Notes

Tuesday, 26 July 2022 3:22 PM

STATISTICS AND DATA

Relevance of statistics

- Enables businesses to make better decisions
- Evidence based management
- Differentiate between sound statistical conclusions and questionable conclusions, e.g.
  - Based on one data point
  - Independent events
  - Sample bias
  - Selective data
  - Correlation to causation fallacy

What is statistics?

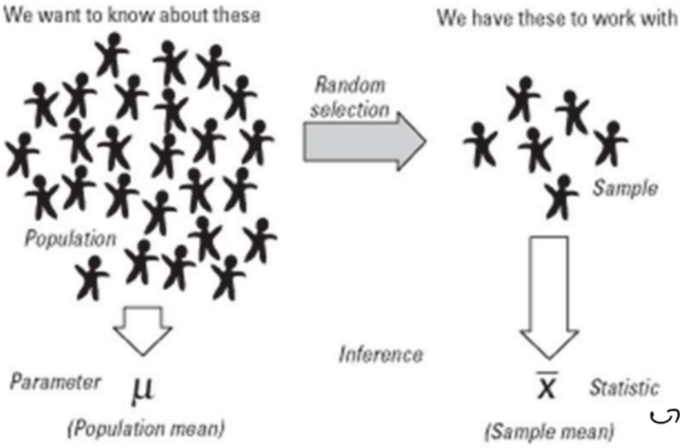
- Statistics - methodology of extracting useful information from a data set
- To do good statistical analysis, you must:
  1. Find the right data (complete and representative)
  2. Use appropriate statistical tools
  3. Clearly communicate the numerical information into written language

- Data: raw facts, measures of certain phenomena
- Information: facts in a form suitable for businesses to base decisions on
  - Must be accurate, relevant, timely and complete

- Two branches of statistics:

1. Descriptive statistics
   - Summary of aspects of data
   - Collecting, organising and presenting data
   - Numerical and graphical tools (charts and tables)
2. Inferential statistics
   - Drawing conclusions about population based on sample data
   - Organisational research is mostly based on sample data

- Inferential procedures



  - Population: all items of interest in a statistical problem
    - Population parameters often unknown
  - Sample: subset of the population
    - Sample statistic calculated from sample and used to make inferences about the population
    - Needs to be representative of the population for sample results to be generalisable
    - Representative when random sampling used

Sampling

- Probability samples
  1. Simple random sampling
     - Every member of population has equal chance of being selected
  2. Cluster sampling
     - Population divided into clusters, then randomly sampled within chosen clusters
     - E.g. high schools, primary schools
  3. Stratified sampling
     - Number of observations per stratum is proportional to the stratum's size in the population
     - Within each stratum, observations randomly selected
     - E.g. Full-time, part-time and casual employees; age

- Non-probability samples
  1. Non-probability sample
     - Little or no attempt made to get representative cross-section of population
     - E.g. Snowball sampling - study participants recruit future subjects
  2. Convenience sample
     - Respondents convenient or readily accessible to researcher

  - Need for sampling
    - Conducting a census of the whole population is expensive and takes time
    - May be impossible or not practicable

  - Central Limit Theorem
    - As sample size increases, you are more likely to get a sample mean closer to the population mean (representative)
    - For most things, if sample size is 30 or more then it forms a normal distribution

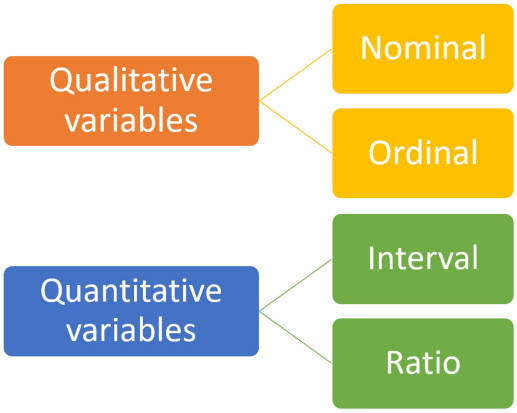
  - Bias: tendency of sample statistic to systematically over or underestimate a population parameter

- Data collection issues
  - Data accuracy (e.g. data entry issues)
  - Interviewer bias (e.g. leading questions, body language)
  - Nonresponsive bias (some of sample don't respond)
  - Selection bias
  - Observer bias (if someone knows they're being watched, they may change their behaviour)
  - Measurement error
  - Internal validity
    - Relates to how well an experiment's extraneous variables were controlled, e.g.
      - Randomly assign subjects to control and experimental groups
      - Match subjects in control and experimental groups in terms of age, gender, race, health
      - Blinding where possible
  - External validity
    - Relates to whether results can be generalised to population
    - If results be replicated at a different location, sample or time
    - If sampling done well, sample should represent population and results should generalise

Data types

- Variable: attribute or measurement on members of population/sample (field in database)
- Observation: list of all variable values for a member (case, record)

1. Cross-sectional data
   - Data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time (snapshot)
2. Time series (longitudinal) data
   - Data collected by recording a characteristic of a subject over several time periods (including daily, weekly, monthly, quarterly, annual observations)



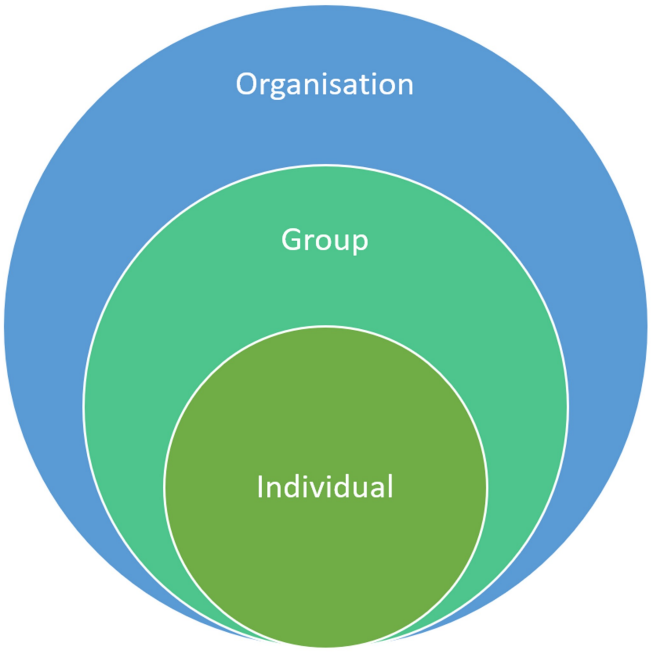
1. Qualitative
    - Descriptive information (labels and names)
2. Quantitative
    - Numerical information
        - Discrete (countable, but not necessarily whole numbers)
        - Continuous (uncountable, measured, unlimited number of values between intervals)

- Scales of measurement

1. Nominal
    - Data are simply categories for grouping data
    - No ordering to responses, differ by name or label only
    - Least sophisticated level of measurement
    - E.g. Gender, company names
2. Ordinal
    - May be categorised and ranked with respect to some characteristic or trait
    - Difference between categories meaningless as they are arbitrary
    - E.g. Rating of excellent, good, fair, poor
3. Interval
    - May be categorised and ranked
    - Differences between intervals are equal and meaningful. Addition and subtraction are meaningful
    - No meaningful ratios, as 0 is arbitrarily chosen (i.e. no absolute 0)
    - E.g. Temperature °C
4. Ratio
    - Categorised, ranked, difference between intervals meaningful
    - Meaningful ratios, absolute 0 exists
    - Strongest level of measurement
    - E.g. Weight, time, distance, sales, profits, inventory levels

	Nominal	**Ordinal**	**Interval**	**Ratio**
Categorised	Y	Y	Y	Y
Ranked	N	Y	Y	Y
Differences between intervals equal and meaningful	N	N	Y	Y
Meaningful ratios, absolute 0	N	N	N	Y
Level of analysis	Lowest (basic)	Higher (mid-level)	Highest (complete)	Highest (complete)

- Data levels

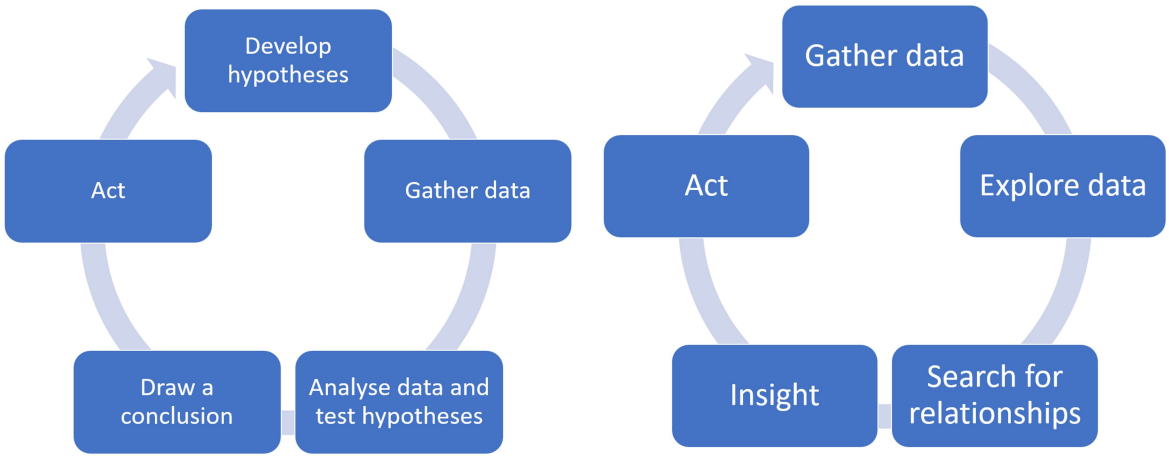


Big data

- Relates to large data sets, but is mostly data that is difficult to manage, process and analyse using traditional data analysis techniques
    - No defined size (data is produced exponentially, so what is big today may be small tomorrow)
    - Not necessarily complete (population) data

- Differences between traditional data and big data

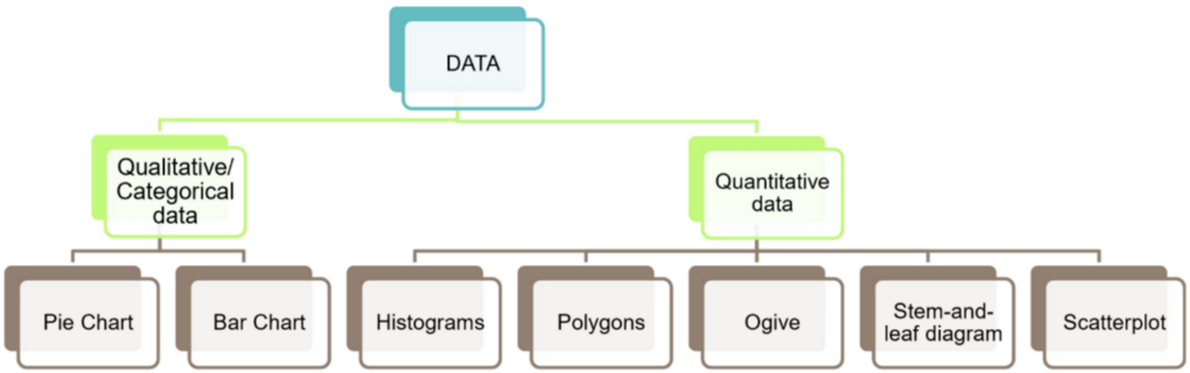
Traditional	Big Data
Analyse subsets of data	Analyse ALL available data



- Importance of big data
  - Has value only if we convert data into information and use it for decision making
  - 95% is unstructured data, so new approaches (mixed-methodologies) are needed
  - It's everywhere
  - Use of large scale data to predict human behaviour is gaining currency in business and government policy practice, and the intersect of physical and social science

- Challenges
  - Technological
    - Multidisciplinary teams needed as the key components of machine learning (expertise, computing power, data, algorithms) are not concentrated in one domain
    - Data scientists, business needs, etc.
    - Industry, academia, government play significant roles
  - Educational
    - Data literacy (everyone should care about quantitative sciences)
    - Ethical issues
  - Societal
    - Fairness, privacy, consent, cybersecurity
    - Trust, transparency, interpretability
    - Engagement of stakeholders
    - Open distribution of tools and knowledge
    - Living alongside machine learning

PRESENTING AND REPORTING DATA



Qualitative (categorical) data

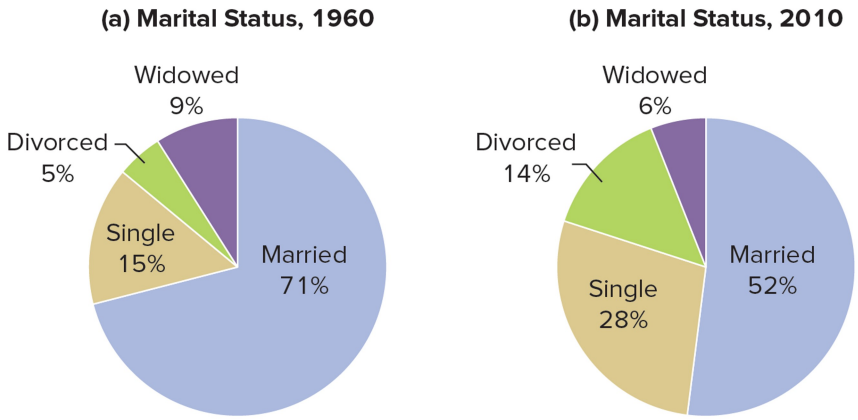
- Frequency distribution
  - Groups data into categories and records number observations in each category
  - Relative frequency - frequency divided by sample size, sum = 1, easier to compare different data sets
  - Percent frequency - relative frequency multiplied by 100

Weather	Frequency	Relative frequency	Percent frequency

Cloudy	1	1/28 = 0.0357	3.57%
Rainy	20	20/28 = 0.7143	71.43%
Sunny	7	7/28 = 0.25	25%
Total	28	1	100%

- Pie chart
  - Segmented circle
  - Relative frequency represented by segments of a circle
  - Aesthetically pleasing, more dramatic and simplifies interpretation
  - Good for a few categories

FIGURE 2.1 Pie charts for marital status



- Bar chart
  - Frequency or relative frequency represented by horizontal or vertical bars
  - Lengths proportional to values depicted
  - Displays data in a way to make it easy to compare categories and read data
  - Vertical bar chart or column chart (category on x-axis, frequency on y-axis)

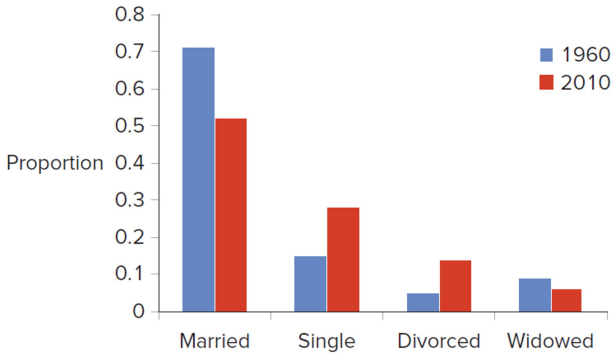


FIGURE 2.2
Marital status of U.S. adults, 1960 versus 2010

  - Horizontal bar chart (category on y-axis, frequency on x-axis)

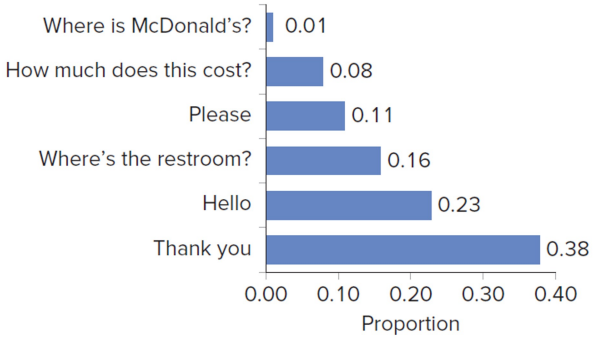
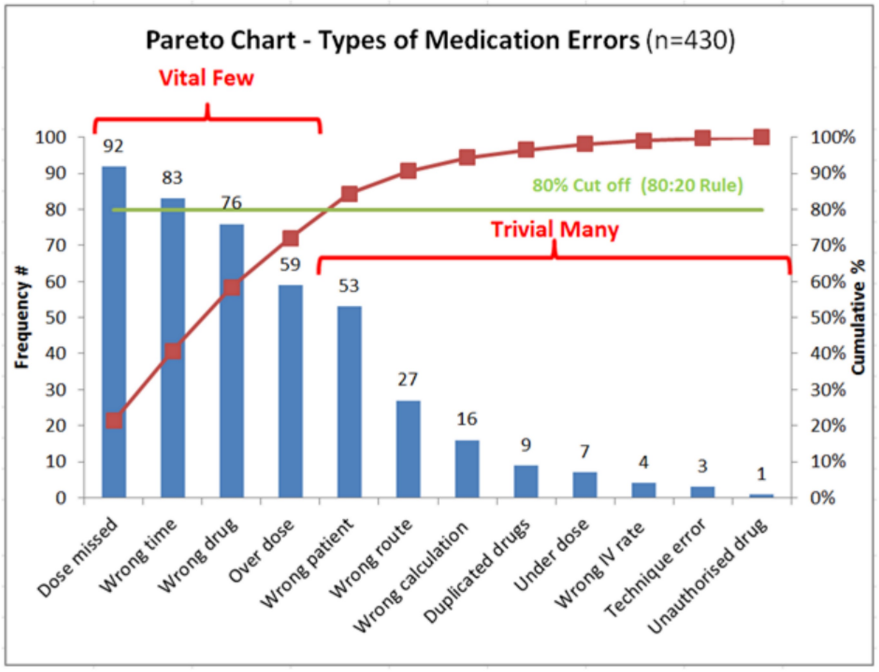


FIGURE 2.3
Results to question: "When traveling in a non-English-speaking country, which word or phrase is most important to know in that country's language?"

- Pareto diagram
  - Bar chart arranged with longest bars on the left and shortest bars on the right
  - Cumulative total represented by curved line

- 80% cut off line
- 80/20 rule (critical few vs trivial many) - roughly 80% of effects come from 20% of the causes
- Enables us to concentrate on factors that have the greatest impact



- To avoid misleading graphical displays:
  - Present data in clear, simple way. Strive for clarity
  - Axes should be clearly marked with numbers and clearly labelled
  - Bars should be same width in bar chart
  - Do not compress vertical axis, i.e. A very high value should not be given as an upper limit (underemphasises trends)
  - Do not stretch vertical axis (overemphasises trends)
  - Vertical axis should (in most cases) include zero
  - Avoid chart junk (anything that doesn't relate to data, e.g. Pictures)
  - Give chart a title
  - Interpret chart in the body of the report, don't assume reader understands
  - Include key for different categories

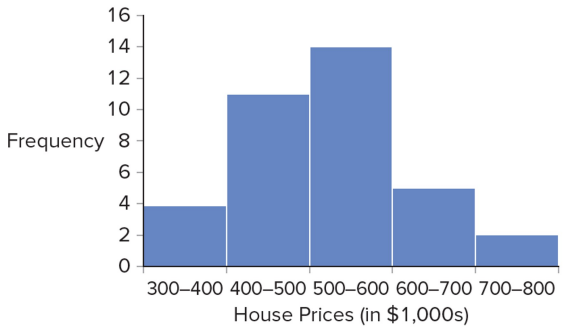
Quantitative data

- Frequency distribution
  - Groups data into intervals called classes (or bins), and records number of observations in each class
  - Data is more manageable, but some detail lost
  - Guidelines for constructing:
    - Classes are mutually exclusive (do not overlap)
    - Classes are exhaustive (total number of classes covers entire sample)
    - Class limits should be easy to recognise and interpret
    - Number of classes usually ranges from 5 to 20 (smaller data sets have fewer than larger)

- Cumulative frequency distribution
  - Number of observations fall below upper limit of a particular class
    - Sum successive relative frequencies, or
    - Divide cumulative frequency by sample size

- Histogram
  - Counterpart to vertical bar chart
  - Bar height represents class (or bin) frequency (or relative frequency)
  - Bar width represents class width
    - Level of detail determines how easy it is to see the shape of the data
  - Allows us to quickly see spread and shape of data

FIGURE 2.5
Frequency histogram for house prices



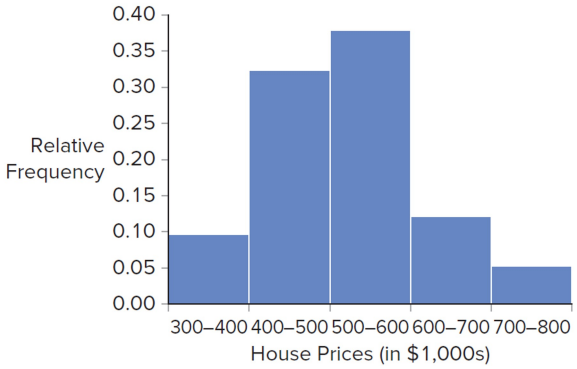
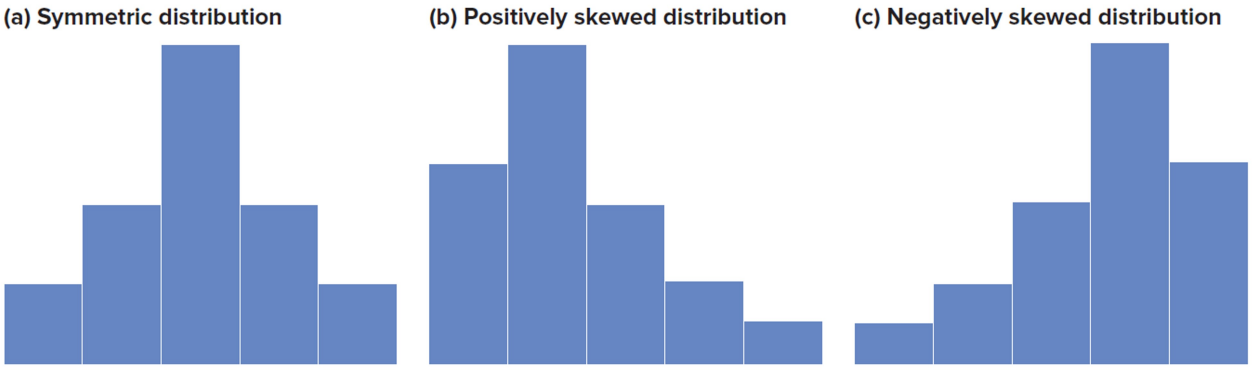


FIGURE 2.6 Relative frequency histogram for house prices

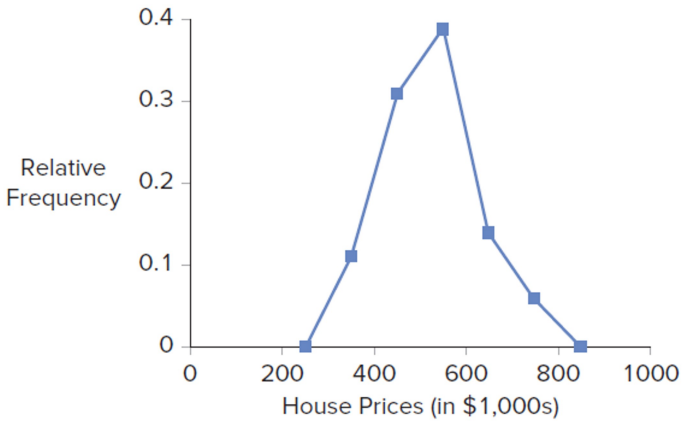
- Skewness
  - Symmetric
    - Mirror image on both sides of centre
    - Bell shaped curve, normal distribution
    - Mean = median
  - Positively skewed
    - Narrow tail to the right (i.e. Outliers to the right or positive)
    - Mean > median
  - Negatively skewed
    - Narrow tail to the left (i.e. Outliers to the left or negative)
    - Mean < median

FIGURE 2.7 Histograms with differing shapes



- Polygon
  - Plot class midpoints (medians) of each class on x-axis, associated frequency (or relative frequency) on y-axis
  - Neighbouring points connected with a straight line
  - To close off graph at each end, add one interval below the lowest interval and one above the highest interval, then assign value of 0
  - Easy to visualise difference between data sets

FIGURE 2.8
Polygon for the house-price data



- Ogive
  - Cumulative frequency (or relative frequency) distribution
  - Plot upper limit of the corresponding class on x-axis, cumulative frequency (or relative frequency) of each class on y-axis
  - Neighbouring points connected with a straight line
  - Close off only at lower end by intersecting x-axis with lower limit of first class

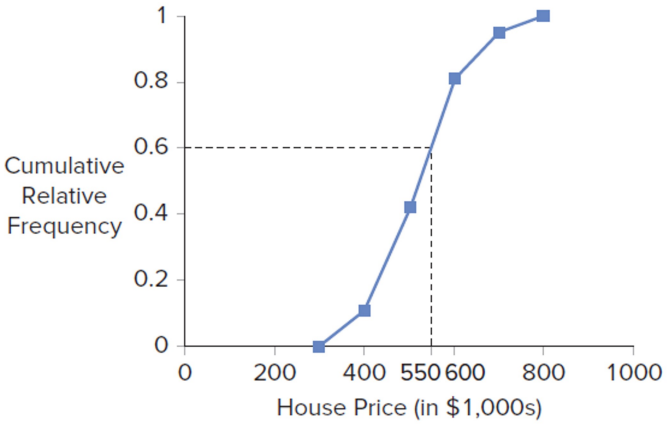


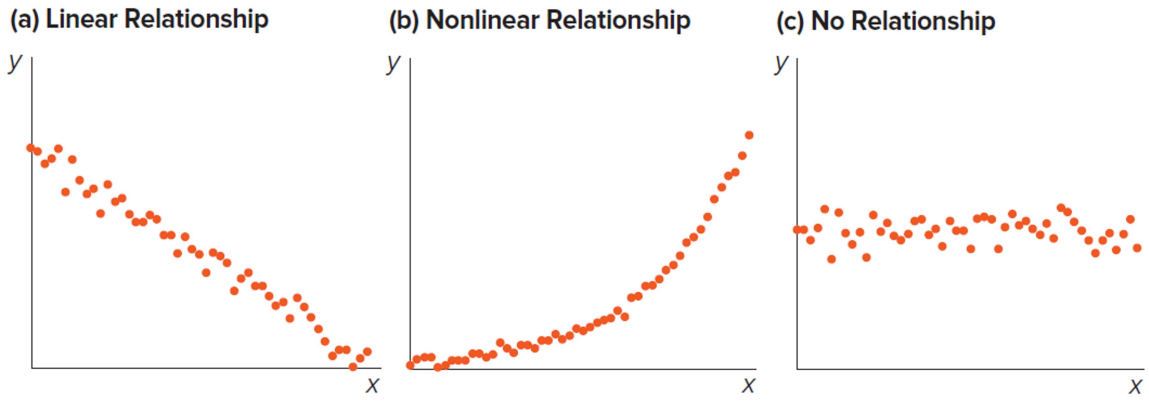
FIGURE 2.9
Ogive for the house-price data

- Stem-and-leaf diagrams
  - Separate each value of data set into:
    - Stem - leftmost digits
    - Leaf - last digit
  - Often preliminary step in data analysis, overall picture of where data centred and how dispersed from the centre

Panel C	
Stem	Leaf
3	6
4	
5	2 2 3 4 4 5 9
6	0 1 2 2 5 6 6 8
7	0 4 4 9
8	1 2 3 7
9	0

- Scatter plot
  - Determines if two variables are related
  - Independent variable on x-axis, dependent variable on y-axis

FIGURE 2.13 Scatterplots depicting relationships between two variables



a. Linear relationship (positive or negative)
b. Curvilinear
c. No relationship

Measures of central tendency

- Arithmetic mean or average

Sample mean

$$\overline{x} = \frac{\Sigma x_i}{n}$$

Population mean

$$\mu = \frac{\Sigma x_i}{N}$$

  - Sensitive to outliers

- Median
  - Middle value of data set
  - Not affected by outliers

- Mode
  - Most frequently occurring value in data set
  - For qualitative data, it's the only meaningful measure of central tendency
  - Can have no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal)
  - Usefulness diminishes if >3 modes

Percentiles and boxplots

- Generally, *p*th percentile divides data set into two parts
  1. Approximately *p* percent of observations have values less than the *p*th percentile
  2. Approximately (100 - *p*) percent of observations have values greater than the *p*th percentile
- Most meaningful for large data sets

- Calculating *p*th percentile
  1. Arrange data in ascending order
  2. Locate approximate position of percentile

     $L_p = (n+1)\frac{p}{100}$
  3. If L_p is an integer, then it denotes the location of the *p*th percentile

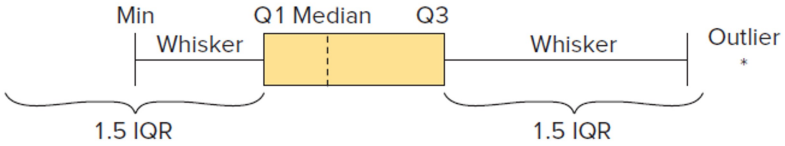
     E.g. If L_{20} is 2, then the 20th percentile is the second observation of the data set
  4. If L_p is not an integer, need to interpolate between two observations

     E.g. If L_{20} is 2.25, then the 20th percentile is 25% of the distance between the second and third observation

- The 25th percentile is also referred to as the first quartile (Q1), 75th percentile is the third quartile (Q3)

- Five-number summary is the Min, Q1, Median (or Q2), Q3 and Max
- Box plot displays the five-number summary
  - Useful for comparing similar information gathered at another place or time
  - Can identify outliers and skewness

FIGURE 3.3 A sample boxplot



- Interquartile range $IQR = Q3 - Q1$ (length of box)

- Outliers may arise from bad data (errors) or random variations
  - Outliers exist if they are >1.5 x IQR away from Q1 or Q3
  - Asterisk indicates outlier

- Informally gauges shape of distribution
  - Positively skewed - median left of centre, right whisker longer than left whisker
  - Negatively skewed - median right of centre, left whisker longer than right whisker

Measures of dispersion

- Gauge variability of data set

- Range

  $$Range = \text{Max} - Min$$

  - Simplest measure
  - Focuses on extreme values
  - Affected by outliers
  - E.g. Temperature forecast

- Variance

  Sample variance

  $$s^2 = \frac{\Sigma (x_i - \overline{x})^2}{n - 1}$$

  Population variance

  $$\sigma^2 = \frac{\Sigma (x_i - \overline{x})^2}{N}$$

  - Squared units
  - n -1 is an adjustment factor for samples (mathematicians determined this results in an unbiased answer closer to population value)
  - Emphasises larger differences than smaller ones

- Standard deviation

  Sample standard deviation

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

  Population standard deviation

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

  - Same units as the data
  - With financial data, is most common measure of risk

- Coefficient of variation (CV)
  - Relative measure of dispersion
  - Adjusts for differences in the magnitudes of means
  - Unitless, allows direct comparisons of mean-adjusted dispersion across different data sets

  $$Sample\ CV = \frac{s}{x}$$

  $$Population\ CV = \frac{\sigma}{\mu}$$

  - Sometimes multiplied by 100 to return a percentage
  - Higher is more disperse

- Sample data set

	Age	Height	Work Experience	Travel Time to UWA	Comfort with Analytics
Mean	24.891	170.685	2.191	28.413	6.307
Standard Error	0.407	0.938	0.328	1.862	0.179
Median	24	171	1	25	6
Mode	23	163	0	15	6
Standard Deviation	3.904	8.995	3.148	17.856	1.717
Sample Variance	15.241	80.916	9.911	318.839	2.947
Kurtosis	3.942	-0.685	6.357	1.650	1.474
Skewness	1.896	-0.166	2.235	1.195	-0.467
Range	19	41	17	89	9
Minimum	20	150	0	1	1
Maximum	39	191	17	90	10
Sum	2290	15703	201.6	2614	580.25
Count	92	92	92	92	92

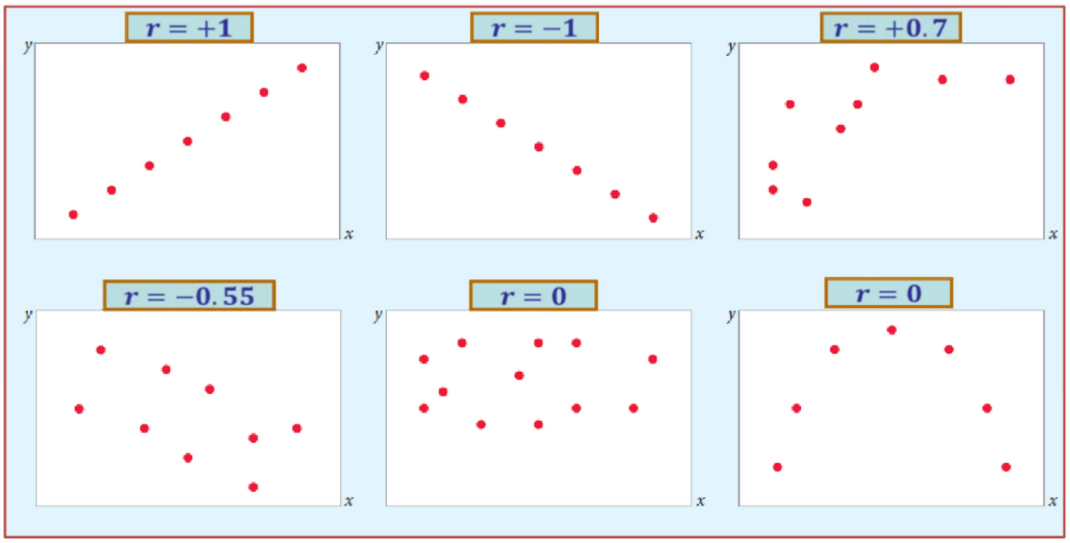
- Kurtosis - how tall and skinny or fat and flat the data set is
- Skewness - 0 is evenly distributed, positive implies outliers to the right tail, negative implies outliers to the left tail

Correlation coefficient

- Sample r_{xy} or population ρ_{xy} (rho)
- Describes both direction and strength of linear relationship between x and y
- Scatterplot visually displays this relationship

- Interpreting correlation coefficient
  - Sign (+/-) indicates direction of relationship (direct/inverse)
  - Value indicates strength of relationship. As a rule of thumb:

$\|0.71\| - \|1.00\|$	Strong
$\|0.41\| - \|0.70\|$	Moderate
$\|0.01\| - \|0.40\|$	Weak
0	No relationship



  - When r = 0 - as x increases, y stays the same
  - r only measures linear correlation, thus r = 0 for curvilinear relationship

Difference between population and samples

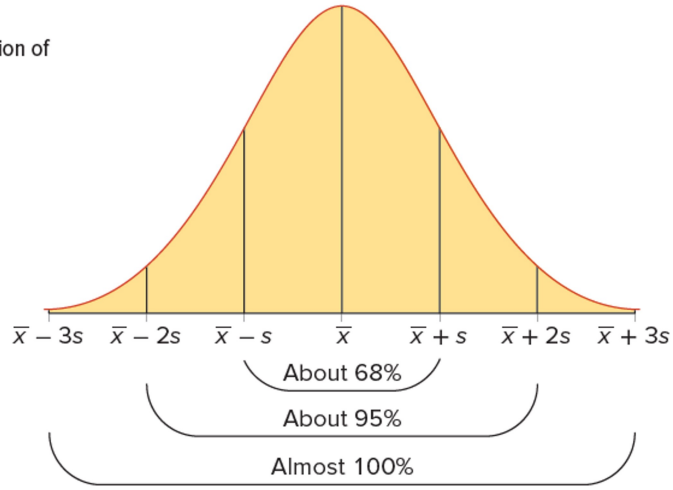
Measure	Population parameter	Sample statistic
Mean	μ	$\overline{x}$

Variance	σ^2	s^2
Standard deviation	σ	s
Correlation	ρ	r
Size	N	n

Empirical rule

- 68-95-99.7 rule

FIGURE 3.5
Graphical description of the empirical rule



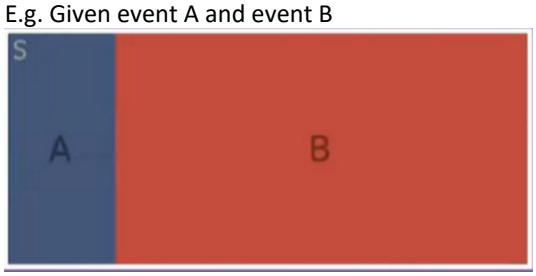
- Given sample mean $\bar{x}$ and sample standard deviation s
  - Approximately 68% of all observations fall within $\bar{x} \pm s$
  - Approximately 95% of all observations fall within $\bar{x} \pm 2s$
  - Approximately 99.7% of all observations fall within $\bar{x} \pm 3s$

PROBABILITY

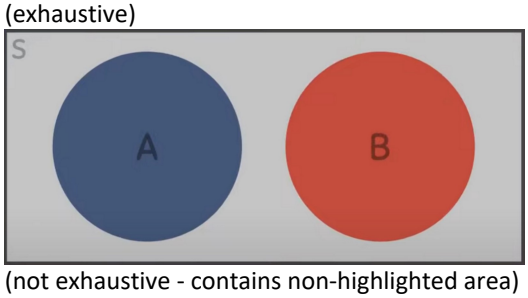
- Probability: likelihood an uncertain event will occur (0 to 1)
  - 0 - impossible events
  - 1 - definite events

- Experiment: process that produces outcomes for uncertain events
- Sample space *(S)*: collection of all possible experimental outcomes
- Event: subset of sample space
  - Simple event if only one outcome

  - Exhaustive: events that contain all outcomes in sample space
    E.g. Given event A and event B

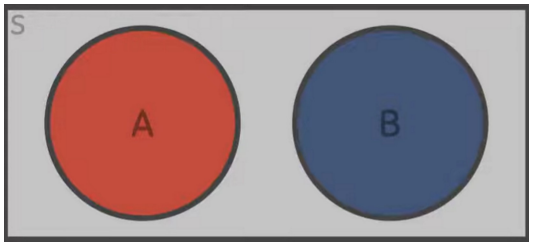


(exhaustive)



(not exhaustive - contains non-highlighted area)

  - Mutually exclusive: do not share any common outcomes



- $A \cup B$ - union (or)
- $A \cap B$ - intersection (and)
- A^c or $\overline{A}$ - complement (not)

Assigning probabilities

- For any event A, $0 \leq P(A) \leq 1$
- For mutually exclusive and exhaustive events, sum of probabilities = 1

1. Subjective
   - Based on personal and subjective judgement (knowledge and experience)
2. Objective
   a. Empirical
   - Relative frequency of occurrence (research)

   b. Classical
   - Logical analysis (understand process and reasoning about the problem)

- Empirical probability approaches classical probability if experiment is run a very large number of times (law of large numbers)

Visualising events

- Contingency table
  - Frequencies for two qualitative variables, x and y
  - Each cell represents a mutually exclusive combination of the pair of x and y values
  - Sample size = total of row (or column) totals

TABLE 4.4a A Contingency Table Labeled Using Event Notation

	Brand Name			
Age Group	B_1	B_2	B_3	Total
A	174	132	90	396
A^c	54	72	78	204
Total	228	204	168	600

- Joint probability table
  - Joint probabilities represented by each cell, e.g. $P(A \cap B_1)$
  - Marginal probabilities, e.g. $P(A)$

TABLE 4.4b Converting a Contingency Table to a Joint Probability Table

	Brand Name			
Age Group	B_1	B_2	B_3	Total
A	0.29	0.22	0.15	0.66
A^c	0.09	0.12	0.13	0.34
Total	0.38	0.34	0.28	1.00

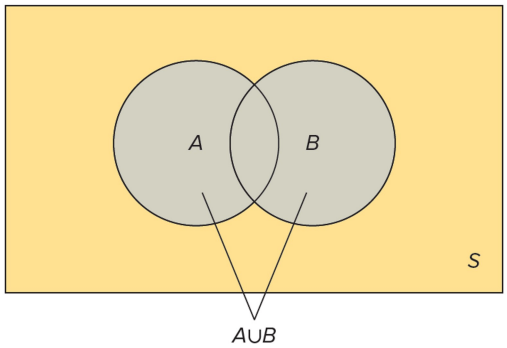
- Complement rule

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

- Addition rule
  - If events are NOT mutually exclusive:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

FIGURE 4.4
Finding the probability of the union of two events, $P(A \cup B)$



- If events are mutually exclusive, $P(A \cap B) = 0$

  $P(A \cup B) = P(A) + P(B)$

Conditional probability

- Probabilities always assessed relative to information currently available. As new information becomes available, it often changes

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

- Conditional is on the denominator

- Dependence
  - Independent - occurrence of one event does not affect probability of occurrence of other event
    - Not feasible to represent on Venn diagram

    $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$

  - Dependent - occurrence of one event related to probability of occurrence of the other

- Multiplication rule
  - Rearrange conditional probability formula to:

    $P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$

  - For independent events (can use this to check dependence as well)

    $P(A \cap B) = P(A) \cdot P(B)$

NORMAL DISTRIBUTION

Random variable (usually denoted by X): exact value unknown, but can be described by a set of possible values from random event

1. Discrete random variable: countable (finite) number of possible values
2. Continuous random variable: uncountable number of values in an interval (continuum)

- Every random variable is associated with probability distribution that describes variable completely
  - Probability mass function describes discrete random variables
    - P(X = x) where x refers to a possible outcome
  - Probability density function describes continuous random variables
    - P(a < X < b) where a and b refer to values of a specific interval
    - Probability is area under the curve between points a and b
  - Cumulative distribution function may be used to describe both discrete and continuous random variables
    - P (X ≤ x) where x refers to an upper limit

- Discrete random variables
  - E(X) - expected value of random variable X

    $E(X) = \Sigma\, xP(x)$
    - Where x = possible value of random variable X, and P(x) is probability of random variable having value x, i.e. Weighted average of all possible values of X
    - Same as population mean, μ, but not called mean as expected value indicates uncertainty of value
    - May not equal most probable value

  - SD(X) - standard deviation of random variable X

    $SD(X) = \sqrt{\Sigma x(-E(X)x)^2 P(x)}$

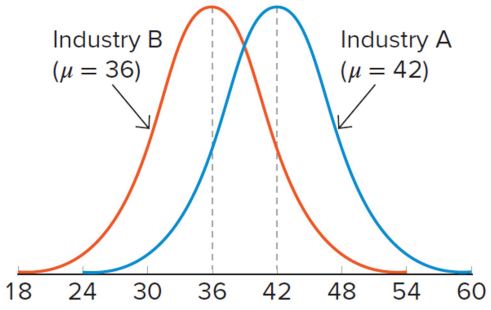
    $Var(X) = SD(X)^2$

- Continuous random variables
  - Probability random variable assumes particular value x is 0, i.e. P(X = x) = 0

Normal distribution

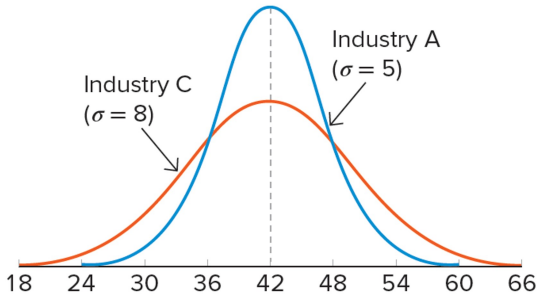
- Characteristics of normal distribution
  - Bell-shaped
  - Symmetric about its mean
    - Mean = mode = median
  - Asymptotic - tails get close to horizontal axis but never touch it (values range from negative infinity to positive infinity)
    - In reality often touches horizontal axis at ± 6 standard deviations
  - Completely described by two parameters - μ and σ^2
    - Changes in mean (same shape, different placement)

      FIGURE 6.4 Normal probability density function for two values of μ along with $\sigma = 5$

    - Changes in standard deviation (different shape, same placement)

      FIGURE 6.5 Normal probability density function for two values of σ along with $\mu = 42$

  - Closely approximates probability distribution of wide range of random variables
  - Statistical inference generally based on assumption of normality

- Standardising normal variables

$$z = \frac{x - \mu}{\sigma}$$

  - Z value measures distance of value x from mean in terms of standard deviations
    - Positive z (z > 0) indicates how many standard deviations corresponding x lies

above mean
    - Zero z (z = 0) indicates x = mean
    - Negative z (z < 0) indicates how many standard deviations x lies below mean
  - Z values can also detect outliers (more than 3 or less than -3)

- Standard normal (Z) distribution
  - Normal distribution where $\mu = 0$ or E(Z) = 0
  - and $\sigma = 1$ or SD(Z) = 1

  - Standard normal table (z-table)
    - Gives probabilities (area under the curve) for positive and negative values of z
    - Random variable Z is symmetric about mean of 0, so P(Z < 0) = P(Z > 0) = 0.5
    - For any z value > 3.99, it's acceptable to treat P(Z ≤ z) = 1

  - To calculate probability of given z value:
    1. Draw probability distribution and shade area for calculation
    2. Transform normally distributed random variable into standard normal random variable
    3. Use z table to obtain P(Z < z), by reading down z column first then across the top
    4. Probabilities given for P(0 ≤ z), so may need to subtract or add 0.5

Population and sampling distributions

- One population, but many possible samples of given size (n) can be drawn
  - Population parameter is a constant, but value may be unknown
  - Sample statistic is random variable whose value depends on chosen random sample

- Sampling distribution of the means
  - Each random sample drawn has sample mean $\overline{X}$ which provides estimate of population mean μ
    - Drawing many samples of size n results in many different sample means, one for each sample
    - Sampling distribution of mean $\overline{X}$ is probability distribution of all possible samples of given size from population
    - Average value of sample means = population mean, i.e. $E(\overline{X}) = \mu$ (unbiased estimator)

  - Standard error is standard deviation of sampling distribution for all samples of size n

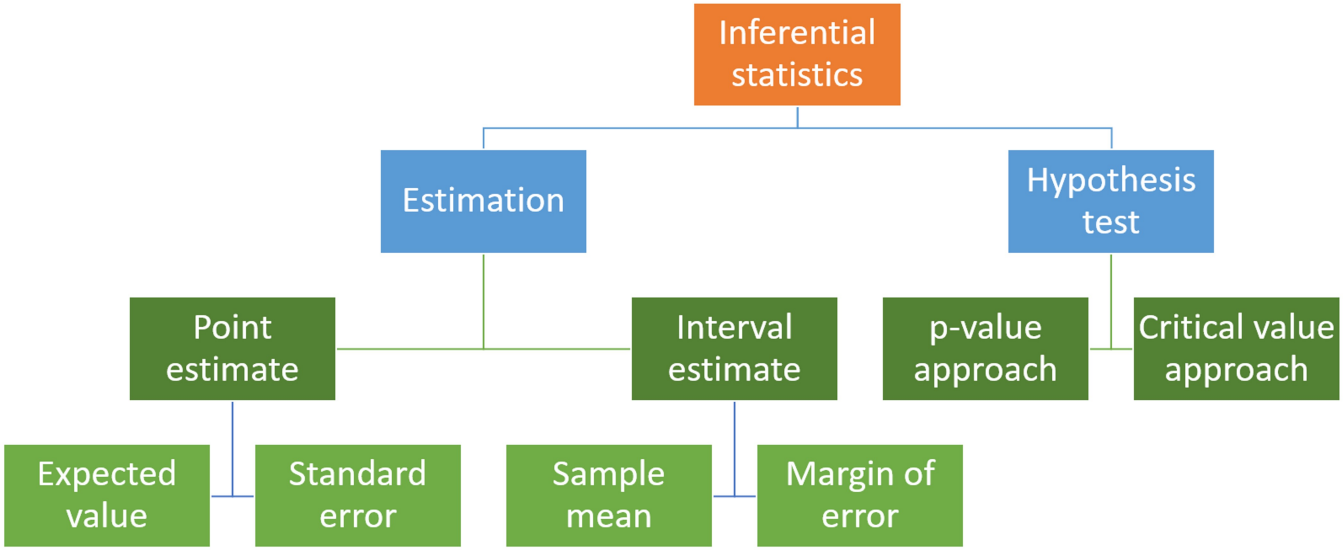
    $$\text{se}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

    - Variability between sample means is less than variability between observations
    - As sample size increases, standard error decreases (larger sample sizes approximate μ better)
  - Variance of $\overline{X}$

    $$Var(\overline{X}) = \frac{\sigma^2}{n}$$
  - Estimation error is difference between point estimate and true value of population parameter being estimated

- Sampling from normal distribution
  - For any sample size n, sampling distribution of $\overline{X}$ is normal if population X from which sample drawn is normally distributed
  - Otherwise if sample is sufficiently large (generally n ≥ 30), then sampling distribution will be normal (as per Central Limit Theorem)
  - If $\overline{X}$ is normal, can transform it to standard normal random variable

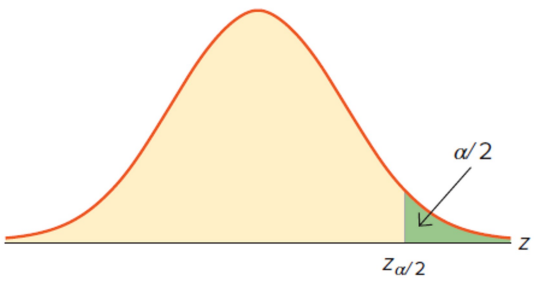
    $$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$



Interval estimation

- Confidence interval - range of values that, with a certain level of confidence, contains population parameter of interest
  - Provides additional information about variability of estimate
  - Interpretation, e.g. 95% confidence interval
    - For 95% of samples, formula produces interval that contains μ
    - Report with 95% confidence that μ lies in given interval
    - Not correct to say there is 95% chance that μ lies in given interval as it either falls in the interval (probability = 1) or does not fall in interval (probability = 0)
  - Point estimate ± margin of error
    (sample mean) (desired confidence level * standard error)
  - Factors influencing width of confidence level
    - Greater σ, wider interval
    - Smaller sample size n, wider interval
    - Greater confidence level, wider interval
    - Wider interval, lower precision

- When σ is known
  - α denotes allowed probability of error
    - Confidence coefficient = 1 - α
    - Confidence level = 100(1 - α)%
  - Confidence interval
    $$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
  - $z_{\alpha/2}$ associated with probability of α/2 in upper tail of standard normal distribution



  - For 95% confidence interval, α = 0.05, α/2 = 0.025, $z_{\alpha/2} = z_{0.025} = 1.96$
  - For 90% confidence interval, α = 0.10, α/2 = 0.05, $z_{\alpha/2} = z_{0.05} = 1.645$
  - For 99% confidence interval, α = 0.01, α/2 = 0.005, $z_{\alpha/2} = z_{0.005} = 2.576$

- When σ is unknown
  - t Distribution
    - Uses estimator S (sample standard deviation) in place of σ

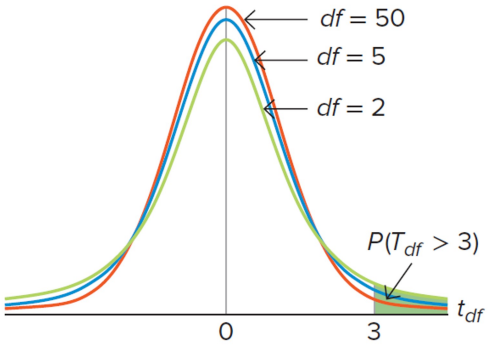
- Random variable T follows Student's t distribution / t distribution / t_{df} distribution

  $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$

  with $(n-1)$ degrees of freedom, df

- Summary of t_{df} distribution
  - Bell shaped and symmetric around 0 with asymptotic tails
  - Slighter broader tails than z distribution
  - Shape of each distribution depends on degrees of freedom df
  - As df increases, becomes similar to z distribution

FIGURE 8.3
The t_{df} distribution with various degrees of freedom



- Probabilities in table correlate to probabilities in upper tail of distribution

- Confidence interval

  $\overline{X} \pm t_{\alpha/2,df} \frac{s}{\sqrt{n}}$

  - Uncertainty increased when estimate population standard deviation with sample standard deviation
  - Confidence interval wider, especially for smaller samples (captured by wider tail of t_{df} distribution

HYPOTHESIS TESTING

- Determining validity of belief (unproven proposition or supposition that tentatively explains certain phenomena)
- Hypothesis is claim (assumption) about population parameter (inferential statistics)
  - Never about sample statistic
- Based on sampling distribution of normally distributed sample mean $\overline{X}$, i.e. $E(\overline{X}) = \mu$ and $se(\overline{X}) = \sigma / \sqrt{n}$

- Hypothesis testing used to resolve conflicts between two competing (mutually exclusive and exhaustive) hypotheses on population parameter of interest:
  - Null hypothesis H_O
    - Presumed default state of nature or status quo
    - Analogous to "innocent until proven guilty" or "individual is free of particular disease"
    - Always contains equality sign =, ≤, or ≥
    - May or may not be rejected
  - Alternative hypothesis H_A
    - Opposite of H_O and challenges status quo
    - Generally what we want to establish
    - Never contains =, ≤ or ≥; can only be ≠, < or >
    - May or may not be demonstrated/proven
- On basis of sample evidence we either
  - "Reject the null hypothesis"
    - Sample evidence inconsistent with null hypothesis
    - Conclude with alternative hypothesis
  - Or "do not reject the null hypothesis"
    - Sample evidence not inconsistent with null hypothesis
    - Not correct to "accept null hypothesis" because sample evidence does not necessarily prove null hypothesis is true (cannot be certain based on sample)

- One sample or two sample tests
  - One sample hypothesis tests - sample mean compared with hypothesised population mean
  - Two sample hypothesis tests - comparison of two means or matched pairs

- One-tailed or two-tailed tests
  - One-tailed
    - Has direction
    - Null hypothesis can only be rejected on one side of hypothesised value
    - H_O: $\mu \leq \mu_O, H_A$: $\mu > \mu_O$ (right-tailed test)
    - H_O: $\mu \geq \mu_O, H_A$: $\mu < \mu_O$ (left-tailed test)
  - Two-tailed
    - Either direction
    - Null hypothesis can be rejected on either side of hypothesised value
    - H_O: $\mu = \mu_O, H_A$: $\mu \neq \mu_O$

- When formulating hypotheses
  1. Identify relevant population parameter of interest (e.g. μ or p)
  2. Determine whether one- or two-tailed test
  3. Include some form of equality sign in null hypothesis and use alternative hypothesis to establish claim

Type I and Type II errors

- As decision of hypothesis test based on limited sample information, bound to make errors
  - Type I error: reject null hypothesis when it is actually true
    - Probability denoted as α
    - For medical test, false positive
    - Verdict of guilty when innocent
  - Type II error: do not reject null hypothesis when it is actually false
    - Probability denoted as β
    - For medical test, false negative
    - Verdict of innocent when guilty

Decision	Null hypothesis is true	Null hypothesis is false
Reject null hypothesis	Type I error (α)	Correct decision (1-β) Statistical power
Do not reject null hypothesis	Correct decision (1-α) Level of confidence	Type II error (β)

- Not always easy to determine which of two errors has more serious consequences
  - Optimal choice of α and β depends on context and relative cost of errors
- Reducing likelihood of Type I error increases likelihood of Type II error (↓α = ↑β), and vice versa (↑α = ↓β)
- Only way to lower both Type I and Type II errors is by increasing sample size n (providing more evidence)

One sample hypothesis tests

- Hypothesis test for population mean when σ known
  - σ is rarely known, but there are instances when it is considered fairly stable and determined from prior experience
  - First assume null hypothesis is true, then determine if sample evidence contradicts this assumption

  - Two approaches with the same outcome
    1. Critical value approach
       - Preferred when calculations done by hand
    2. P-value approach
       - Statistical software packages (e.g. Excel)

Steps for conducting hypothesis test with p-value approach:

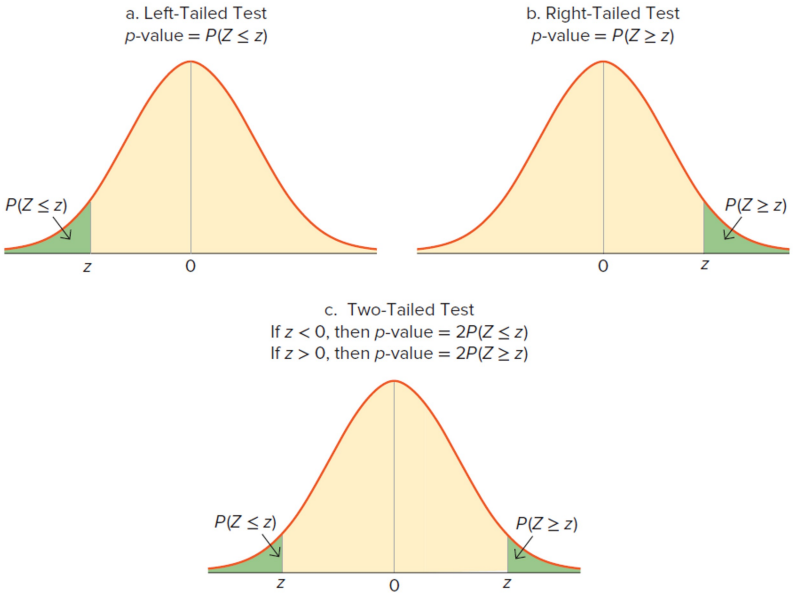
1. Specify null and alternative hypotheses
2. Specify significance level
   - Allowed probability of Type I error (α)

3. Calculate test statistic and p-value
   - Test statistic calculated by z distribution

$$z = \frac{\overline{x} - \mu_O}{\sigma / \sqrt{n}}$$

   - Find p-value using z value from test statistic (done by Excel or can look up in z table)
     - Indicates likelihood of obtaining sample mean at least as extreme as the one derived from given sample
     - E.g. If p-value = P(Z > 2.22) = 0.0132, then there is 1.32% chance that the sample mean will be greater than $\overline{x}$
     - Calculation depends on specification of alternative hypothesis:

FIGURE 9.2 The *p*-values for one- and two-tailed tests



Alternative Hypothesis	*p*-value
$H_A: \mu > \mu_0$	Right-tail probability: $P(Z \geq z)$
$H_A: \mu < \mu_0$	Left-tail probability: $P(Z \leq z)$
$H_A: \mu \neq \mu_0$	Two-tail probability: $2P(Z \geq z)$ if $z > 0$ or $2P(Z \leq z)$ if $z < 0$

     - If two-tail test, multiply p-value for one-tail test by 2 because P(Z > z) + P(Z < z) = 2P(Z > z) due to symmetry
     - If using Excel, need to be careful interpreting values, i.e. Correct signs and p-values
   - Reject null hypothesis if p-value < α,
   - Do not reject null hypothesis if p-value ≥ α. Rejecting at:
     - 10% significance level (α = 0.10) indicates some evidence null hypothesis false
     - 5% significance level (α = 0.05) indicates strong evidence null hypothesis false
     - 1% significance level (α = 0.01) indicates very strong evidence null hypothesis false
   - Significance of p-value

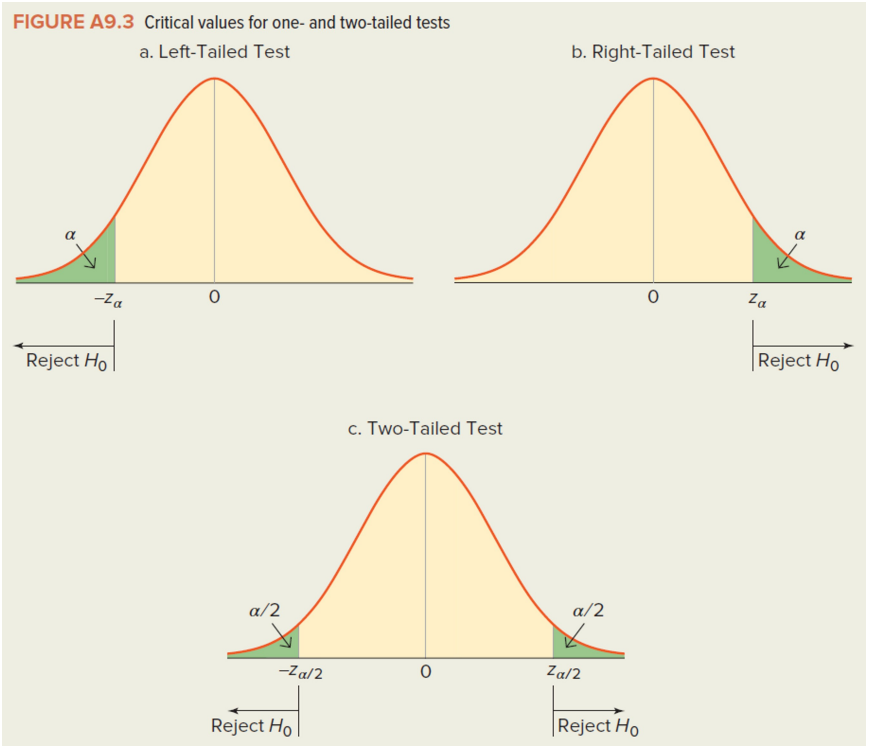
P-value	Evidence
< 0.01	Very strong (convincing) evidence alternate hypothesis is true
0.01 - 0.05	Strong evidence alternate hypothesis is true
0.05 - 0.10	Some evidence alternate hypothesis is true (grey area)
> 0.10	Weak or no evidence in support of alternate hypothesis

4. State conclusion and interpret results
   - Clearly communicate and interpret results meaningfully

Steps for critical value approach:

- Specifies rejection region such that if value of test statistic falls into rejection region then

we reject null hypothesis

- Critical value is point that separates rejection region from non-rejection region
- 

- For two-tail tests, significance level split in half to determine two critical values

1. Specify null and alternative hypotheses
2. Specify significance level and find critical value(s)
   - If test statistic follows z distribution, critical value is:
     - z_α where $P(Z \geq z_\alpha) = \alpha$ for right-tailed test
     - $-z_\alpha$ where $P(Z \geq z_\alpha) = \alpha$ for left-tailed test
     - $-z_{\alpha/2}$ and $z_{\alpha/2}$ where $P(Z \geq z_{\alpha/2}) = \alpha/2$ for two-tailed test
3. Calculate value of test statistic
   - Standardised value as in p-value approach
   - Reject if test statistic falls in rejection region
4. State conclusion and interpret results
   - Identical to p-value approach

- Commonly used z-values for specified p-values

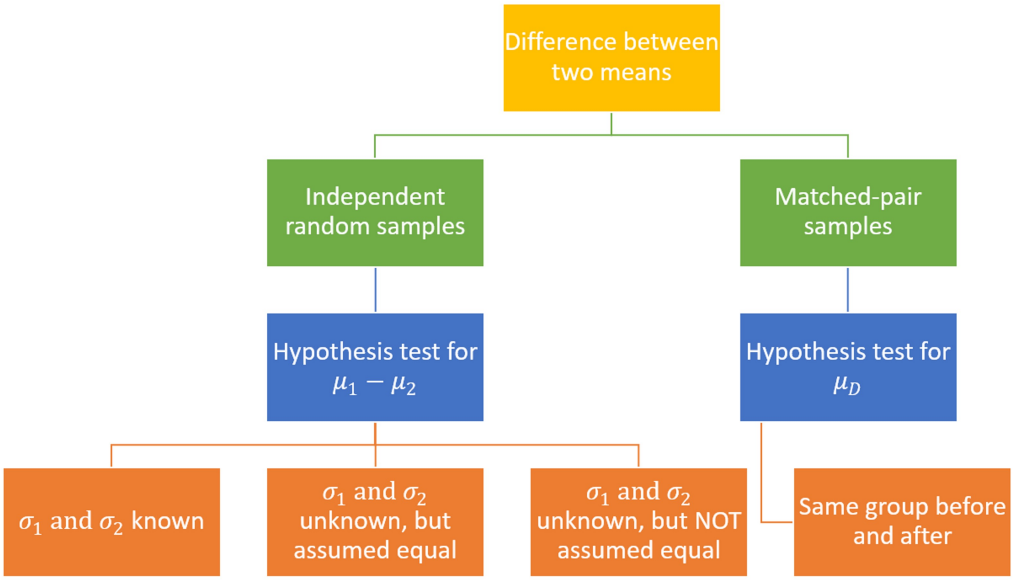
Alpha levels	0.10	0.05	0.01
Upper one-tail test	1.28	1.645	2.33
Lower one-tail test	-1.28	-1.645	-2.33
Two tail test	±1.645	±1.960	±2.575

- Hypothesis test for population mean when σ unknown
  - Occurs in most business applications, so replace σ with sample standard deviation s
  - Steps are the same as above, but test statistic is calculated with t distribution

$$t_{df} = \frac{\overline{x} - \mu_o}{s/\sqrt{n}}$$

with $(n-1)$ degrees of freedom, df

Two sample hypothesis tests



- Independent random samples
  - Process that generates one sample completely separate from (has no effect on) process that generates other sample
  - No natural pairing, samples clearly delineated
  - Sample sizes frequently different
  - Population standard deviations may be equal or not equal
  - E.g. Difference in average wage between male and female employees

  - Hypotheses:

Lower-tail test	Upper-tail test	Two-tail test
H_O: $\mu_1 - \mu_2 \geq d_o$ H_A: $\mu_1 - \mu_2 < d_o$	H_O: $\mu_1 - \mu_2 \leq d_o$ H_A: $\mu_1 - \mu_2 > d_o$	H_O: $\mu_1 - \mu_2 = d_o$ H_A: $\mu_1 - \mu_2 \neq d_o$

    - Where d_o is hypothesized difference between two population means
    - d_o commonly 0, so hypotheses can be expressed as:

Lower-tail test	Upper-tail test	Two-tail test
H_O: $\mu_1 - \mu_2 \geq 0$ H_A: $\mu_1 - \mu_2 < 0$	H_O: $\mu_1 - \mu_2 \leq 0$ H_A: $\mu_1 - \mu_2 > 0$	H_O: $\mu_1 - \mu_2 = 0$ H_A: $\mu_1 - \mu_2 \neq 0$
H_O: $\mu_1 \geq \mu_2$ H_A: $\mu_1 < \mu_2$	H_O: $\mu_1 \leq \mu_2$ H_A: $\mu_1 > \mu_2$	H_O: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$

  - Test statistic
    - Calculated by dividing point estimate ($(\overline{x_1} - \overline{x_2}) - d_o$) by standard error of estimator ($se(\overline{X_1} - \overline{X_2})$)
    - When σ_1 and σ_2 known

      $$z = \frac{(\overline{x_1} - \overline{x_2}) - d_o}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

    - When σ_1 and σ_2 unknown, but assumed equal

      $$t_{df} = \frac{(\overline{x_1} - \overline{x_2}) - d_o}{\sqrt{{s_p}^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

      Where ${s_p}^2 = \frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}$
      And $df = n_1 + n_2 - 2$

      - Pooled estimate of common variance (${s_p}^2$): estimate of population variance by weighted average of sample variance (${s_1}^2$ and ${s_2}^2$)

    - When σ_1 and σ_2 unknown, but NOT assumed equal

$$t_{df} = \frac{(\overline{x_1} - \overline{x_2}) - d_o}{\sqrt{(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2})}}$$

Where $df = \frac{(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2})^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}}$

- □ Cannot calculate pooled estimate of population variance
- □ Generally round df down

- Excel calculates these values for us, we just need to know which test to pick

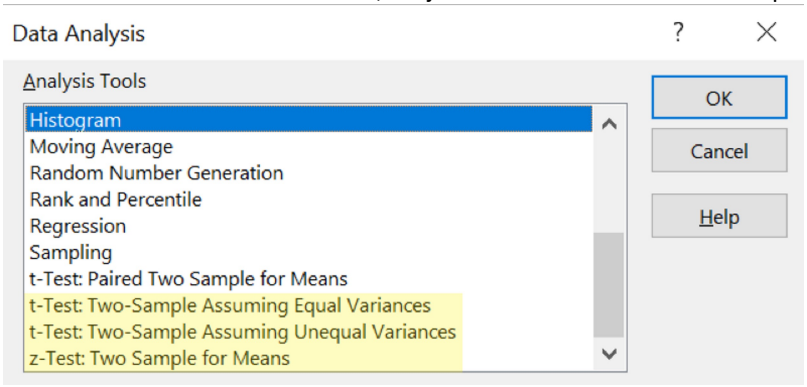


TABLE 10.3 Excel's Output for *t*-Test concerning $\mu_1 - \mu_2$

	Gold	Oil
Mean	16	17.3
Variance	70.6667	114.2333
Observations	10	10
Hypothesized Mean Difference	0	
Df	17	
t Stat	**−0.3023**	
P(T ≤ t) one-tail	0.3830	
t Critical one-tail	1.7396	
P(T ≤ t) two-tail	**0.7661**	
t Critical two-tail	**2.1098**	

- As with one sample tests, reject null hypothesis when:

Lower-tail test	Upper-tail test	Two-tail test
$t_{stat} < -t_\alpha$	$t_{stat} > t_\alpha$	$t_{stat} < -t_{\alpha/2}$ or $t_{stat} > t_{\alpha/2}$
	P(T>=t) < α	P(T>=t) < α

- Matched-pair samples
  - Samples naturally paired or matched
  - Sample sizes must be the same
  - Parameter of interest is mean difference μ_D, where $D = X_1 - X_2$
  - Much of noise is controlled for in these experiments

1. Observations on same subject - measurement, intervention, another measurement
   - Before and after studies
   - E.g. Measuring weight before and after diet plan
2. Observations on units that are similar or identical, where not on same subject that gets sampled twice
   - E.g. Matching 20 adjacent plots of land, then using organic fertiliser on one half and non-organic fertiliser on the other half

  - Hypotheses (hypothesised difference often 0):

Lower-tail test	Upper-tail test	Two-tail test
$H_O: \mu_D \geq d_o$ $H_A: \mu_D < d_o$	$H_O: \mu_D \leq d_o$ $H_A: \mu_D > d_o$	$H_O: \mu_D = d_o$ $H_A: \mu_D \neq d_o$
$H_O: \mu_D \geq 0$ $H_A: \mu_D < 0$	$H_O: \mu_D \leq 0$ $H_A: \mu_D > 0$	$H_O: \mu_D = 0$ $H_A: \mu_D \neq 0$
If before is less than after	If before is greater than after	If there is a difference

- Test statistic assumed to follow t_{df} distribution with df = n-1

$$t_{df} = \frac{\overline{d} - d_o}{s_D/\sqrt{n}}$$

  - $\overline{d}$ is mean of sample differences, s_D is standard deviation of sample differences

- Using Excel

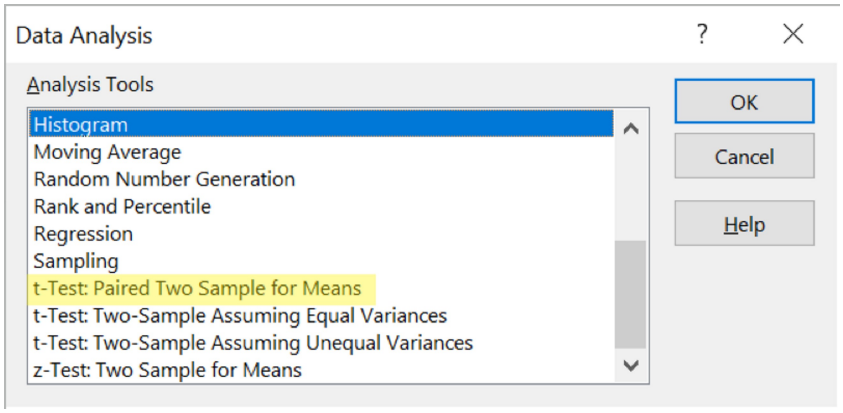


TABLE 10.5 Excel's Output for *t*-Test concerning μ_D

	Before	After
Mean	400.275	391.475
Variance	49.9481	42.3583
Observations	40	40
Pearson Correlation	0.27080	
Hypothesized Mean Difference	0	
Df	39	
t Stat	**6.7795**	
P(T ≤ t) one-tail	**2.15E-08**	
t Critical one-tail	1.6849	
P(T ≤ t) two-tail	4.31E-08	
t Critical two-tail	2.0227	

  - Reject null hypothesis if
    - p-value < α, or

Lower-tail test	Upper-tail test	Two-tail test
$t_{stat} < -t_\alpha$	$t_{stat} > t_\alpha$	$t_{stat} < -t_{\alpha/2}$ or $t_{stat} > t_{\alpha/2}$

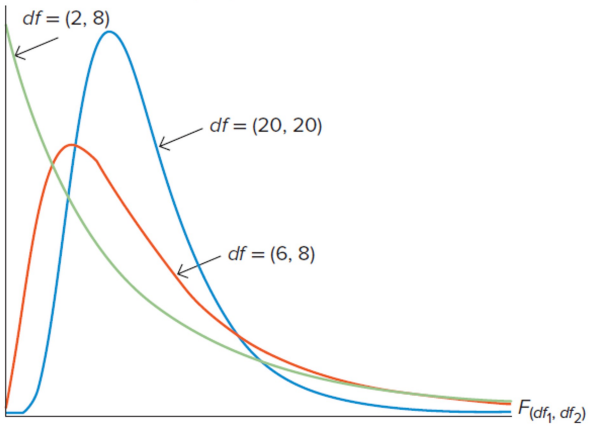
  - NB: Although p-value calculated correctly, expression Excel uses to denote p-value is not always correct

- Inference concerning difference among man means
  - One-way analysis of variance (ANOVA) test determines if differences exist between means of 3 or more population means
  - Assumptions
    - Samples selected independently
    - Populations normally distributed
    - Population variances unknown but assumed equal

  - Based on F distribution

- Depends on two degrees of freedom
- Positively skewed. Values range from 0 to infinity
- As df_1 and df_2 increase, becomes increasingly symmetrical

FIGURE 10.3 The $F_{(df_1, df_2)}$ distribution with various degrees of freedom



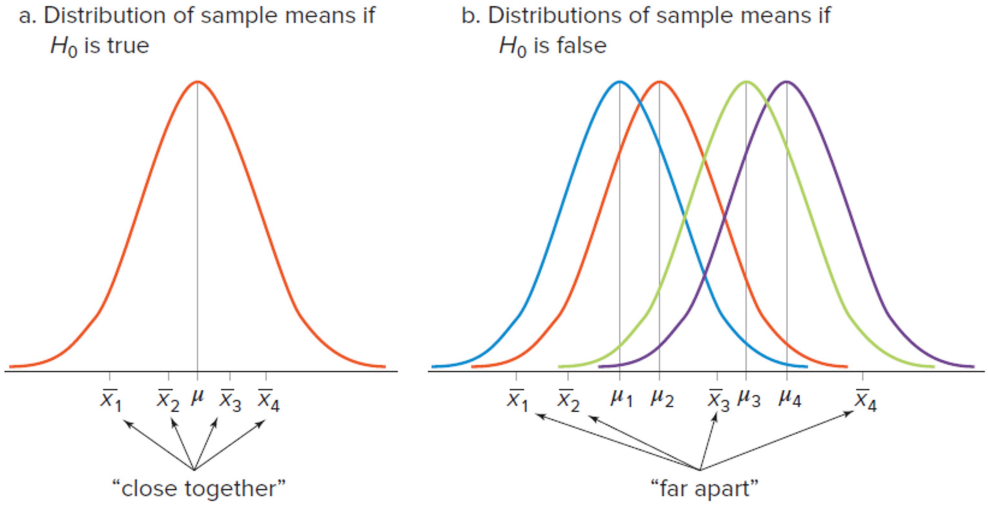
- Hypotheses

  H_O: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n$

  H_A: Not all population means are equal.

- Logic:
  - If population means are equal (H_O true), sample means should be relatively close to each other
  - If populations means differ (H_A true), sample means relatively far apart and variability cannot be explained by chance

FIGURE 10.5 The logic of ANOVA



- Why we don't set up a series of two-sample t tests with equal variances
  - Cumbersome and flawed - inflate risk of Type I error α, resulting significance level greater than 5%
  - ANOVA is one test that simultaneously evaluates equality of several means

- Steps to interpret
  1. Specify null and alternative hypothesis
  2. Specify significance level (α = 0.05)
  3. Generate output from Excel

TABLE 10.10 Excel-Produced ANOVA Table for Public Transportation Example

SUMMARY					
Groups	Count	Sum	Average	Variance	
Boston	5	63110	12622	7707.5	
New York	8	100680	12585	6464.3	
San Francisco	6	70320	11720	7050	
Chicago	5	53650	10730	8212.5	
ANOVA					
Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	13204720	3	4401573	**610.566**	**7.96E-20**
Within Groups	144180	20	7209		
Total	13348900	23			

- Reject null hypothesis if p-value < α
- Output shows variation between and within groups, test statistic F and p-value

4. State conclusion and interpret results
   - If we reject null hypothesis, conclude not all population means are equal
   - Does not allow us to infer which individual means differ, just at least one differs

PREDICTIVE MODELLING - REGRESSION ANALYSIS

- Regression analysis one of most important statistical methodologies
- Analyse linear relationship between dependent variable and independent variable
  - Predict and/or describe changes in dependent variable
  - Dependent variable (y) - explained variable, predicted variable, regressand
  - Independent variable (x_k) - explanatory variable, predictor variable, control variable, regressor
- Detects correlation, does not prove causality
- Develop mathematical model that captures relationship between dependent variable y and k independent variables $x_1, x_2, x_3, \cdots, x_k$

Simple linear regression

- Involves one independent variable

- Simple linear regression model

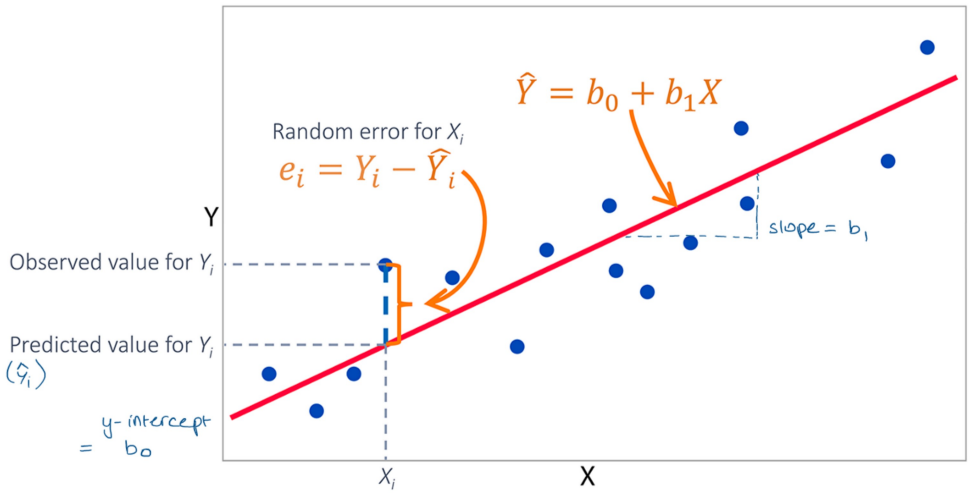
$$y = \beta_0 + \beta_1 x + \epsilon$$

  - y is dependent variable, x is independent variable
  - ϵ is random error term as observed value may differ from expected value
    - Captures omitted variables that may be difficult to measure
  - Coefficients β_0 and β_1 are unknown parameters to be estimated
    - β_0 y-intercept
    - β_1 slope parameter (positive or negative linear relationship)

- Simple linear regression equation for model

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

  - $\hat{y}$ (y-hat) predicted value of dependent variable
  - b_0 and b_1 estimates of β_0 and β_1 respectively
    - b_1 represents change in $\hat{y}$ when x increases by one unit
    - b_0 represents value of $\hat{y}$ when x is 0. Not always possible to provide economic interpretation of , but can be thought of as a starting point or y not explained by x
  - Residual $e = y - \hat{y}$ represents difference between observed and predicted values



- Method of least squares or ordinary least squares (OLS)
  - Used to find linear trend line and estimate parameters β_0 and β_1
  - Minimises error sum of squares (SSE) where $SSE = \sum(y_i - \hat{y_i})^2 = \sum {e_i}^2$
    - Sum of squared vertical distances from observation points to regression line

- Before estimating model, useful to visualise relationship between y and x with scatterplot
- Avoid making predictions based on values outside of sample range, unless we can assume the same linear relationship will be maintained

Multiple linear regression

- Involves more than one independent variable
- Choice of independent variables based on economic theory, intuition and/or prior research

- Multiple linear regression model

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

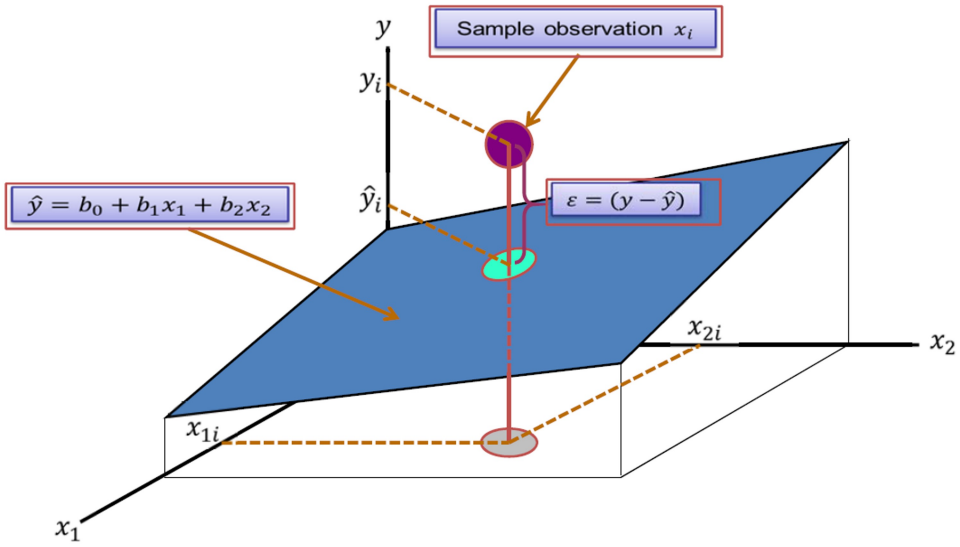
  - $x_1, x_2, \ldots, x_k$ are *k* independent variables

- Sample multiple regression equation

  $$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

  - Slight modification in interpretation of slopes b_1 to b_k as they show partial influences
  - E.g. If k = 3 independent variables, then b_1 estimates how change in x_1 will influence $\hat{y}$, holding x_2 and x_3 constant

  - If independent variable has same scale of measurement, we can compare their relative effects to each other
  - If different scale of measurement, cannot conclude which is the stronger predictor



- Need to view as 3D figure predicting plane of best fit

Goodness-of-fit measures

- Summarise how well sample regression equation fits data
  - If each predicted value $\hat{y}$ equals observed values y, then it's perfect fit

- Standard error of the estimate (s_e)
  - Sample standard deviation of residual - error sum of squares divided by degrees of freedom to get variance of residual, then square root the variance

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

  - s_e can have values from 0 to infinity
  - Smaller indicates less dispersion around regression line, implying model is good fit
  - Magnitude of s_e should be judged relative to size of y values in sample data
  - Excel refers to s_e as Standard Error

- Coefficient of determination (R^2)
  - Ratio of explained variation of dependent variable to total variation
    - E.g. If $R^2 = 0.72$, we say that 72% of sample variation in dependent variable explained by sample regression equation

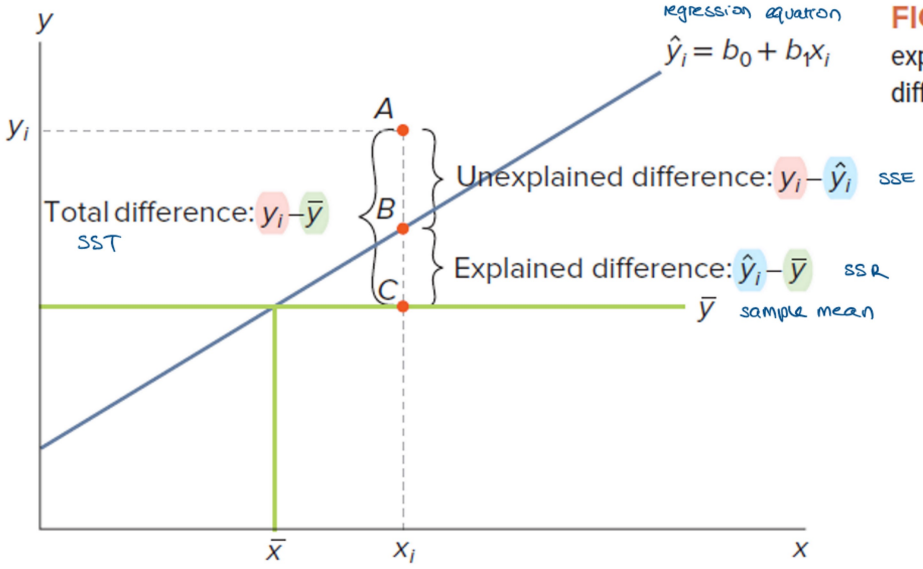


FIGURE 12.5 Total, explained, and unexplained differences

  - SST (total sum of squares) $\sum(y_i - \overline{y})^2$ (total) variation between observed value and sample mean
  - SSR (regression sum of squares) $\sum(\hat{y_i} - \overline{y})^2$ (explained difference)

variation between predicted value and sample mean
    - SSE (error sum of squares) $\sum(y_i - \hat{y}_i)^2$ (unexplained difference)
      variation between observed value and predicted value
        - Due to random error or chance

    - SST = SSR + SSE

  - Formula for coefficient of determination:

    $$R^2 = \frac{SSR}{SST} \text{ or equivalently } R^2 = 1 - \frac{SSE}{SST}$$

    - Value between 0 and 1. Closer to 1, better fit
    - 0.8 is a very good model, 0.6-0.7 is good model

  - Excel refers to it as R Square
    - Also reports Multiple R which is sample correlation coefficient between dependent variable y and its predicted value

- Adjusted coefficient of determination (Adjusted R^2)
  - Used when comparing models with the same *y* but different number of independent variables
    - R^2 never decreases as we add more independent variables, so possible to increase its value unintentionally by including insignificant independent variables

  - Formula for adjusted coefficient of determination

    $$Adjusted\ R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

    - Accounts for number of independent variables *k*
    - May be negative if correlation between dependent variable and independent variables is low enough

Tests of significance

- Tests of individual significance
  - Whether slope coefficient (β_j) is different from 0
    - If $\beta_j = 0$ corresponding x_j drops out of equation, therefore no linear relationship between x_j and *y*
    - Conversely if $\beta_j \neq 0$, then x_j influences *y*
  - Competing hypotheses for two-tailed test:
    - H_O: $\beta_j = 0$
    - H_A: $\beta_j \neq 0$

  - Test statistic is calculated as

    $$t_{df} = \frac{b_j - \beta_{jO}}{se(b_j)}$$

    Where $df = n - k - 1$

  - If p-value < α, then we reject null hypothesis and conclude with alternative hypothesis that slope coefficient is not 0 and is significant

- Test of joint significance
  - Overall usefulness of regression
  - Whether independent variables have joint statistical influence on dependent variable

  - Hypotheses:
    - H_O: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
    - H_A: At least one $\beta_j \neq 0$

  - Right-tailed F test

TABLE 12.9 General Format of an ANOVA Table for Regression

ANOVA	df	SS	MS	F	Significance F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{(df_1, df_2)} = \frac{MSR}{MSE}$	$P\left(F_{(df_1, df_2)} \geq \frac{MSR}{MSE}\right)$
Residual	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$		
Total	$n - 1$	SST			

- Excel reports F test statistic as F, and p-value as Significance F
- Generally if F value is large, indicates large portion of sample variation explained by regression model, and model is useful
- If p-value < α, reject null hypothesis and conclude at least one of slope coefficients does not equal 0 and model is significant

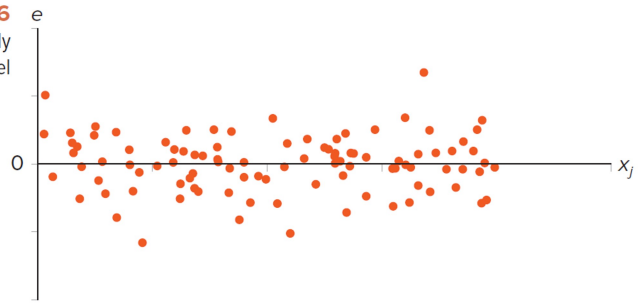
Required assumptions for regression analysis model (L.I.N.E.)

1. Linearity - slope parameters $\beta_0,\ \beta_1,\ \beta_2, \cdots,\ \beta_k$ are linear (i.e. Relationship between *x* and *y* is linear)
2. Independence of errors
   - Error values are independent, no correlations across observations
   - No serial correlation - error term ϵ uncorrelated across observations
   - No endogeneity - error term ϵ not correlated with any independent variables
3. Normality of error
   - Error term ϵ normally distributed - allows us to construct interval estimates and tests of significance. If ϵ not normally distributed, tests valid only for large sample sizes
4. Equal variance (homoscedasticity)
   - No heteroskedasticity - variability of error term is same for all observations. Violated if observations have changing variability
5. No perfect multicollinearity - no exact linear relationship among independent variables (independent from each other)
6. Error term ϵ has expected value of zero, i.e. $E(\epsilon) = 0$, this implies $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

Residual plots analysis

- Informal analysis of regression model - detects some common violations to model assumptions
- Can also detect outliers, residual will appear distinct in plot
- Residual plot with no violations

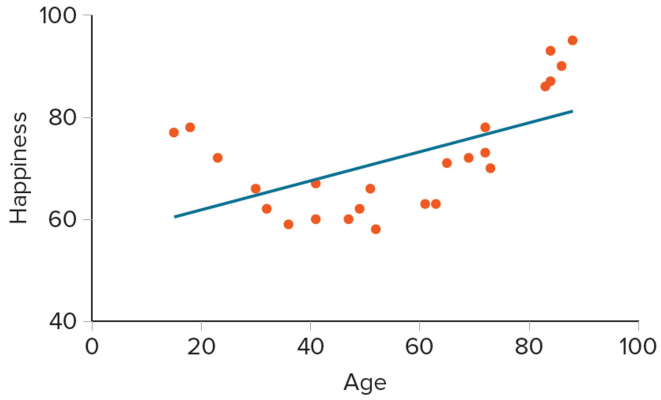
FIGURE 12.6 Residual plot of a correctly specified model



   - All points randomly dispersed around 0, no discernible patter
   - No evidence of outliers

1. Nonlinear patterns
   - When relationship cannot be represented by straight line
   - Discernible trend indicates nonlinear patterns

FIGURE 12.7 Scatterplot and the superimposed trendline (Example 12.8)

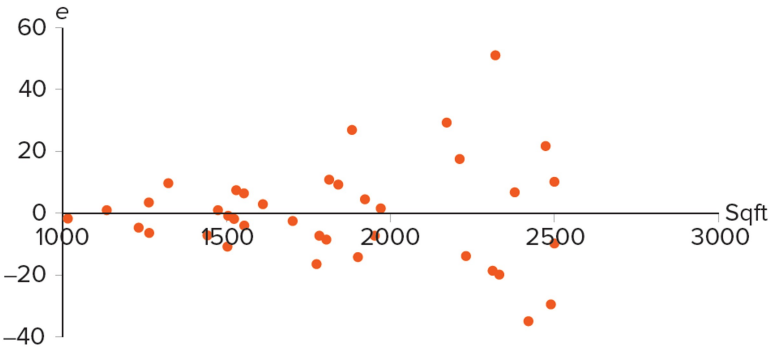


- Remedy by applying nonlinear regression methods

2. Multicollinearity
   - Two or more independent variables have exact linear relationship
     - Imprecise estimates of slope coefficients
   - Detection
     - Seemingly wrong signs on coefficients
     - Variable previously significant in model (very strong significance, not on cusp of 0.05) becomes non-significant when new independent variable added
     - High R^2 with individually insignificant independent variables
     - Examine correlations between independent variables, if correlation coefficient >0.80 or <-0.80
     - Sizeable change in values of previously estimated coefficients when new variable added to model
     - Estimate of standard deviation of model increases when variable added to model
   - Remedy
     - Drop one of collinear variables
     - Obtain more data and increase sample size. Sample correlation may get weaker if more observations included
     - Apply Bayesian or ridge regression, transform variables or combine two variables
     - Do nothing, especially when model yields high R^2

3. Changing variability (heteroscedastic)
   - Results in inefficient estimators, hypothesis tests for significance no longer valid
   - Residuals plotted against each independent variable and predicted value
     - If violation, variability increases or decreases over values of *x*. Residuals fan out across horizontal axis

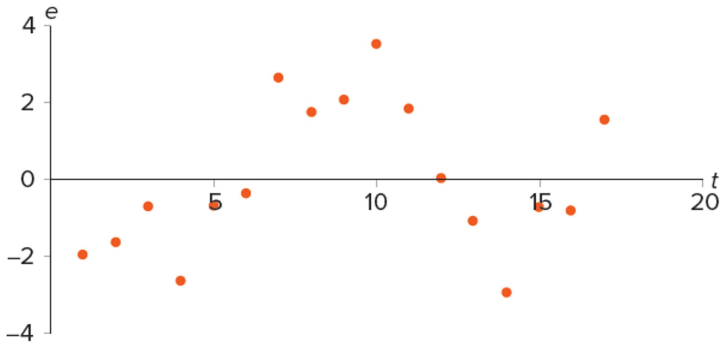
FIGURE 12.9 Residual plot against square footage (Example 12.10)



   - Slope coefficients still useful, but t and F tests no longer valid
     - May use slope coefficients along with correction for standard errors

4. Serially correlated observations
   - Often occurs with time series data where variables exhibit cycles
   - Residuals plotted sequentially over time
     - If violation, shows wavelike movement

FIGURE 12.10 Scatterplot of residuals against time *t*



5. Excluded variables (endogeneity)
    - Commonly occurs when important independent variables omitted
    - Variables incorporated in error term
    - Results in unreliable coefficient estimates
    - May be difficult to remedy
        - Before running regression model, compile comprehensive list of potential independent variables
        - Due to data limitations, may be difficult to include all variables, so can use instrumental variable technique

Dummy variables

- For independent variables that are qualitative/categorical
    - E.g. Gender (male/female), nationality (Australian/non-Australian), education (secondary, undergraduate, postgraduate)
    - Also known as indicator variable

- Two categories
    - Dummy variable *d* takes on values of 0 or 1
    - Regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 d$$

        - If d = 1

$$\hat{y} = b_0 + b_1 x_1 + b_2$$

        - If d = 0

$$\hat{y} = b_0 + b_1 x_1$$

        - d affects intercept but not slope, resulting in two parallel lines

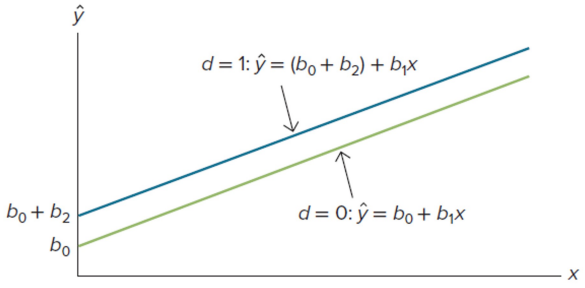


FIGURE 13.1 Using *d* for an intercept shift

    - Interpretation: keeping all other variables constant, the difference between d=0 and d=1 is b_2

- Multiple categories
    - If qualitative variable has more than 2 categories, we use multiple dummy variables
    - E.g. Mode of transport - public transport, car driving, park-and-ride
        - Define two dummy variables d_1 and d_2 that collectively represent one categorical variable

$$\hat{y} = b_0 + b_1 x_1 + b_2 d_1 + b_3 d_2$$

        - $d_1 = 1$ and $d_2 = 0$ for public transport
        - $d_1 = 0$ and $d_2 = 1$ for car driving
        - $d_1 = 0$ and $d_2 = 0$ for park-and-ride
            - Reference category
            - Choice of reference does not matter for making predictions

- Dummy variable trap
  - Number of dummy variables should be one less than number of categories of qualitative variable
  - Otherwise creates perfect multicollinearity

Chi-square χ^2 test

- Used to predict qualitative dependent variables with logistic regression
- Frequency differences across groups

Model building

- Goal is to develop model with best set of independent variables
  - Easier to interpret if unimportant variables removed
  - Lower probability collinearity
- Stepwise regression procedure
  - Final model only includes significant variables (principle of parsimony - simpler model is better, balancing with maintaining high R^2)
  - Overall goodness-of-fit
  - Predictors - what is significant in model
  - Diagnostics (residual analysis) - are modelling assumptions met

- Always think about expected relationship between independent and dependent variables

- Common problems
  - Conclusions and inferences made from regression line statistically valid only over range of data contained in sample
  - Do not assume correlation implies causation
  - High coefficient of determination does not guarantee model is good predictor
  - Confidence and prediction errors may be too wide for model to be used in many situations

Complex modelling techniques

- Hierarchical modelling
  - Used to compare models (R^2, adjusted R^2, standard error, significance of independent variables)
  - Logic on how to step down one variable at a time from full model
  - May also compare two models of interest instead of starting with full model
  - Not good to remove variables one at a time randomly as in stepwise regression
  - Other purposes for models instead of prediction
    - Determining if independent variable is important predictor (may not reduce to smallest number of predictors)
    - Compare models

- Control variables
  - Similar to hierarchical models
  - Model control variable along with independent and dependent variables to isolate control variable's effects from relationship between variables of interest
  - Partials out noise from variables that we can't necessarily change

- Dealing with non-linear relationships
  - Relationship forms curve instead of straight line on scatterplot
  - Quadratic regression model
    - Include x2 in regression equation
    - Allows one sign change of slope
  - Polynomial regression models
    - Various numbers of sign changes
    - E.g. Cubic regression model includes x, x2 and x3 and allows for two sign changes to slope

- Mediation
  - Mediator variables are hidden/lurking/third variable that explain how and to some extent why two variables are related
  - Independent variable has indirect relationship with dependent variable rather than a

direct one
    - To establish mediation
        1. Show X predicts mediator M
        2. Show X predicts outcome Y
        3. Show M affects outcome Y
        4. Establish M completely mediates X-Y relationship. Effect of X on Y controlling for M becomes insignificant

- Moderation
    - When two independent variables combine or interact to account for variation in dependent variable
    - Moderator variable M can enhance or reduce relationship between independent variable and dependent variable
    - Relationship between X and Y changes depending on level of M
    - Interaction term (product of X and M, XM) is significant and included in model