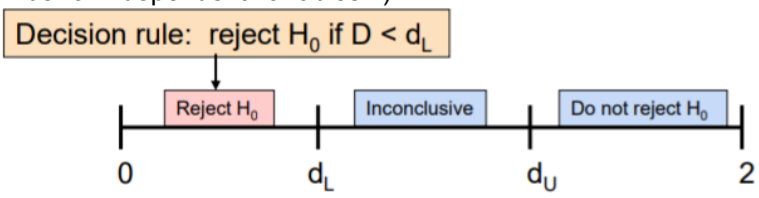
- Measuring autocorrelation – the Durbin-Watson statistic:
  - Used when data are collected over time to detect if autocorrelation is present
  - Autocorrelation exists if errors in one time-period are related to errors in another period
  - Existence of autocorrelation violates the independence assumption
  - The Durbin-Watson statistic:

H_0: residuals are not correlated
H_1: autocorrelation is present

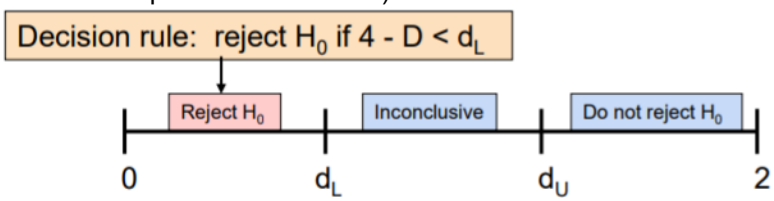
$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

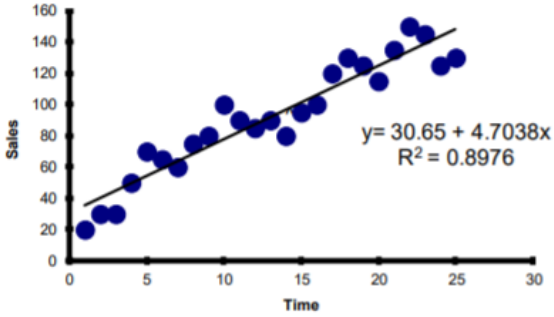
  - The possible range is $0 \le D \le 4$
  - D should be close to 2 if H_0 is true
  - D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation
  - Note that this statistic only looks at autocorrelation in successive time periods
  - D < 2 signals positive autocorrelation as errors must be closely linked successively for a low D value, as shown by the D formula

- Testing for positive autocorrelation:
  - 1. Calculate the Durbin-Watson test statistic (D)
  - 2. Find the values d_L and d_U from the Durbin-Watson table (for sample size n and number of independent variables k)
  - 3.

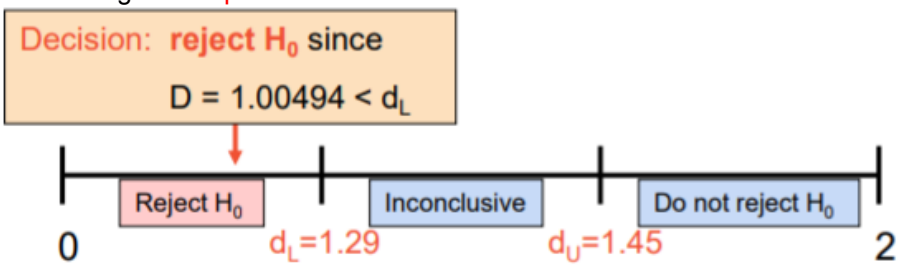


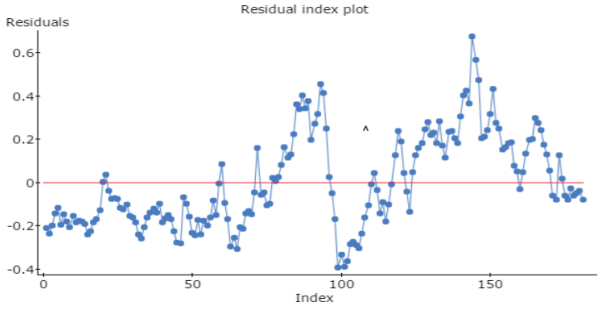
- Testing for negative autocorrelation:
  - 1. Calculate the alternative Durbin-Watson test statistic = 4 – D
  - 2. Find the values d_L and d_U from the Durbin-Watson table (for sample size n and number of independent variables k)
  - 3.



- Autocorrelation test e.g.1:
  - Suppose we have the following time series data:
    - 

  - Durbin-Watson statistic calculations:
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$
  - Testing for positive autocorrelation:
    - We are testing for positive autocorrelation since D < 2
  - Rejection regions:
    - Here, n = 25 and there is k = 1 one independent variable
    - Using the Durbin-Watson table, d_L = 1.29 and d_U = 1.45
  - Result:
    - D = 1.00494 < d_L = 1.29, so reject H_0 and conclude that statistically significant positive autocorrelation exists
  - 

- Autocorrelation test e.g.2:
  - Suppose we have the following data:
    Recall the residuals from the Oil vs Gas price data:
    - 

- Durbin-Watson statistic calculations:

Here n = 181

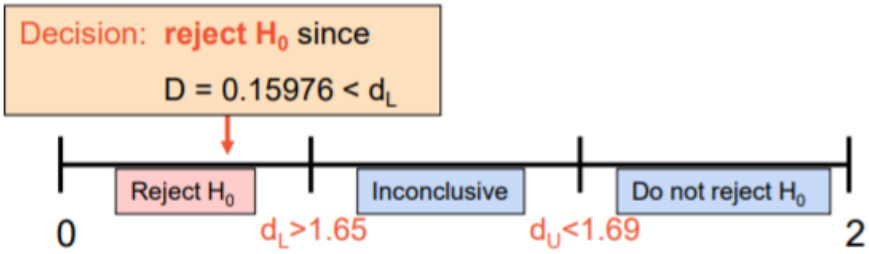
Index	e(i)-e(i-1)	(e(i)-e(i-1))^2	ei^2
1			
2	-0.02475	0.00061245	0.055792
3	0.037099	0.00137636	0.039642
4	0.054869	0.003010626	0.020803
5	0.028503	0.000812405	0.013394
6	-0.07969	0.006349754	0.038188

Excel output:

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	1.315754
Sum of Squared Residuals	8.235871
Durbin-Watson Statistic	0.159759

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{1.315754}{8.235871} = 0.159759$$

- Testing for positive autocorrelation:
  - We are testing for positive autocorrelation since D < 2
- Rejection regions:
  - Here, n = 181 and there is k = 1 one independent variable
  - Using the Durbin-Watson table, d_L > 1.65 and d_U < 1.69
- Result:
  - D = 0.15976 < (d_L > 1.65), so reject H_0 and conclude that statistically significant positive autocorrelation exists



-

Pitfalls of regression analysis:

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

Strategies for avoiding the pitfalls of regression:

- Start with a scatter plot of Y vs. X to observe possible relationship
- Perform residual analysis to check the assumptions
  - Plot the residuals vs. X (or vs. predicted Y values) to check for violations of assumptions such as a violation of homoskedasticity for example
  - Use a histogram, boxplot, or normal probability plot of the residuals to uncover possible non-normality (if low sample size or prediction intervals needed)
- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range