BUSS1020 NOTES

BUSS1020 Week 1 Lecture – Introduction and Discipline of Business Analytics

Introduction

- Business analytics is the discovery and communication of meaningful patterns in data.
- It is especially valuable in areas rich with recorded information, relying on the simultaneous application of statistics, computer programming and operations research to quantify performance.
- Often favours data visualisation to communicate insights, with firms applying analytics to describe, predict and improve business performance.

Why Study Business Analytics?

- Statistics is the study of the collection, analysis, interpretation, presentation and organisation of data.

The Data Deluge

- More and more processes generate data and soon all problems will as well.
- More and more companies use and need analytics.
- Everyone will need analytics eventually.

Applicable Areas

- New customer acquisition
- Measuring and boosting customer loyalty
- Pricing tolerance
- Cross-sell/up-sell/targeted advertisements
- Supply, staffing optimisation
- Financial forecasting
- Fraud detection
- Product placement
- Churn measurement and reduction
- Website design and management

Business Analytics and Statistics

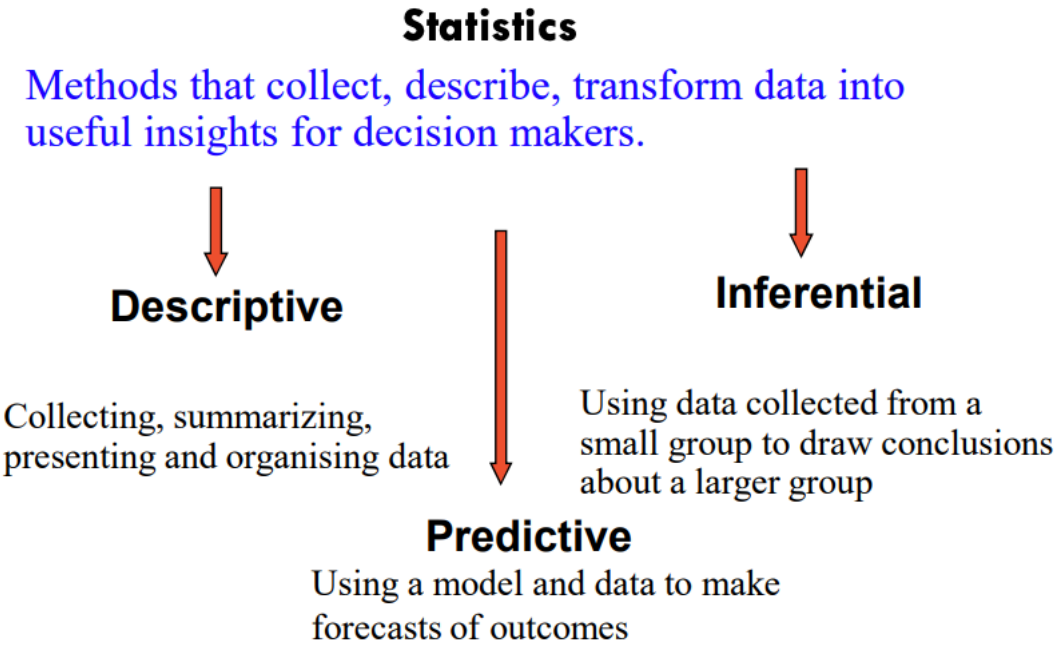
Framework for Conducting Statistical Analyses: DCOVA

- 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 analysis, tabulate and summarise it
- Visualise: the data
- Analyse: the data

Three Different Branches of Statistics Used in Business

- These branches are used to improve business processes through better understanding of the data they generate, and by allowing better decisions to be made.

- Descriptive – Collecting, summarising and organising data.
- Inferential – Using data collected from a small group to draw conclusions about a larger group.
- Predictive – Using a model and data to make forecasts of outcomes.



Descriptive Statistics

- Collect data e.g. survey.
- Present data e.g. tables and surveys.
- Summarise and visualise data e.g. sample average.

Inferential Statistics

- Drawing conclusions about a large group of individuals based on a smaller group.

  → Estimation e.g. estimate the population average amount spent using the sample average spent.

  → Hypothesis testing e.g. test the claim that the population average amount spent in one group is larger than that in another group.

Predictive Statistics

- Making reliable predictions based on a sample of data and a model.

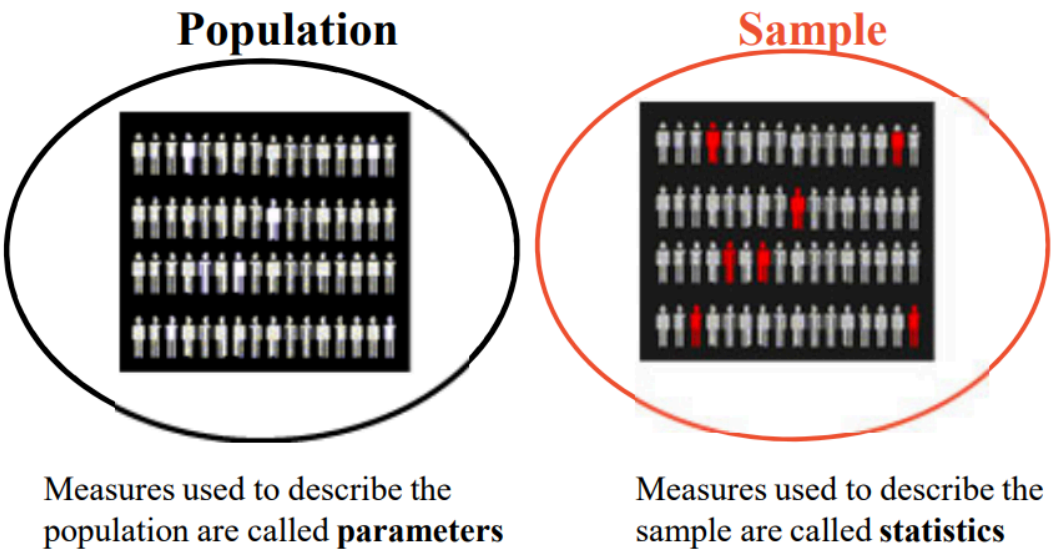
  → Predict the amount a new customer will spend based on their attributes.

  → Use a statistical model.

Basic Vocabulary of Statistics

- Variables - characteristics of an item or individual. Data on a variable(s) is what you analyse when you use a statistical method (variables are often called attributes) e.g. age.
- Data - the observed values or outcomes of one or more variables e.g. 32, 12, 40.

- Operational Definition - Variables should have universally accepted meanings that are clear to all associated with an analysis; that clearly defined meaning is the operational definition e.g. age in whole years on Jan 1st, 2019.
- Population - A population consists of all the items or individuals which you want to draw a conclusion. The population is the 'large group'.
- Sample - A portion of a population selected for analysis (small group).
- Parameter - A parameter is a numerical measure that describes a relevant characteristic of a population.
- Statistic - A statistic is a numerical measure that describes a characteristic of a sample. Often a statistic estimates a parameter.



Types of Variables

- At each step of a statistical analysis, the type of data must be known.
- The data type strongly influences how the analysis proceeds, the choice of methods etc → affects all parts of DCOVA.

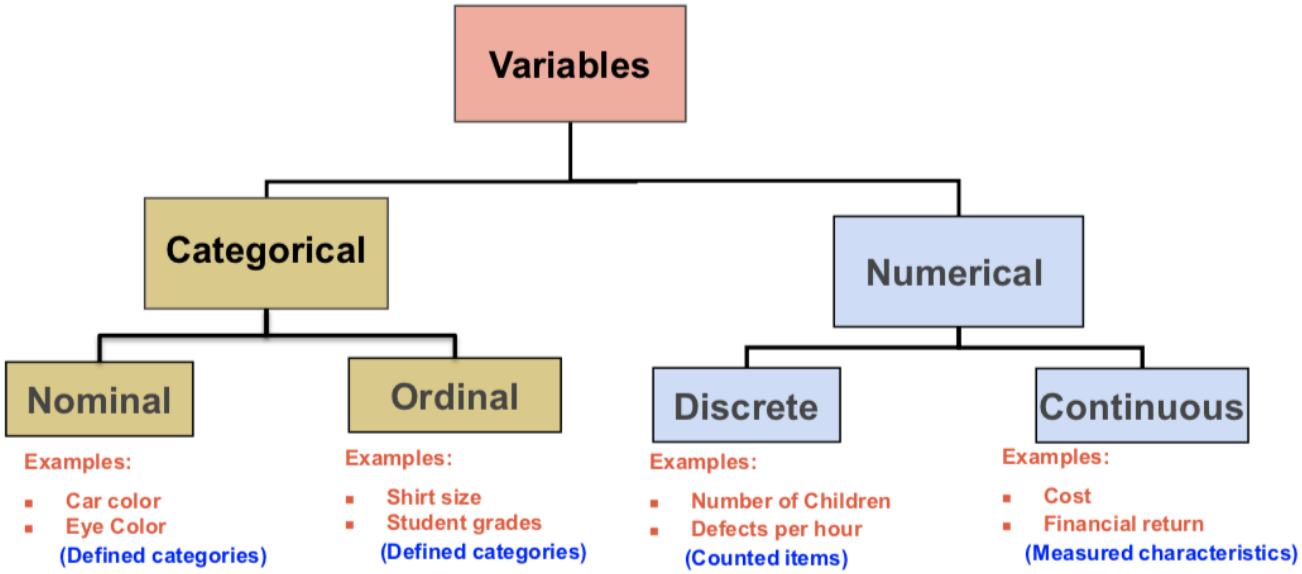
Categorical Variables

- Categorical (qualitative) variables have values that can only be placed into categories e.g. yes or no.
  - → Nominal – car colours (red, blue, black etc).
  - → Ordinal – shirt sizes (small, medium, large etc).

Numerical Variables

- Numerical (quantitative) variables have values that represent actual number quantities.
  - → Discrete variables arise from a counting process
  - →Continuous variables arise from a measuring process: can be assigned a value within a given interval.
  - → Sometimes discrete variables with many outcomes are treated like continuous variables e.g. prices.

Types of Variables



Levels of Data Measurement

Nominal (Lowest Level)

- Classify or categorize e.g. employment classification.
- Labels are used to distinguish different categories that have no order.

Ordinal

- Labels are used to classify and to indicate rank or order.
  - → Often represent an underlying scale e.g. quality.
  - → Differences between levels are not comparable.
- E.g. feedback surveys with options (e.g. not helpful, somewhat helpful, moderately helpful etc).

Interval

- 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.
- E.g. temperature, calendar time, monetary utility, scaled marks.

Ratio (Highest Level)

- Same properties as interval data plus zero has a true meaning (represents absence of the thing being measured).
  - → Measurements: height, weight, volume.
  - → Price, profit, loss, revenue etc.
  - → Financial ratios, returns, inventory turnover etc.

Data Collection

Sources of Data

Primary Sources

- The analyst collects the data.

Secondary sources

- The analyst is not the data collector.

Data Distributed by Organisations

- 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 etc.
- Data sets cover large spatial areas and/or long time periods.

Data from a Designed Experiment

- Consumer testing of different versions of a product.
- Quality testing.
- Market testing etc.
- The researcher subjects different groups of people to different conditions and observes the result → tests a theory or hypothesised outcome.

Survey Data

- Political polls.
- Determine customer satisfaction with a recent product or service experience.
- Internet polls.
- USE.
- People are asked questions about their beliefs, attitudes, behaviours, and other characteristics.

Data from Observational Studies

- Market researchers utilizing focus groups to elicit unstructured responses to open ended questions.
- Measuring the time it takes for customers to be served in a fast food establishment.
- Measuring the volume of traffic through an intersection to determine if some form of advertising at the intersection is justified.
- The researcher collects data by directly observing a behaviour, usually in a natural or neutral setting.

Automated and Streaming Data

- Mobile phone data usage etc.
- Financial markets.
- GPS data.
- 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 Format

- Traditionally, data is stored in easy to use Excel-type tables, one column for each clearly defined variable, one row for each item etc → structured (can be tabulated).
- Lots of modern data is unstructured → cannot be tabulated, messy, not stored easily or in an Excel file, stored in several different locations etc.

    → Such data needs lots of: data linking, preparation, data cleaning etc. before any organisation or analysis.

    → Video streams, audio, voice, email, tweets, texts, posts, interviews, blogs, pictures etc.

Cleaning Data

- Modern data often contains errors, missing values and outliers.
- Data often goes through a 'cleaning' process:

    → Remove errors (e.g. negative trade volumes, 0 heart rates).

    → Fill in or delete missing data (raises questions of which one?)

    → Flag strange but possible data points – outliers.

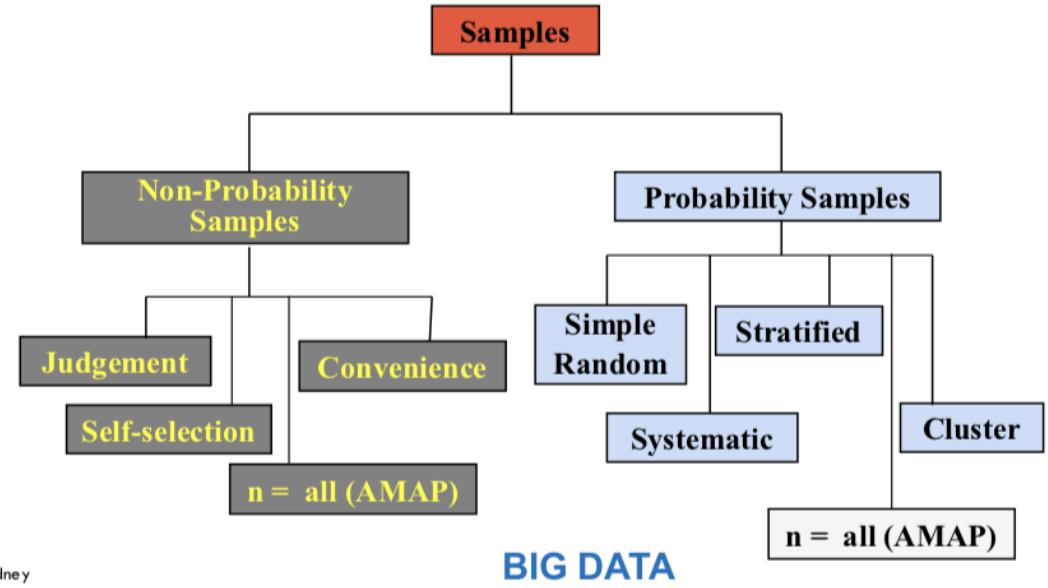
Sampling

- Often, we can't get the whole population → collecting information from a sample is less time consuming and less costly than selecting every item from the population (census).
- An analysis of a sample is often less cumbersome and more practical than an analysis of the entire population.

Sampling: Begin with a Sampling Frame

- List of items that are in the population and can be sampled.
- Includes population lists, directories, customer databases etc.
- Inaccurate or biased results can result if parts of the population are excluded.

Types of Samples



Non-Probability Sampling

- Items are chosen without regard to their probability of occurrence.
  - → Convenience sample – easy selection, quick, inexpensive.
  - → Judgement sample – 'experts' select most appropriate items/people, by convenience.
  - → Self-selected – individuals choose to participate.
  - → Quota sample – pre-set quotas of groups chosen, by convenience.

Probability Sampling

- Items are chosen randomly, sometimes using known probabilities that (closely) match those in the population.

Simple Random Sampling

- 'drawing names out of a hat'.
- Every individual or item (in the frame) has equal chance of being selected.
- May be with replacement or without replacement.
- Often obtained via a random number generator or via software.

SRS: Using A Random Number Generator

Sampling frame with 850 Items

Item Name	Item #
Bev R.	001
Ulan X.	002
.	.
.	.
.	.
.	.
Joann P.	849
Paul F.	850

E.g. in Excel:

Randomly select a number between 1 and 850. Match the number to the item and include it in your sample.

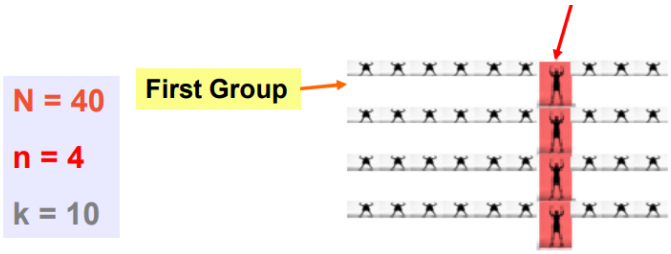
Excel in-cell function

=RANDBETWEEN(1,850)

493

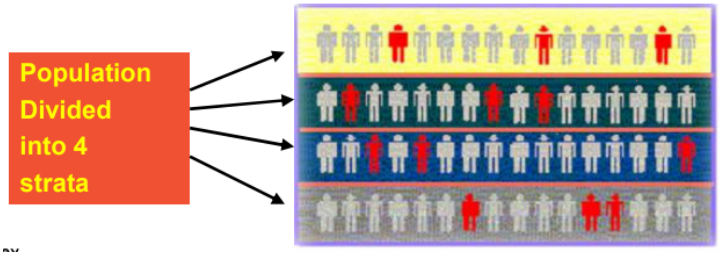
Systematic Sample

- Decide on sample size: n
- Divide frame of N individuals into groups of K individuals whereby k=N/n
- Randomly select one individual from the 1st group
- Select every kth individual after



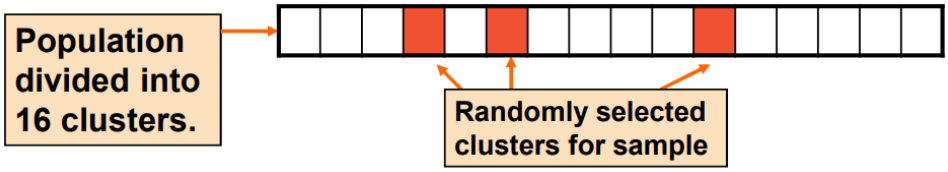
Stratified Sample

- Divide frame into strata according to a characteristic e.g. gender.
- An SRS is selected from each strata, with sample size proportional to each strata's size.
- Samples from each strata are combined into one sample.
- Common technique when sampling voters e.g. stratify across socio economic variables.
- Ensures proportionate representation by ensuring that minority groups are included.



Cluster Sample

- Population is divided into several clusters, each representative of the population.
- An SRS of clusters is selected.
- 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 (fully) sampled.



Comparing Probability Sampling Methods

SRS and Systematic Sampling

- Simple, cheap to use, effective against many types of bias.
- May not give the best representation of the population's underlying characteristics.

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

Coverage Error or Selection Bias

- Exists if some groups are excluded from the frame and have little or no chance of being selected.

Non-Response Error or Selection Bias

- People who choose not to respond may be different from those who do respond.
- Follow-up or non-responses.

Sampling Error

- Variation from sample to sample → will always exist

Measurement Error

- Due to the weakness in question design e.g. ambiguous, unclear or leading question, respondent error, and interviewer's effect on the respondent ("Hawthorne Effect")

Evaluating Survey Worthiness

- What is the purpose of it?
- Is it based on a probability sample?
- Coverage error – appropriate frame?
- Non-response error – how are non-responders followed up?
- Measurement error – clear, unambiguous questions?
- Sampling error – always exists unless n = N.

BUSS1020 Week 2 Lecture – Organising and Visualising Data

- Data is organised and visualised so as to reveal and glean any insights hidden within it, then communicate that information, especially the main features and patterns.
- First step: explore, find the 'story' in the data.
- Last step: Communicate the 'story'.

Organising Categorical Data

Organising One Variable Data (Summary Table)

- A summary table indicates frequency, amount, percentage or proportion in each category.
- We can see:
  - → The relative frequency of each category
  - → Differences between categories

Frequency table results for Type: Count = 316

Type	Frequency	Percent of Total
Growth	227	71.8
Value	89	28.2

See textbook for Excel instructions for these sort of tables

Visualising Categorical Data

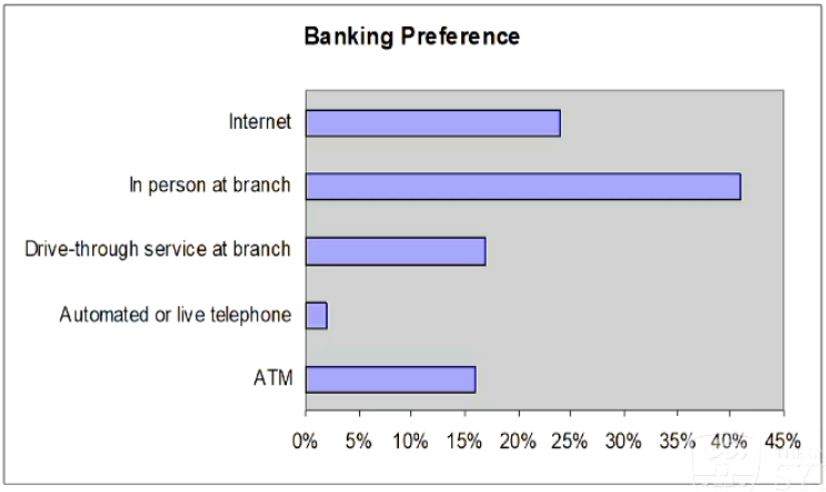
Visualising One Variable Data

Bar Chart

- In a bar chart, a bar shows each category, the length of which represents the frequency (amount, count), relative frequency or percentage of values falling into that category.

A Survey of 1000 Bank Customers:

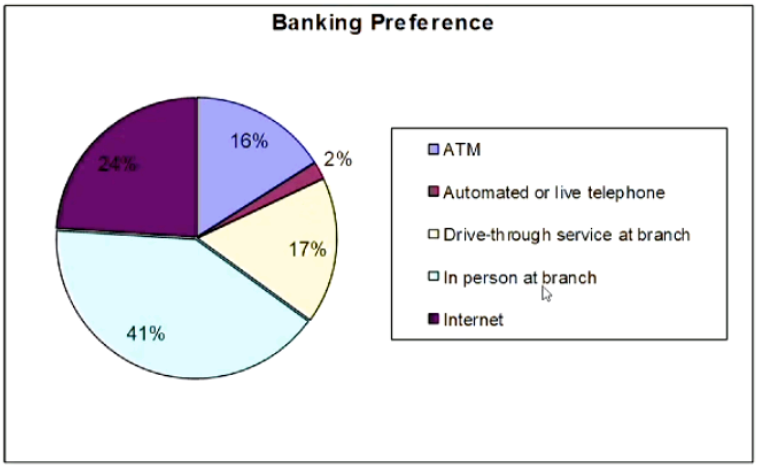
Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%



Pie Chart

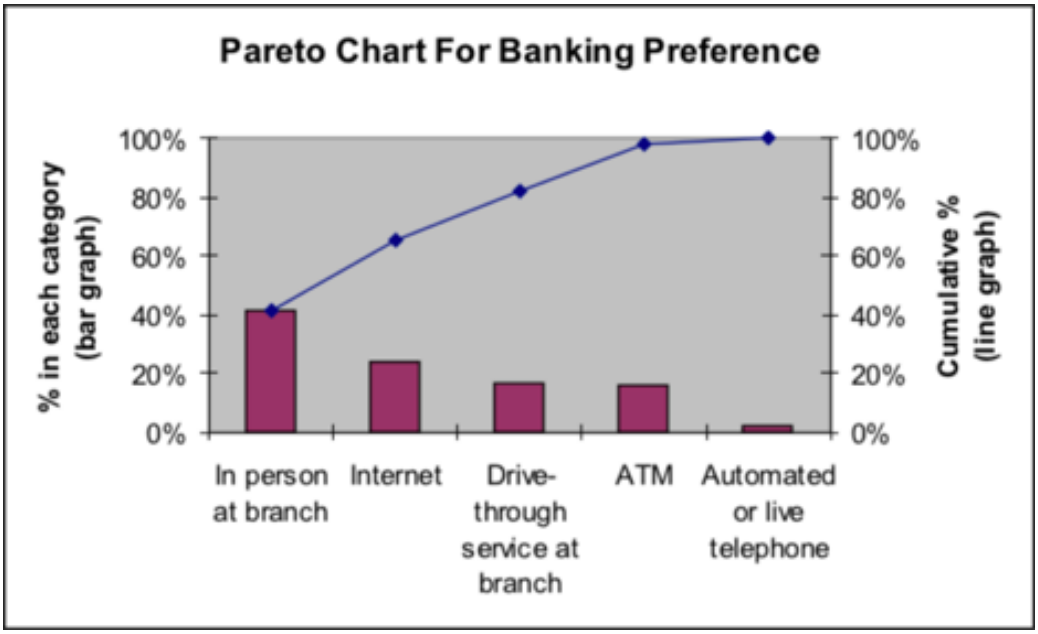
- The pie chart is a circle broken up into slices that represent categories, where the size of each slice of the pie varies according to the percentage in each category.

Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%



Pareto Chart

- Used to portray categorical data (nominal scale).
- A vertical bar chart, where categories are shown in descending order and frequency.
- A cumulative polygon is shown in the same graph.
- Used to separate the 'vital few' from the 'trivial many'.
- See textbook for excel instruction.



Organising Multiple Variable Data

Contingency Table

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

Contingency table for Retirement Fund data: “Type” vs “Risk”

Contingency table results:
Rows: Type
Columns: Risk

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

Is there a pattern or relationship?
If so, what is it?

Contingency table results:
Rows: Type
Columns: Risk

Is there a pattern or relationship?

If so, what is it?

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

	Average	High	Low	Total
Growth	81.32%	76.92%	67.45%	71.84%
Value	18.68%	23.08%	32.55%	28.16%
Total	100.00%	100.00%	100.00%	100.00%

	Average	High	Low	Total
Growth	32.60%	4.41%	63.00%	100.00%
Value	19.10%	3.37%	77.53%	100.00%
Total	28.80%	4.11%	67.09%	100.00%

Pivot Table Version of Contingency Table For Bond Data

First Six Data Points In The Bond Data Set

Fund Number	Type	Assets	Fees	Expense Ratio	Return 2009	3-Year Return	5-Year Return	Risk
FN-1	Intermediate Government	7268.1	No	0.45	6.9	6.9	5.5	Below average
FN-2	Intermediate Government	475.1	No	0.50	9.8	7.5	6.1	Below average
FN-3	Intermediate Government	193.0	No	0.71	6.3	7.0	5.6	Average
FN-4	Intermediate Government	18603.5	No	0.13	5.4	6.6	5.5	Average
FN-5	Intermediate Government	142.6	No	0.60	5.9	6.7	5.4	Average
FN-6	Intermediate Government	1401.6	No	0.54	5.7	6.4	6.2	Average

	A	B	C	D
1	PivotTable of Type and Fees			
2				
3	Count of Fees	Fees		
4	Type	Yes	No	Grand Total
5	Intermediate Government	34	53	87
6	Short Term Corporate	20	77	97
7	Grand Total	54	130	184

Can Easily Convert To An Overall Percentages Table

DCO

	A	B	C	D
1	Contingency Table of Type and Percentages of Fees			
2				
3	Count of Fees	Fees		
4	Type	Yes	No	Grand Total
5	Intermediate Government	18.48%	28.80%	47.28%
6	Short Term Corporate	10.87%	41.85%	52.72%
7	Grand Total	29.35%	70.65%	100.00%

Intermediate government funds are much more likely to charge a fee, compared to Short term corporate.

Can Easily Add Variables To An Existing Table

D

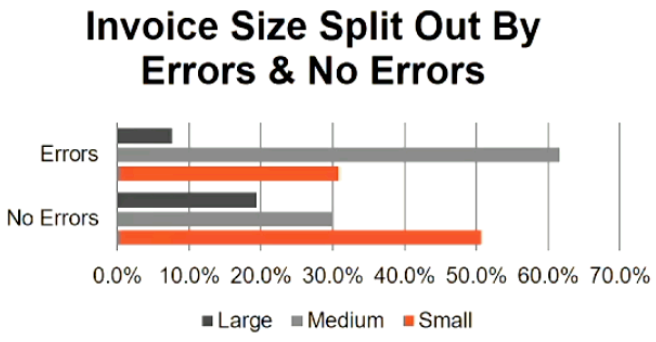
	A	B	C	D	E
1	Multidimensional Contingency Table of Type, Risk, and Fees				
2					
3	Count of Fees		Fees		
4	Type	Risk	Yes	No	Grand Total
5	Intermediate Government	Above average	15	14	29
6		Average	13	19	32
7		Below average	6	20	26
8	Intermediate Government Total		34	53	87
9	Short Term Corporate	Above average	7	23	30
10		Average	7	30	37
11		Below average	6	24	30
12	Short Term Corporate Total		20	77	97
13	Grand Total		54	130	184

Is the pattern of risk the same for all combinations of fund type and fee charge?

Visualising Multiple Variable Data

Side by Side Bar Chart

	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%



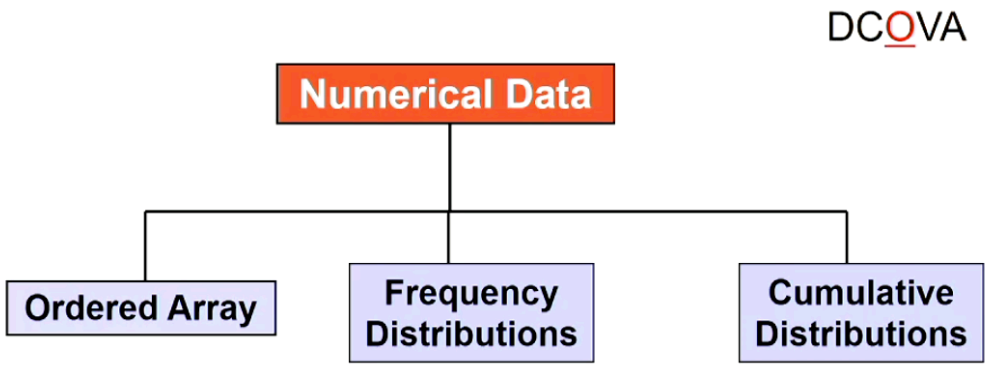
Invoices with errors are much more likely to be of medium size (61.54% vs 30.77% and 7.69%)

- Can do the same with side by side pie chart, however harder to directly compare.
- The side by side bar chart represents the data from a contingency table.

Some Principles of Graphing

- Maximise message, minimise noise.
- Include a title and label axes.
- Include a reference to the source.
- Keep things in correct proportions.
- 3D pie charts are never a good idea.

Organising Numerical Data



Ordered Array

- An ordered array is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value).
- May help identify outliers (unusual observations).

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	

Frequency Distribution

- The frequency distribution is a summary table in which the data are arranged into numerically ordered classes.
- Must give attention to selecting appropriate number of class groupings for the table, determining a suitable width of a class grouping, and establishing the boundaries of each class grouping to avoid overlapping.
- Number of classes depends on number of values on the data.
- To determine the width of a class interval, you divide the range of the data by the number of class groupings desired.

- 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

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

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

DCOVA

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

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%

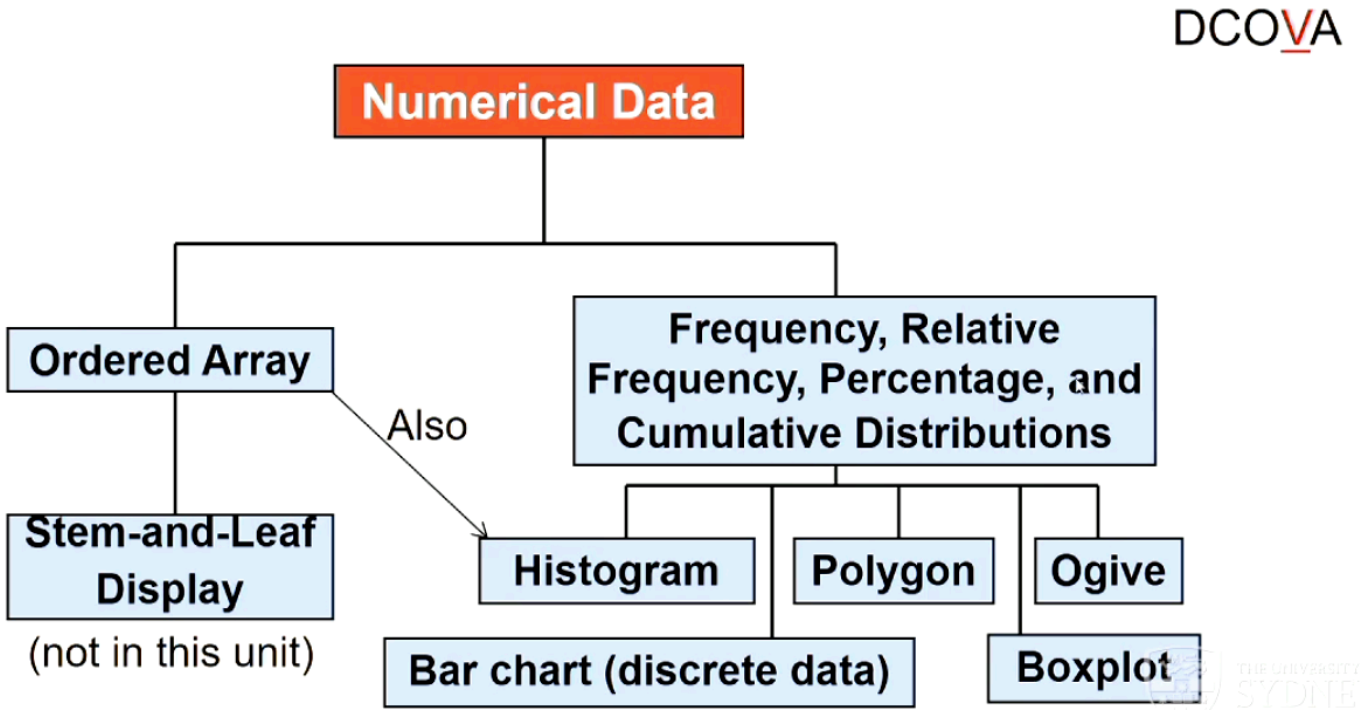
Why Use a Frequency Distribution?

- It condenses the raw data into a more useful form.
- It allows for a quick visual interpretation of the data.
- It enables the determination of the major characteristics of the data set including where the data are concentrated.
- Allows to make a histogram.

Frequency Distribution Tips

- Different class boundaries may provide different pictures for the same data (especially for smaller data sets).
- Shifts in data concentration may show up when different class boundaries are chosen.
- As the size of the data set increases, the impact of alterations in the selection of class boundaries is greatly reduced.
- When comparing two or more data sets, with different sample sizes, you must either use a relative frequency or a percentage distribution.

Visualising Numerical Data

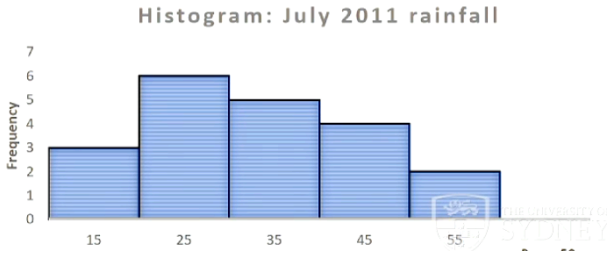


Histogram

- A histogram organises data into groups (called bins) so that the size of the bin reflects the amount or percentage of data points in each group.
- A vertical bar chart of the data in a frequency distribution is called a histogram.
- No gaps between bars, may be gaps for discrete data.
- Class boundaries shown on horizontal axis.
- Vertical axis is either frequency, relative frequency or percentage.
- Height of bars represent the frequency.

Class	Frequency	Relative Frequency	Percentage
>10 but less than 20	3	0.15	15
>20 but less than 30	6	0.30	30
>30 but less than 40	5	0.25	25
>40 but less than 50	4	0.20	20
>50 but less than 60	2	0.10	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

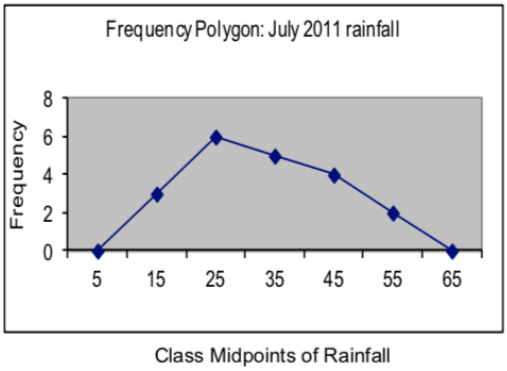
- A frequency or percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages
- The cumulative frequency or cumulative percentage polygon, or ogive, displays the variable of interest along the x axis and the cumulative percentages along the Y axis

Visualizing Numerical Data: The Frequency Polygon

DCOVA

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
Total		20

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

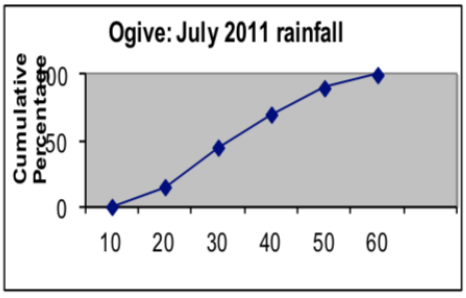


Visualizing Numerical Data: The Ogive (Cumulative % Polygon)

DCOVA

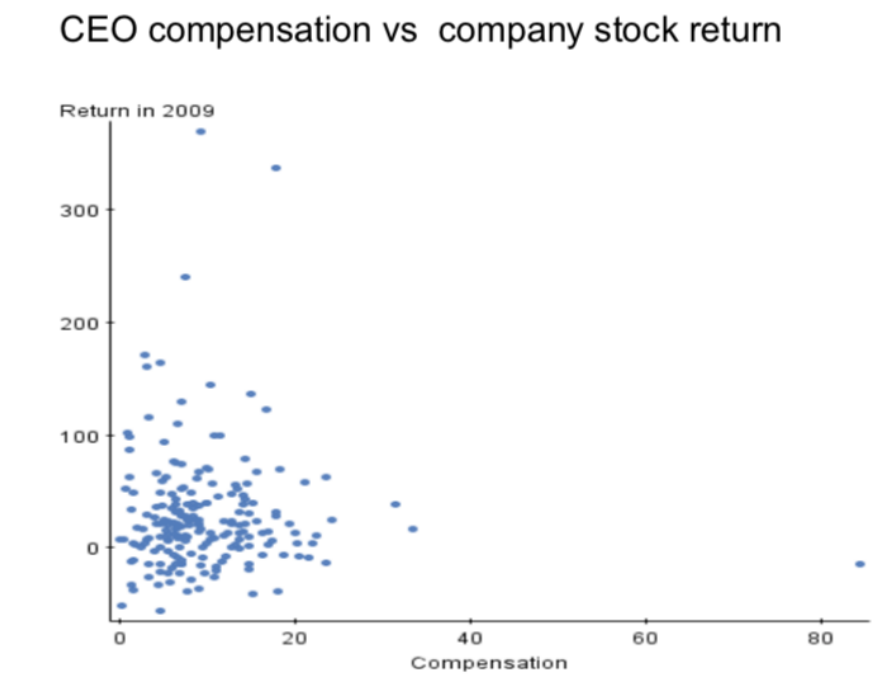
Class	Lower class boundary	% less than lower class boundary
>10 but less than 20	10	0
>20 but less than 30	20	15
>30 but less than 40	30	45
>40 but less than 50	40	70
>50 but less than 60	50	90
>60 but less than 70	60	100

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



Scatter Plots

- Scatter Plots are used for numerical data consisting of paired observations taken from two numerical variables.
- One variable is measured on the vertical axis and the other variable is one the horizontal axis.
- Scatter plots are used to examine possible relationships between two numerical variables.

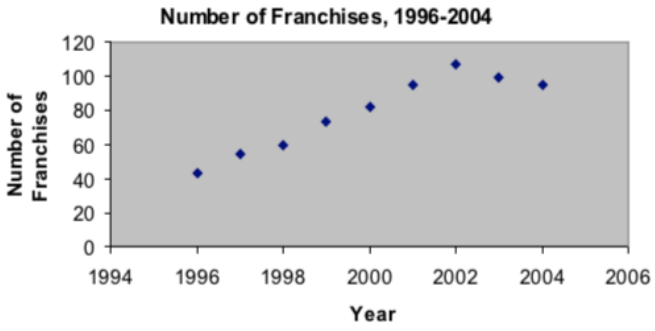


The Time Series Plot

- Time series plots are used to study patterns in values of a numeric variable over time.
- The numeric variable is measured on the horizontal axis.

Time Series Plot Example

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 Data

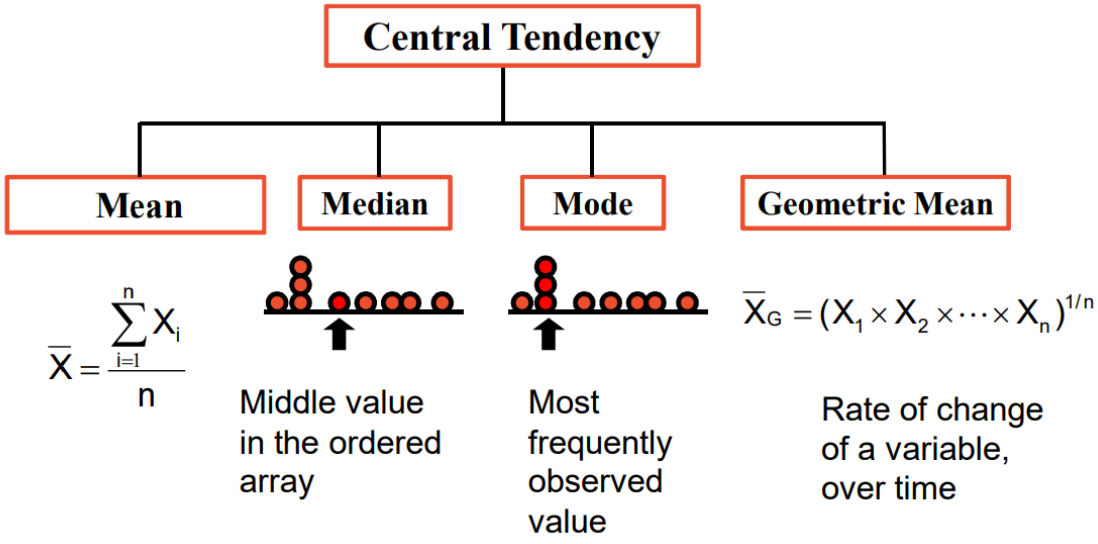
- The graph should not distort the (information in the) data.

- The graph should not contain (too many) unnecessary adornments.
- The scale on vertical axis should (usually) begin at zero.
- All axes should be properly and clearly labelled.
- The graph should contain an informative title.
- The simplest possible graph should usually be used for a given set of data.
- The graph should contain the source of the data.
- 3D graphs should have a meaningful 3rd dimension, otherwise 2D.
- Graphs should be used to objectively and clearly convey the message (relevant info) in the data.

BUSS1020 Week 3 Lecture – Numerical Descriptive Measures

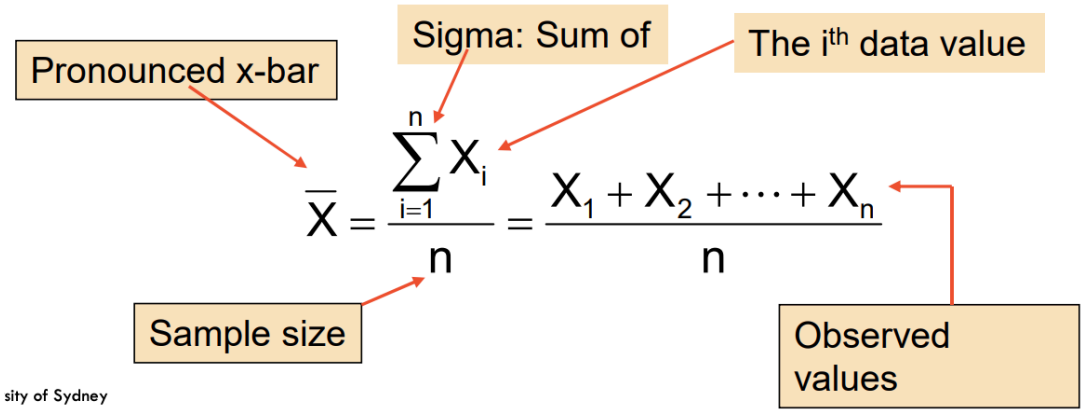
- The central tendency is the extent to which all the data values group around a typical or central value.
- The variation is the amount of dispersion or scattering of values around the central value.
- The shape is the pattern in the distribution of values from the lowest value to the highest value.

Measures of Central Tendency



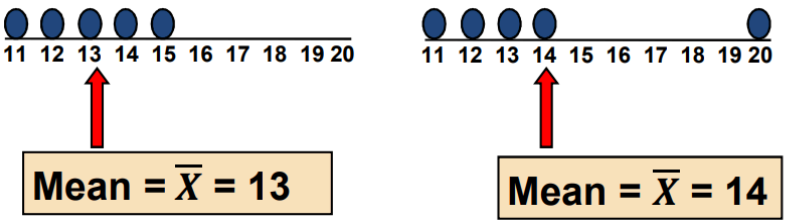
Mean or Average

- The arithmetic mean is the most common measure of central tendency → for a sample of size n, the sample mean is:



sity of Sydney

- It is only useful for numerical data and can be affected by extreme values like outliers.



$$\overline{X} = \frac{11 + 12 + 13 + 14 + 15}{5} = \frac{65}{5} = 13 \qquad \frac{11 + 12 + 13 + 14 + 20}{5} = \frac{70}{5} = 14$$

Median

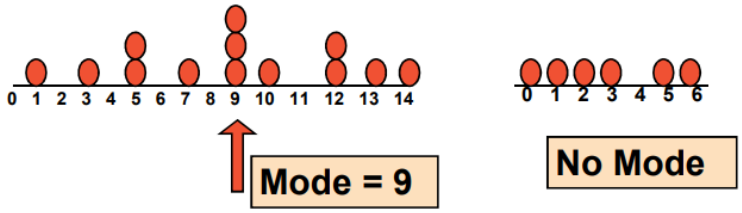
- In an ordered array, the median is the 'middle' number (50% above, 50% below).
- It is not affected by extreme values and can be used for numerical or categorical ordinal data.
- The location of the median when the values are in numerical order (smallest to largest) is, noting that it only gives the position and not the value:

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- 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.

Mode

- Values that occurs most often and is not affected by extreme values.
- Used for either numerical or categorical data → sometimes there is no mode or there is more than one mode.



Geometric Mean and Geometric Rate of Return

- Geometric mean is often used to measure the rate of change of a variable over time.

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{\frac{1}{n}}$$

- Geometric mean for rate of return measures the status of an investment over time.

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{\frac{1}{n}} - 1$$

Where R_i is the rate of return in time period i

The Geometric Mean Rate of Return: Example

DCOVA

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

The overall two-year per year return is zero, since it started and ended at the same level. Right?

The Geometric Mean Rate of Return: Example *(continued)*

DCOVA

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\bar{X} = \frac{(-.5) + (1)}{2} = 0.25 = 25\%$$

Misleading result

Geometric mean rate of return:

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{\frac{1}{n}} - 1$$

$$= \left[\left(1 + (-0.5)\right) \times \left(1 + (1)\right)\right]^{1/2} - 1$$

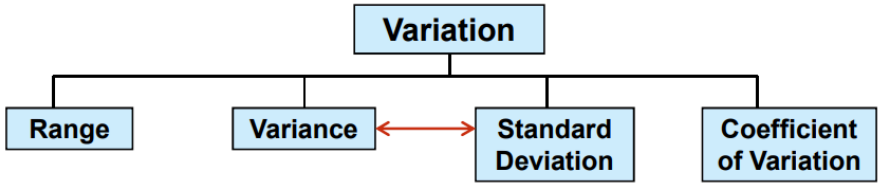
$$= [(0.5) \times (2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%$$

More representative result

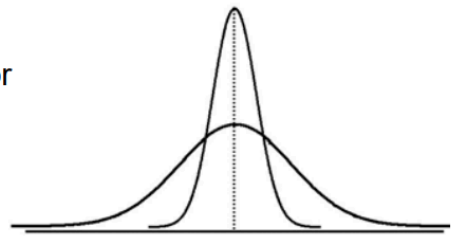
Which Measure to Choose?

- Mean is most often used, unless extreme values (outliers) exist. Can't be used for categorial data.
- Median is the next most popular, since it is not sensitive to outliers e.g. median house prices may be reported for a region. Can also be used for ordinal data.
- Sometimes both the mean and median are used.
- Mode is usually reported for discrete or categorical data only.
- Geometric mean (return) is useful to measure and track percentage changes.

Measures of Variation



- give information on the **spread** or **variability** or **dispersion** of the data values.



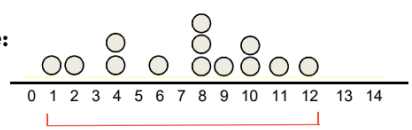
Same centre, differing variation

Range

- The simplest one, it can be used for all numerical data and is the difference between the largest and smallest values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:

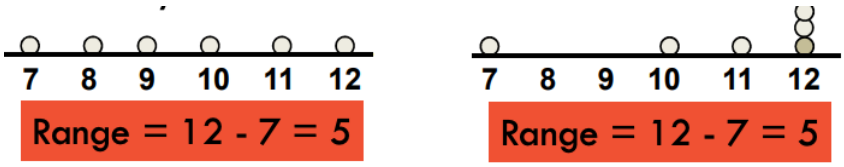


Student test scores

Range = 12 - 1 = 11

Why Range Can be Misleading

- Ignore the way in which data is distributed.



- Highly sensitive to outliers.

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,4,**120**

Range = 120 - 1 = 119

The Sample Variance

- Average of squared deviations of values from the sample mean.

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

- Variance forms a major component of modern financial investment and risk management practice.

  → Variance measures the variability (volatility) from an average. Volatility is a measure of risk, so this statistic can help determine the risk an investor might take on when purchasing a specific security.

  → Modern portfolio theory says than portfolio variance can be reduced by choosing asset classes with a low or negative correlation, such as stocks and bonds. This type of diversification is used to reduce risk.

  → Hedging is an investment that reduces the risk of adverse price movements in an asset.

The Sample Standard Deviation

- Most commonly used measure of variation, it shows variation about the sample mean.
- It has the same units as the original data, and is the square root of the variance.

– Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

The standard deviation for daily asset returns is typically close to 1%, in most periods.

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

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

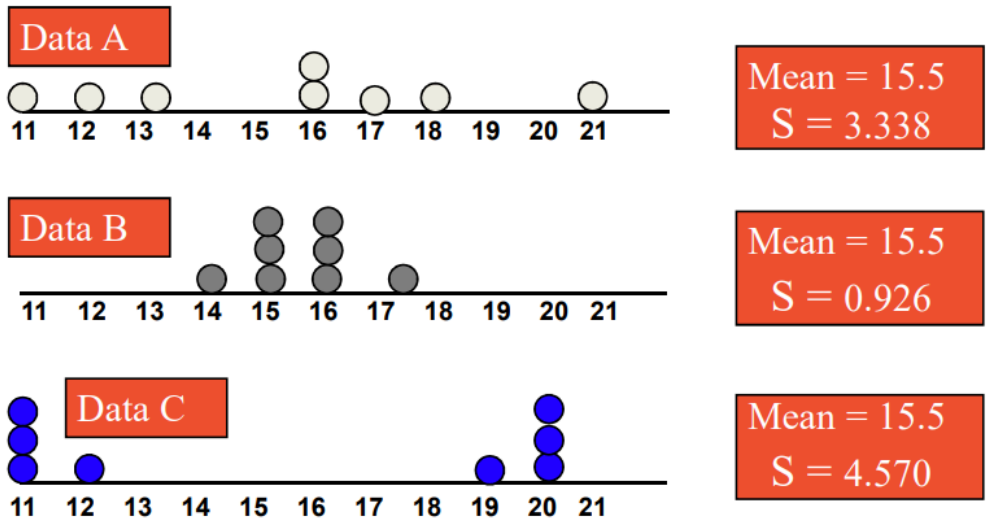
$$S = \sqrt{\frac{(10-\overline{X})^2+(12-\overline{X})^2+(14-\overline{X})^2+\cdots+(24-\overline{X})^2}{n-1}}$$

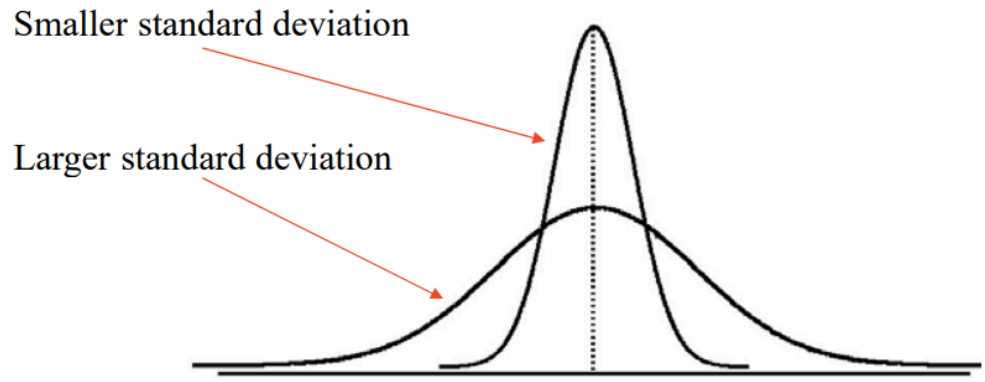
$$= \sqrt{\frac{(10-16)^2+(12-16)^2+(14-16)^2+\cdots+(24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

→ A measure of the "average"/typical scatter around the mean

The University of Sydney





Summary Characteristics

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

The Coefficient of Variation

- Measures relative variation, always in percentage (%).
- Shows variation relative to mean, and can be used to compare the variability of two or more data sets measured in different units → CV = (S/X) . 100%

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its average price

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock C:
  - Average price last year = $8
  - Standard deviation = $2

$$CV_C = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

Stock C has a much smaller standard deviation but a much higher coefficient of variation

Assessing Extreme Observations: Sample Z-Score

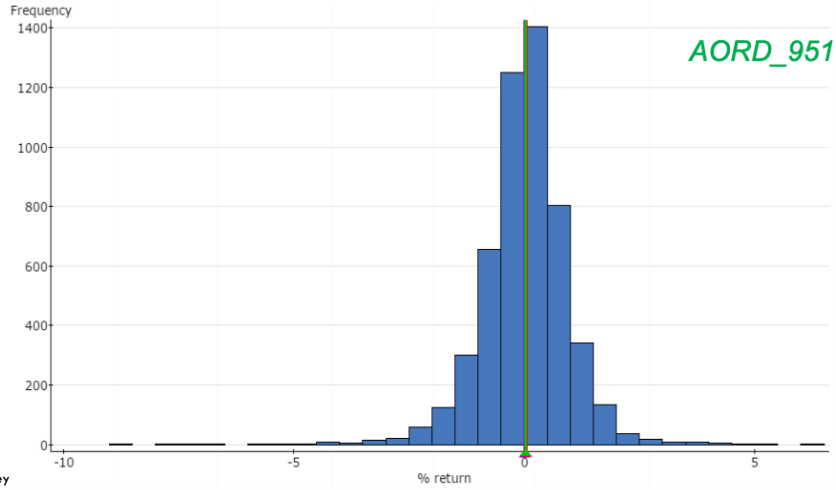
- To compute the Z-score of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value could be considered extreme if its X-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the X-score, the farther the data value is from the mean.

$$Z = \frac{X - \bar{X}}{S}$$

where X represents the data value
$\bar{X}$ is the sample mean
S is the sample standard deviation

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Median	Range	Min	Max	Coef. of var.	Mode
% return	5210	0.0205	0.886	0.941	0.044	14.62	-8.55	6.07	4580.3	0



AORD_9515.xls

Locating Extreme Outliers: Z-Score

Sample Question:

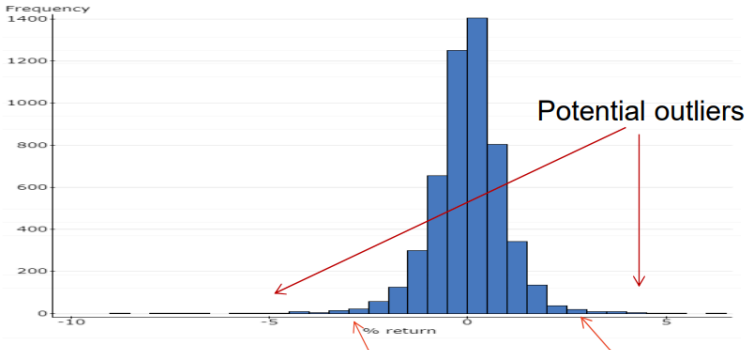
- The return on 24th July, 2015 was -0.4399%. Compute its' Z-score and hence determine if it is an outlier.
- The sample mean is 0.0205, S = 0.941

Answer:

$$Z = \frac{X - \overline{X}}{S} = \frac{-0.4399 - 0.0205}{0.941} = -0.489$$

A return of -0.4399 is 0.489 standard deviations below the mean. It would **not** be considered an outlier or extreme return.

- The return on 30th September 2008 was -4.39%. It's Z-score is -4.687, hence meaning it is a potential outlier.



AORD_9515.xls

Mean: 0.022 % Standard deviation: 0.94 %

$\overline{X} - 3s = -2.80\%$ $\overline{X} + 3s = 2.84\%$

0.8% of returns were below this value

0.6% of returns were above this

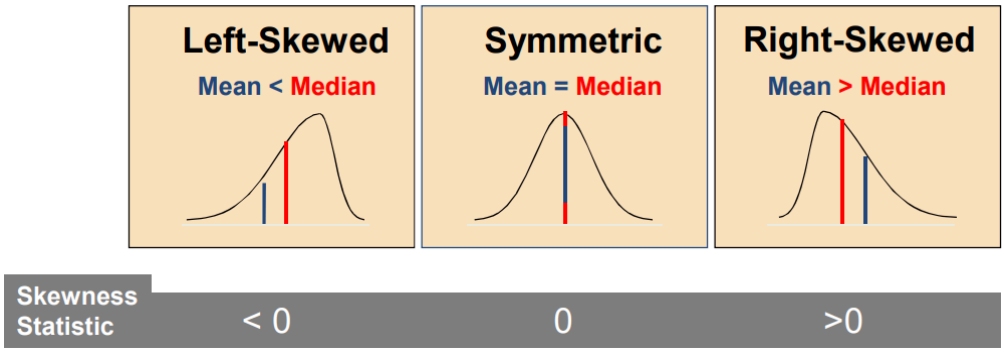
Shape of a Distribution

- Describes how data are distributed, with two useful shape related statistics being:
- Skewness → measures the amount of asymmetry in a distribution.
- Kurtosis → measures the relative concentration values in the centre of a distribution as compared with the tails.

$$Skewness = \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})^3}{S^3} \quad ; \quad Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})^4}{S^4} - 3$$

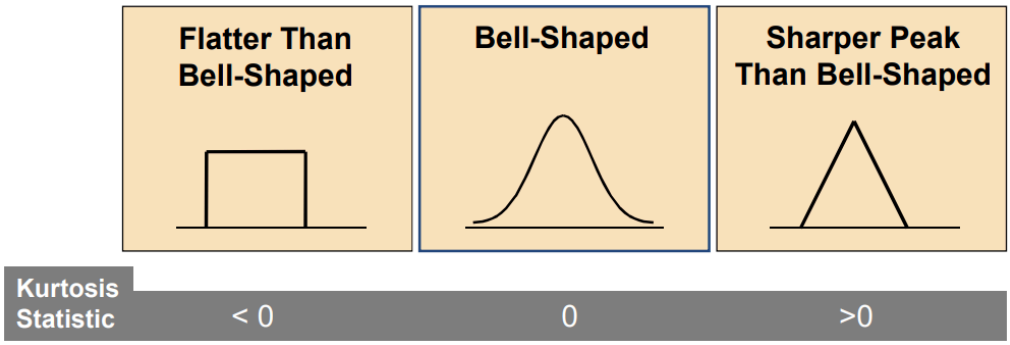
Skewness

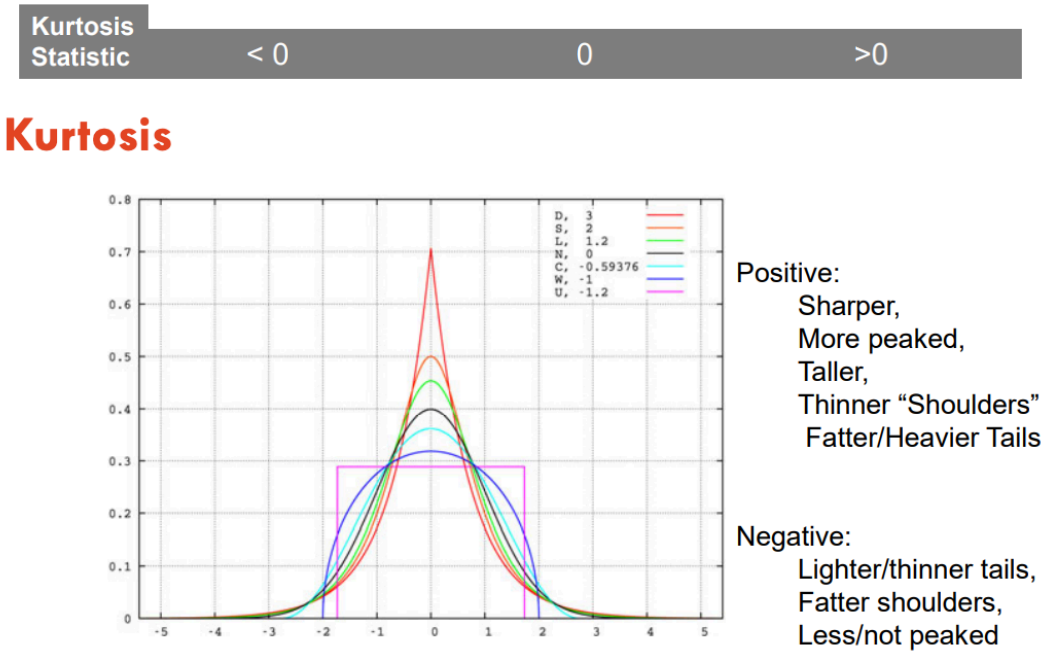
- Describes the amount of asymmetry in a distribution → symmetric or skewed.



Kurtosis

- Describes relative concentration of values in the centre as compared to the tails.





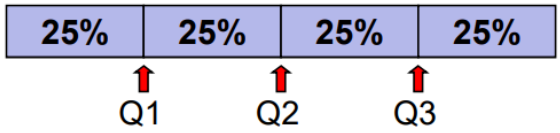
General Descriptive Stats Using Microsoft Excel Functions

DCOVA

	A	B	C	D	E
1	House Prices		Descriptive Statistics		
2	$ 2,000,000		Mean	$ 600,000	=AVERAGE(A2:A6)
3	$ 500,000		Standard Error	$ 357,770.88	=D6/SQRT(D14)
4	$ 300,000		Median	$ 300,000	=MEDIAN(A2:A6)
5	$ 100,000		Mode	$ 100,000	=MODE(A2:A6)
6	$ 100,000		Standard Deviation	$ 800,000	=STDEV(A2:A6)
7			Sample Variance	6.4E+11	=VAR(A2:A6)
8			Kurtosis	4.130126953	=KURT(A2:A6)
9			Skewness	2.006835938	=SKEW(A2:A6)
10			Range	$ 1,900,000	=D12 - D11
11			Minimum	$ 100,000	=MIN(A2:A6)
12			Maximum	$ 2,000,000	=MAX(A2:A6)
13			Sum	$ 3,000,000	=SUM(A2:A6)
14			Count	5	=COUNT(A2:A6)

Quartile Measures

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



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than Q_3

- Find a quartile by determining the value in the appropriate position in the ranked data, where:

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

- When calculating the ranked position, use the following rules:

→ If the result is a whole number, then use its ranked position.

→ If the result is a fractional half (2.5, 7.5, 8.5 etc.), then average the two corresponding data values.

→ If the result is not a whole number nor a fractional half, then round the result to the nearest integer to find the ranked position e.g. 1.25 → 1, 5.75 → 6.

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 = 5th 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

The Interquartile Range (IQR)

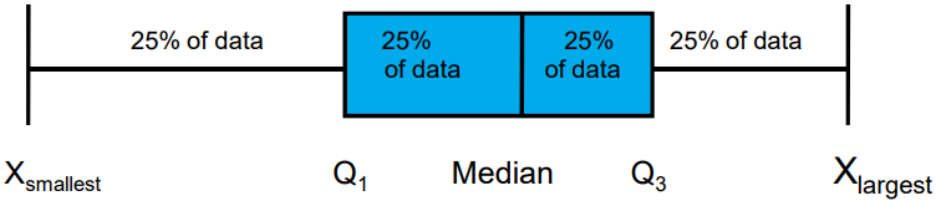
- The IQR, or mid-spread, is $Q_3 - Q_1$, measuring the spread in the middle 50% of the data.
- It is a measure of variability, which is not influenced by outliers or extreme values.
- Measures like Q_1, Q_3, and IQR that are not influenced by outliers are called resistant or robust measures.
- The IQR of the example above is 19.5 – 12.5 = 7.

Five Number Summary and the Boxplot

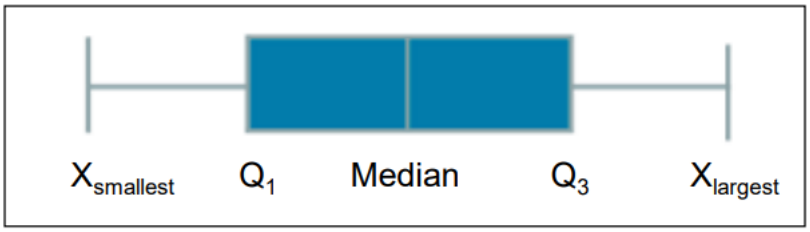
- The five numbers that help describe the centre, spread and shape of the data are:
  - $X_{smallest}$
  - First Quartile (Q_1)
  - Median (Q_2)
  - Third Quartile (Q_3)
  - $X_{largest}$
- The boxplot is a graphical display of the data based on the five-number summary:

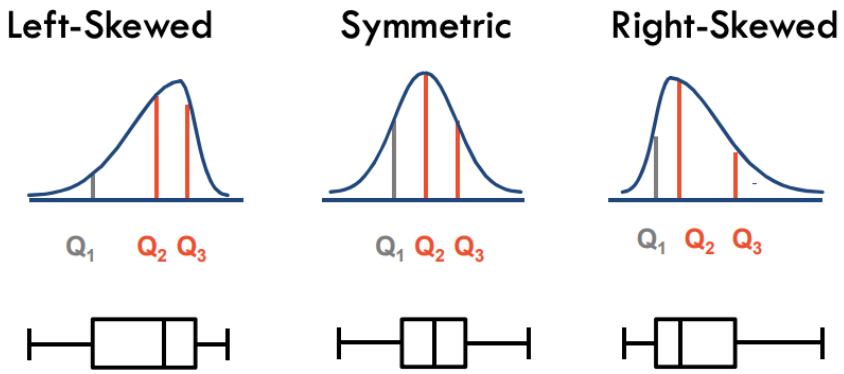
$X_{smallest}$ – Q_1 – Median – Q_3 – $X_{largest}$

Example:



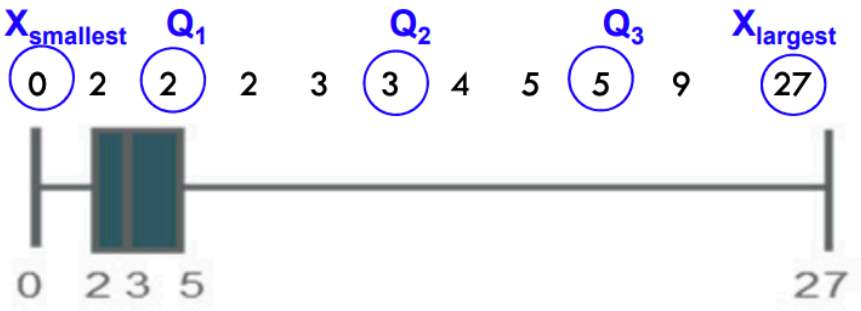
- If the data is symmetric around the median, then the box and central line are centred between the endpoints.
- A boxplot can be shown in either a vertical or horizontal orientation.





Mean<Median

- Below is a Boxplot for the following data:



- The data are right skewed, as the plot depicts (Mean > Median)

Population Measures

Numerical Descriptive Measures for a Population

- Descriptive statistics discussed previously described a sample, not the population.
- Summary measures describing a population, called parameters, are denoted with Greek letters.
- Important population parameters are the population mean, population variance and the population standard deviation.
- Note that n (or n-1) is replaced by N (=population size).

– The **population mean**

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

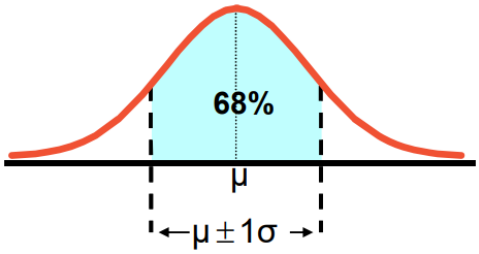
– The **population variance**

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

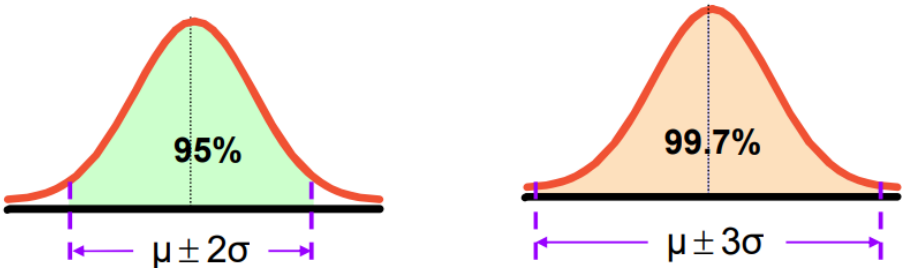
Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

The Empirical Rule

- The empirical rule approximates the variation of data in a bell-shaped distribution.
- Approximately 68% of the data in a bell shaped distribution is within $\pm$ one standard deviation of the mean or $\mu \pm 1\sigma$.



- Approximately 95% of the data in a bell-shaped distribution lies within $\pm$ two standard deviations of the mean, or $\mu \pm 2\sigma$.
- Approximately 99.7% of the data in a bell-shaped distribution lies within $\pm$ three standard deviations of the mean, or $\mu \pm 3\sigma$.



- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then,
  - 68% of all test takers scored between 410 and 590 (500 $\pm$ 90).
  - 95% of all test takers scored between 320 and 680 (500 $\pm$180).
  - 99.7% of all test takers scored between 230 and 770 (500 $\pm$ 270).

Chebyshev's Rule

- Regardless of how the data is distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for k > 1).
  - **Examples:**

	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 $(\mu \pm 3\sigma)$

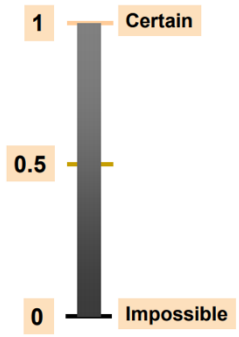
  - **Suppose that a financial return variable is not bell-shaped distributed, but has $\mu = 0.5\%$ and $\sigma = 1\%$. Then,**
    - **(k=2) At least 75% of all returns will lie between -1.5% and 2.5% (0.5 ± 2*1).**
    - **(k=3) At least 89% of all returns will lie between -2.5% and 3.5% (0.5 ± 3*1).**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

Ethical Considerations

- Numerical descriptive measures should be presented in a fair, objective and neutral manner.
- You should include both good and bad results.
- It is unethical to selectively remove relevant information that is detrimental to supporting a particular position.

BUSS1020 Week 4 Lecture – Basic Probability

Probability Concepts



- Probability: the chance 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).

Assessing Probability

- An example of three approaches to assess the probability of an uncertain (discrete or categorical) event:

1. *a priori -- based on prior knowledge of the process*

$$\text{probability of occurrence} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of outcomes}}$$

2. empirical probability -- based on observed data

$$\text{probability of occurrence} = \frac{\text{number of ways the event } \textit{has} \text{ occurred}}{\text{number of trials}}$$

3. subjective probability

Probability of occurrence is based on a combination of an individual's past experience, personal opinion, and/or analysis of a particular situation

A Priori Probability

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 ?

$$\text{Probability of being in top 10} = \frac{X}{T} = \frac{\text{number of players in top 10}}{\text{total number of opponents}}$$

$$\frac{X}{T} = \frac{10 \text{ top 10 players}}{99 \text{ possible opponents}} = \frac{10}{99} = 0.101$$

Empirical Probability

What is the probability your financial asset's price increases from today to tomorrow?

$$\text{Probability of price increasing} = \frac{X}{T} = \frac{\text{number of days price increased}}{\text{total number of days considered}}$$

$$\frac{X}{T} = \frac{121 \text{ days of increases}}{250 \text{ days considered}} = \frac{121}{250} = 0.484$$

Subjective Probability

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 is called an event.

Simple Event

- Described by a single characteristic.
- e.g. a customer plans to purchase a product

Joint Event

- An event described by two or more characteristics.
- e.g. a customer plans to purchase a product AND pay more than $100

Complement of an Event A

- Denoted as A' → All events that are not part of event A.
- e.g. the customer does not plan to purchase the product.

Sample Space

- The collection of all possible events → different, mutually exclusive and collectively exhaustive.
- e.g. all 6 faces of a die, all 52 cards in a deck, a customer plans or does not plan to purchase.

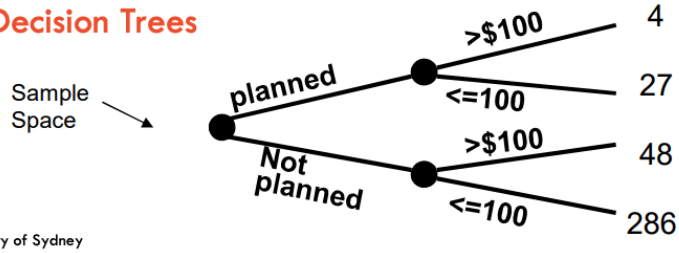
Visualising Events

– Contingency Tables -- For All Customers

Paid	Planned Purchase	Not planned purchase	Sub-total
>$100	4	48	52
<=$100	27	286	313
Sub-Total	31	334	365

Sample size

– Decision Trees



Basic Probability

Marginal (Simple) Probability

- Refers to the probability of a simple event.
- e.g. P(planned to purchase a specific product), P(paid > $100).
  - (found in the *margin* of the table)

Paid	Planned Purchase	Not planned purchase	Sub-total
>$100	4	48	52
<=$100	27	286	313
Sub-Total	31	334	365

P(Paid > $100) = 52 / 365

P(Planned to Purchase) = 31 / 365

Joint Probability

- Refers to the probability of an occurrence of two or more events (joint event).
- e.g. P(planned purchase and paid > $100), P(not planned purchased and paid > $100).

Paid	Planned Purchase	Not planned purchase	Sub-total
>$100	4	48	52
<=$100	27	286	313
Sub-Total	31	334	365

P(Not Planned & paid < $100) = 286 / 365

P(Planned Purchase & paid > $100) = 4 / 365

Mutually Exclusive Events

- Mutually exclusive events are events that cannot occur simultaneously.
- e.g. choosing one day from 2018, A: All days in January, B: All days in March.

Collectively Exhaustive Events

- Collectively exhaustive events are ones where one of the events must occur, with the set of collectively exhaustive events covering the entire sample space.

Example: Randomly choose a day from 2018

A = Weekday; B = Weekend;
C = January; D = Spring;

- Events A, B, C and D are collectively exhaustive (but not mutually exclusive – a weekday can be in January or in Spring)
- Events A and B are **collectively exhaustive and also mutually exclusive**

$P(X) + P(Y) + P(Z) = 1$

If X, Y, and Z are *mutually exclusive* and *collectively exhaustive*

Calculating Probabilities

Computing Joint and Marginal Probabilities

- The empirical probability of a joint event, A and B:

$$P(\text{A and B}) = \frac{\text{number of outcomes satisfying A and B}}{\text{total number of elementary outcomes}}$$

- Computing a marginal (or simple) probability from joint probabilities:

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

  - Where B_1, B_2, …, B_k are k mutually exclusive and collectively exhaustive events

Joint Probability Example

P(PPurchase and spend > $100)

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

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

Marginal Probability Example

P(customer spent > $100)

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

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

Marginal & 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

General Addition Rule

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

General Addition Rule Example

P(PPurch OR paid >$100) = P(PPurch.) + P(>$100) - P(PPurch & >$100)

= 31/365 + 52/365 - 4/365 = 79/365

Paid	Planned Purchase	Not planned purchase	Sub-total
>$100	4	48	52
<=$100	27	286	313
Sub-Total	31	334	365

Don't count the 4 purchases with spend > $100 twice!

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

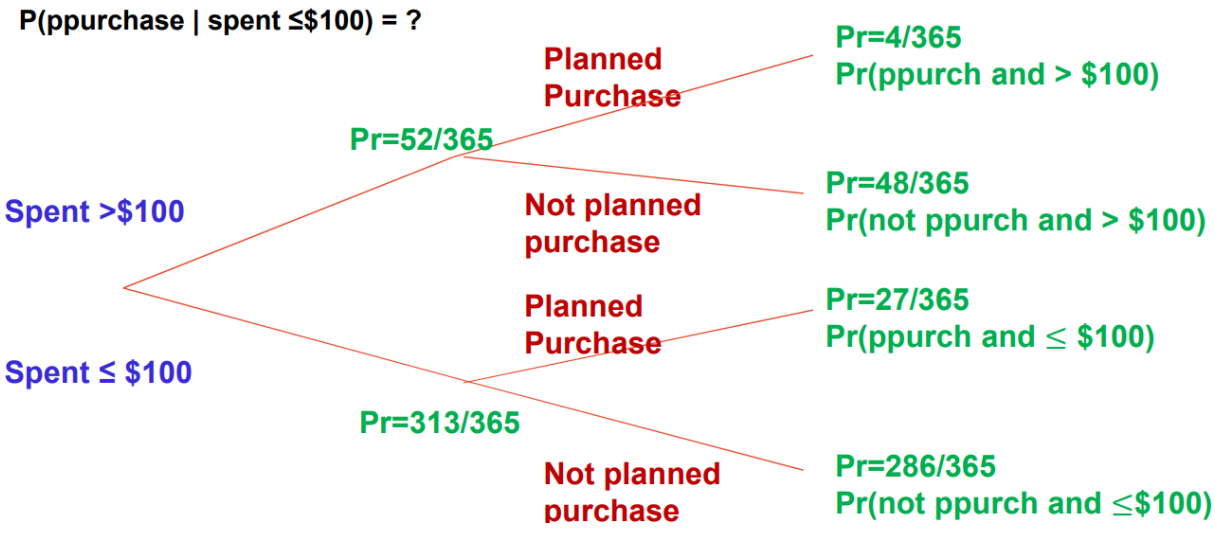
Where P(A and B) = joint probability of A and B
P(A) = marginal or simple probability of A
P(B) = marginal or simple probability of B

- Examples of this include the conditional probability that:
  - → a customer will buy Book B after buying Book A
  - → a customer will clickthrough based on a particular web page design.
  - → an asset price will go up following a positive earnings result.
  - → a customer churns following a customer interaction event by a company.
  - → a customer actually purchases after answering a survey question saying they plan o purchase.

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

	PPurchase	Not PPurchase	Total
> $100	4	48	52
≤ $100	27	286	313
Total	31	334	365

Decision Trees



Conditional Probability Example

- Of the cars on a used car lot, 90% have air conditioning (AC) and 40% have a GPS. 35% of the cars have both.

– What is the probability that a car has a GPS given that it has AC ?

i.e., we want to find P(GPS | AC)

- Of the cars on a used car lot, **90%** have air conditioning (AC) and **40%** have a GPS.
35% of the cars have both.

	GPS	No GPS	Total
AC	0.35	0.55	0.90
No AC	0.05	0.05	0.10
Total	0.40	0.60	1.00

- Given AC, we only consider the top row (90% of the cars). Of these, 35% have a GPS. 35% of 90% is about 38.89%.

	GPS	No GPS	Total
AC	0.35	0.55	0.90
No AC	0.05	0.05	0.10
Total	0.40	0.60	1.00

$$P(GPS \mid AC) = \frac{P(GPS \text{ and } AC)}{P(AC)} = \frac{0.35}{0.90} = 0.3889$$

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 by the fact that the other event has occurred.

Multiplication Rules

- Multiplication rule for two events A and B:

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

- Note: if A and B are independent, then $P(A \mid B) = P(A)$ and the multiplication simplifies to:

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

Marginal Probability

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)$$

- **Where $B_1, B_2, \ldots, B_k$ are k mutually exclusive and collectively exhaustive events**

Bayes' Theorem

- Bayes' Theorem is used to revise previously calculated probabilities based on new information.
- Developed by the Reverend Thomas Bayes in the 18th Century.
- It is an extension of conditional probability.
- It allows us to usefully reverse the conditioning between two events or variables.

Simple Form

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

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

General Form

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

where:

- **B_i = i^{th} event of k mutually exclusive and collectively exhaustive events**
- **A = new event that might impact P(Bi)**

Bayes' Theorem Example

- M&R Electronics has observed a 40% success rate for new model TVs.
- A market research firm has issued a favourable report on the new model TV.
- In the past, 80% of successful new TVs received a favourable report; only 30% of unsuccessful TVs received a favourable report.
- The new model TV has a favourable report.
- What is the probability that the new TV will be successful?

– Let S = A new TV is *successful*

U = A new TV is *unsuccessful*

F = The report is *favourable*

From the question:

– P(S) = 0.4 , P(U) = 0.6 (Prior probabilities)

– P(F | S) = 0.8, (F | U) = 0.3 (Conditional probabilities)

– **Goal is to find P(S | F)** (Revised probability)

Apply Bayes' Theorem:

$$P(S|F) = \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|U)P(U)}$$

$$= \frac{(0.8)(0.4)}{(0.8)(0.4) + (0.3)(0.6)}$$

$$= \frac{0.32}{0.32 + 0.18} = 0.64$$

So the revised probability of success, given that the market research report is favourable, is 0.64

– Given the favourable report, the revised probability of a successful model TV has risen to 0.64, up from the original estimate of 0.4

Event	Prior Prob.	Conditional Prob of F	Joint Prob.	Revised Prob given F
S (successful)	0.4	0.8	(0.4)(0.8) = 0.32	0.32/0.5 = 0.64
U (unsuccessful)	0.6	0.3	(0.6)(0.3) = 0.18	0.18/0.5 = 0.36

Sum = 0.5

Counting Rules

- Often discrete probabilities rely on being able to count all the events possible.
- Recall a priori probability: P(event) = number of ways the event can occur/total number of outcomes.
- When there are many, many possible events, this can be tricky and time consuming.
- Counting rules can help!

Counting Rule 1

- If any one of k events can occur on each of n trials, the number of possible outcomes is equal to:

$$k^n$$

- Example: if you roll a fair die 3 times, then there are $6^3 = 216$ possible outcomes.
- Example: if you consider whether 10 customers will purchase or not, there are $2^{10} = 1024$ possibilities.

Counting Rule 2

- If the 1st trial has k_1 events, 2nd has k_2 events, and the nth trial has k_n events, then the number of possible outcomes in the n trials is:

$$(k_1)(k_2)\cdots(k_n) = \prod_{i=1}^{n} k_i$$

- Example: A CEO needs to attend a meeting, a business lunch and see a presentation. But there are 3 meetings, 4 lunches at different restaurants, and 6 presentations need to be seen. How many different possible combinations of these events are there for the CEO to attend?
- Answer: (3)(4)(6) = 72 different possibilities.
- Example: A customer can either plan to purchase or not and then spend <$100 or >=$100 in total.
- This gives 2*2 = 4 possible outcomes in total.
- Example: In the Monty Hall problem, a player can pick one of 3 doors, then switch or not switch doors, and then either win a car or a goat:
- This is 3*2*2 = 12 possible outcomes

Counting Rule 3

- The number of ways that n items can be arranged in unique order is:

$$n! = (n)(n-1)\cdots(1)$$

- Example: you have five books to put on a bookshelf. How many different ways can these books be placed on the shelf?
- Answer: 5! = (5)(4)(3)(2)(1) = 120 different possibilities.

Counting Rule 4

- Permutations: the number of ways of arranging X objects selected from n objects in order is:

$${}_nP_x = \frac{n!}{(n-X)!} = n(n\text{-}1)\ldots(n\text{-}X+1)$$

- Example: You have five books and are going to put three on a bookshelf. How many different ways can the books be ordered on the bookshelf?
- Answer:

$${}_nP_x = \frac{n!}{(n-X)!} = \frac{5!}{(5-3)!} = 5.4.3 = 60$$

Counting Rule 5

- Combinations: the number of ways of selecting X objects from n objects, irrespective of order, is:

$${}_nC_x = \frac{n!}{X!(n-X)!}$$

- Example: you can have five books and are going to randomly select three to read. How many different sets of 3 books might you select?
- Answer:

$$_nC_x = \frac{n!}{X!(n-X)!} = \frac{5!}{3!(5-3)!} = \frac{120}{(6)(2)} = 10$$

Ethical Issues

- Ethical issues arise when probability statements can be misleading – they are easily misinterpreted by people without an understanding of probability.
- Taking this medical therapy increases your chances of one type of cancer by 20% → chance of getting that cancer?
- Percentages of percentages are problematic without marginal probabilities.
- This investment strategy gives you an 83% chance of increasing your annual return → what happens in the other 17% of cases?
- This test detects condition A with 99% accuracy → gives $P(+|A)$: wrong conditioning, needs Bayes' rule!

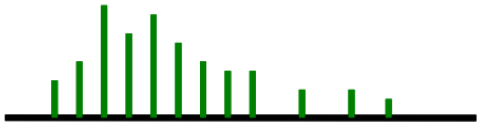
BUSS1020 Week 5 Lecture – Discrete Distributions

Random Variables

- A random variable (rv) represents the possible outcomes from an uncertain event.
- Numerical rvs can be discrete or continuous.

Discrete RV

- Set of all possible outcomes is a finite, or "countably infinite", number of values.
  - → Number of new subscribers to a magazine
  - → Number of fraudulent credit card transactions per month
  - → Number of absent employees per day
  - → Prices and price changes (but sometimes treated as continuous).



Continuous RV

- Takes values at every point in a given interval
  - → Temperature
  - → Elapsed time between arrivals of bank customers
  - → Rate of unemployment
  - → Financial return, percentage changes

Discrete Random Variables

- Can only a countable number of values
  - → Roll a dice twice. Let X be the number of times 4 occurs.
  - → Toss a coin 5 times. Let X be the number of heads.

Probability Distribution for a Discrete Random Variable

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

Outcomes	Probability
X_1	p_1
X_2	p_2
:	:
X_M	p_M

$$\sum_{i=1}^{M} p_i = 1$$

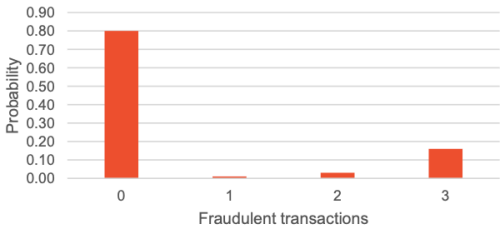
Example: Probability Distribution for a Discrete Random Variable (r.v.)

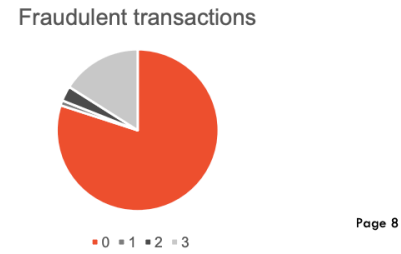
Number of fraudulent credit card transactions per month	Probability
0	0.80
1	0.01
2	0.03
3	0.16

Ed question:

What type of probabilities are these?

Add to 1

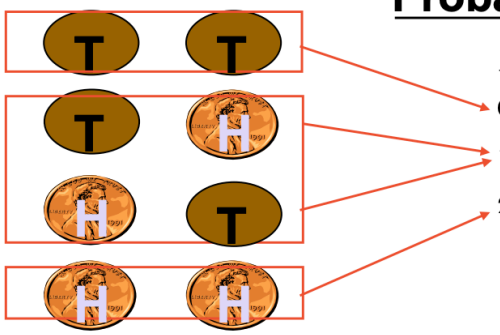




Page 8

Experiment: Toss 2 Coins Let X = # heads.

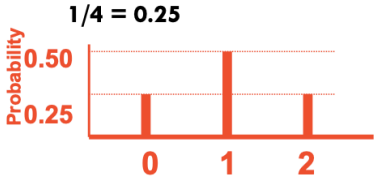
4 possible outcomes



Probability Distribution

X Value	Probability
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25

Type of probabilities ?



The University of Sydney

Page 9

Expected Value (Average, Mean)

- Note that the expected value may not be a possible value for X.

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

Example: Toss 2 coins:

X = # of heads

Compute the expected value of X:

E(X) = (0)(0.25) + (1)(0.50) + (2)(0.25) = 1.0

X	$P(X=x_i)$
0	0.25
1	0.50
2	0.25

X=1 is also the mode and the median

Measuring Dispersion

- **Variance of a discrete rv**

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

- **Standard Deviation of a discrete rv**

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

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

Example: No. Traffic accidents per week, compute the standard deviation

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

$$\sigma = \sqrt{(0 - 1.36)^2(0.2) + (1 - 1.36)^2(0.4) + (2 - 1.36)^2(0.24) + (3 - 1.36)^2(0.16)}$$
$$= \sqrt{0.95} = 0.975 \text{ accidents}$$

X=no. accidents	P(X)
0	0.20
1	0.40
2	0.24
3	0.16

Binomial Probability Distribution

- Used where a rv X counts the number of "events of interest" occurring from a fixed number of observations or trials (denoted n).
  - → n=2 tosses of a coin, counting heads
  - → 10 light bulbs taken from a warehouse, counting defects
  - → 5 trading days counting "high volatility" days
  - → 25 customers, counting "churns"

Requirements

- Each observation is categorised as to whether or not the 'event of interest' occurred or not.
  - → a customer 'churns' or not
  - → a customer purchase or not

The die lands on '6' or not

- There must be a fixed number of trials, denoted n.
- The two categories are mutually exclusive and collectively exhaustive.
- Each observation has constant probability for the event of interest occurring (denoted π):

  → P(head) is the same (0.5) each time we toss a coin

  → Is P(defective) the same each time we test a lightbulb
- Observations are independent

  → The outcome of one observation does not affect any other observation.

  → Two random sampling methods can deliver independence, sampling from an infinite population, without replacement.

  → Finite population, with replacement.

Binomial Probabilities

- For a binomial (n, π), a mathematical rule can find the probability for each of X: 0,1,2,…,n
- E.g.

$$P(X = n) = \pi^n$$
$$P(X = 0) = (1 - \pi)^n \quad \textit{(by independence)}$$

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

$$P(X = 2|n = 2, \pi = 0.5) = 0.5^2$$

- Event of interest: obtaining heads on a coin toss → n = 3 tosses. In how many ways can you get exactly X = 2 heads?
- Possible way: HHT, HTH, THH i.e. three combinations that each have the probability $0.5^3 = 0.125$.
- Thus, P (X = 2 | n = 3) is 3*0.125 = 0.375.

Rule of Combinations

- How many possible 3 scoop combinations could you create at an ice cream parlour if you have 31 flavours to select from? (No repeats! Order not important).
- The total choices are n = 31, and we select X = 3.

$$_{31}C_3 = \frac{31!}{3!(31-3)!} = \frac{31!}{3!28!} = \frac{31\times30\times29\times28!}{3\times2\times1\times28!} = 31 \times 5 \times 29 = 4{,}495$$

- Here, we are choosing positions → in how many ways can we choose 2 positions from 3 for the heads.

Binomial Distribution Formula

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

$P(X = n) = \pi^n$

$P(X = 0) = (1-\pi)^n$

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

$1 - \pi$ = (1 - 0.5) = 0.5

X = 0, 1, 2, 3, 4

- What is the probability of one success in five observations if the probability of an event of interest is 0.1? x = 1, n = 5, π = 0.1.

$$\begin{aligned} P(X = 1|5,0.1) &= \frac{n!}{x!\,(n-x)!}\pi^{x}(1-\pi)^{n-x} \\ &= \frac{5!}{1!\,(5-1)!}0.1^{1}(1-0.1)^{5-1} \\ &= 5 \times 0.1^{1}(1-0.1)^{5-1} \\ &= 0.32805 \end{aligned}$$

Binomial Distribution 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

$(1 - \pi)$ = probability of no event of interest for any trial

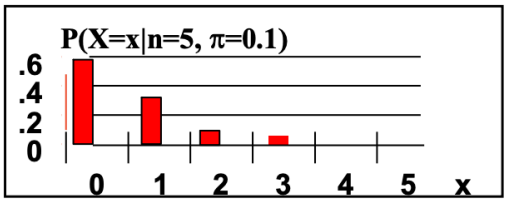
The Binomial Distribution Characteristics

– The shape of the binomial distribution depends on the values of π and n

- Here, n = 5 and π = .1

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

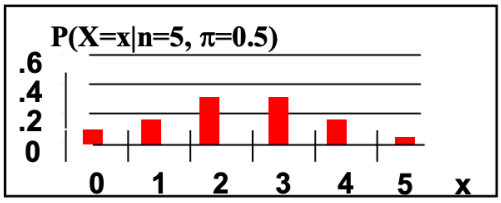
$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.1)(1-.1)} = 0.6708$$



- Here, n = 5 and π = .5

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

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



The Poisson Distribution

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

where:

x = number of events observed

λ = expected number of events

e = base of the natural logarithm system (2.71828...)

- Mean $E(X) = \lambda$

- Variance and Standard Deviation

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

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

where λ = expected number of events

Note: mean = variance!

Assumptions

- The data count is the number of times an event occurs in a given area/time/window of opportunity.
- The probability that an event occurs in one area of opportunity is the same for all areas of opportunity.
- Events occur independently of each other.
- The probability that two or more events occur in an area of opportunity approaches 0, as the area of opportunity becomes smaller and smaller.

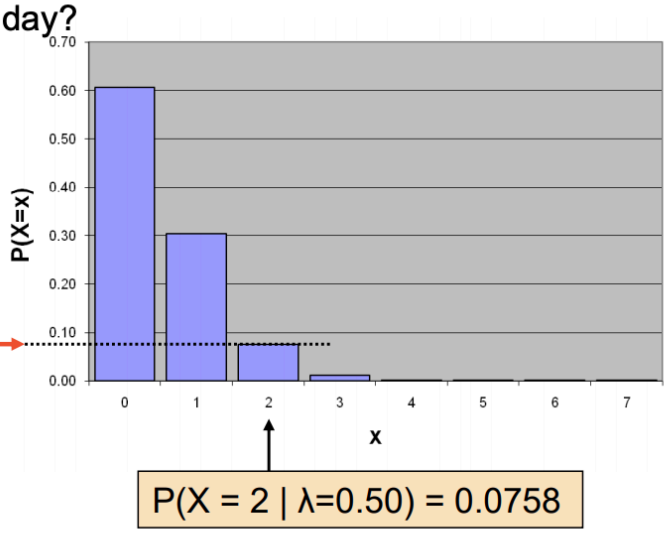
- The average number of events per area of opportunity is denoted by λ (lambda).

On average I receive **one letter every 2 days**. What is the probability I receive 2 letters in one day?

Graphically :

λ = 0.50

X	λ = 0.50
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
6	0.0000
7	0.0000



iversity of Sydney

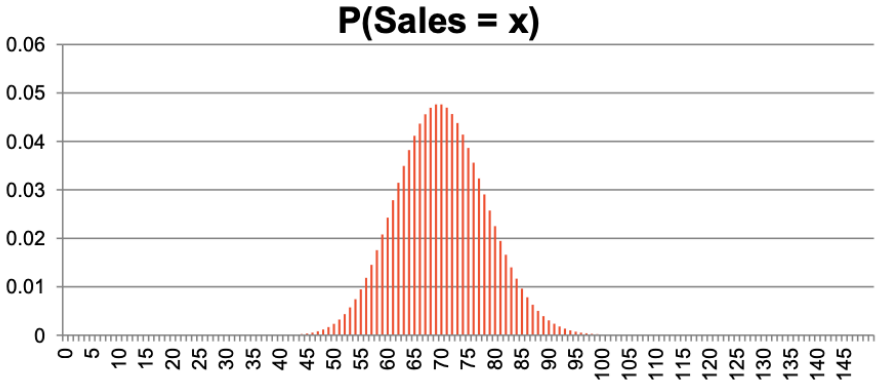
= POISSON.DIST(2,0.5,FALSE)

- The shape of the Poisson Distribution depends on the parameter λ

Example

- 2L Coca-Cola Bottle weekly sales at a Woolworths Store A follow a Poisson distribution with mean 79.

  Calculate:



1.Pr(Weekly sales = 75)=POISSON.DIST(75,70,FALSE)

$$= \frac{e^{-70}70^{75}}{75!} \quad = 0.038648 \quad \text{in Excel}$$

2.Pr(Weekly sales > 75) = 1– P(X≤ 75) $= 1 - \sum_{x=0}^{75} P(X = x)$

= 1-POISSON.DIST(75,70,TRUE)

= 0.251832 In Excel

The Hypergeometric Distribution

- 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 population.

$$P(X = x|n,N,A) = \frac{[{}_{A}C_{x}][{}_{N-A}C_{n-x}]}{{}_{N}C_{n}} = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$$

Where

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

- 'n' trials in a sample taken from a finite population of size N.
- Sample taken without replacement.
- Outcomes of trials are dependent.
- 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, drawing cards from a deck.

– 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

■ **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?

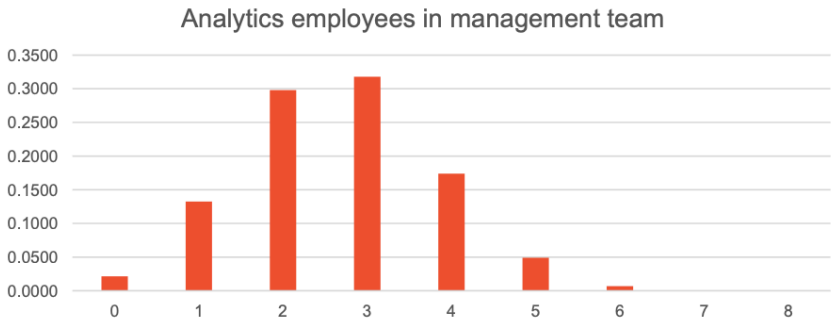
N = 10 A = 4 n = 3 x = 2

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

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

Example: 8 employees are randomly selected from 30 employed by a company to form a management team. 10 of the employees have high level analytics skills. What is the distribution for the number of high-level analytics employees to be selected in the management team?

$N = 30\ A = 10 \quad n = 8 \quad x = 0,1,\ldots,8$



Covariance and Summing Random Variables

Covariance

$$\sigma_{XY} = E([X - E(X)][Y - E(Y)])$$

- The covariance measures the strength of the linear relationship between two numerical random variables X and Y.
- A positive covariance indicates a positive relationship.
- A negative covariance indicates a negative relationship.

– The covariance formula for discrete rvs:

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

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

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

The Mean

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

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

- The Value fund X is averaging a $50 return while the Momentum fund Y is averaging a $95 return per $1000 invested.

The Standard Deviation

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

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

- Even though Momentum fund Y has a higher average return, it is subject to much more variab ility and the amount of loss/gain is higher.

Covariance

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

- Since the covariance is large and positive, 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:

$$\mathrm{E}(\mathrm{X}+\mathrm{Y}) = E(X) + E(Y)$$

- Variance of the sum of two random variables:

$$\mathrm{Var}(\mathrm{X}+\mathrm{Y}) = \sigma^2_{\mathrm{X+Y}} = \sigma^2_{\mathrm{X}} + \sigma^2_{\mathrm{Y}} + 2\sigma_{XY}$$

- Standard deviation of the sum of two random variables:

$$\sigma_{X+Y} = \sqrt{\sigma^2_{\mathrm{X+Y}}}$$

The Weighted Sum of Two Random Variables

- Expected Value of the weighted sum of two random variables:

$$\begin{aligned}\mathrm{E}(a\mathrm{X}+b\mathrm{Y}) &= E(aX) + E(bY)\\ &= aE(X) + bE(Y)\end{aligned}$$

- Variance of the weighted sum of two random variables:

$$\begin{aligned}\mathrm{Var}(a\mathrm{X}+b\mathrm{Y}) &= \sigma^2_{a\mathrm{X}+b\mathrm{Y}} = \sigma^2_{a\mathrm{X}} + \sigma^2_{b\mathrm{Y}} + 2\sigma_{aXbY}\\ &= a^2\sigma^2_{\mathrm{X}} + b^2\sigma^2_{\mathrm{Y}} + 2ab\sigma_{XY}\end{aligned}$$

Portfolio Returns: wX + (1-w)Y

- Expected Value of the weighted sum of two asset returns X, Y:

$$\mathrm{E}(w\mathrm{X}+(1-w)\mathrm{Y}) = wE(X) + (1-w)E(Y)$$

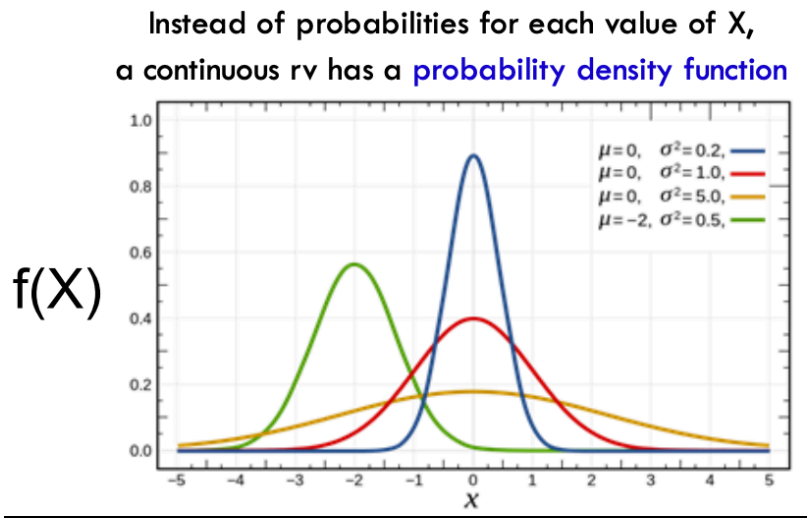
- Variance of the weighted sum of two asset returns X, Y:

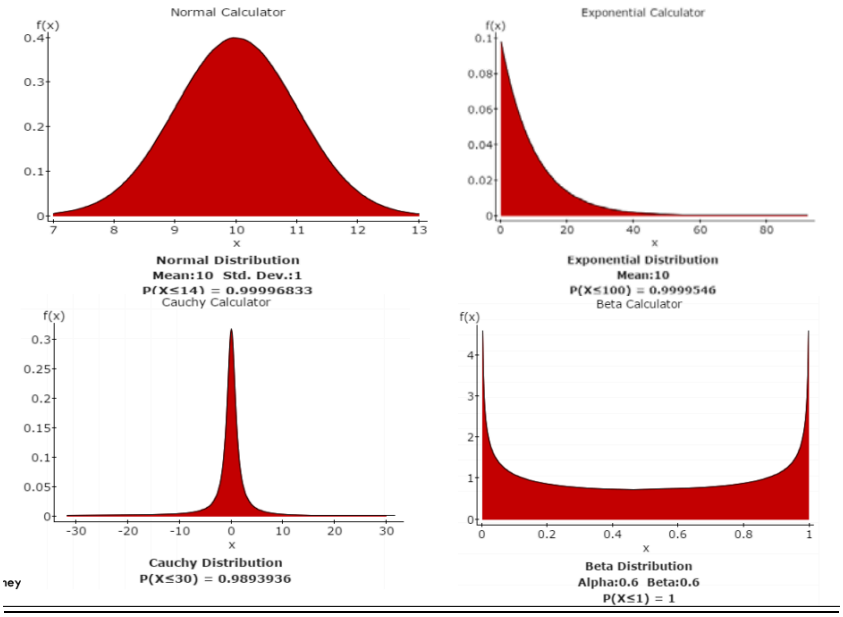
$$\mathrm{Var}(w\mathrm{X}+(1-w)\mathrm{Y}) = w^2\sigma^2_{\mathrm{X}} + (1-w)^2\sigma^2_{\mathrm{Y}} + 2w(1-w)\sigma_{XY}$$

BUSS1020 Week 6 Lecture – Continuous Distributions

- A continuous random variable can potentially take on any value in a range, depending on how accurately and precisely it can be measured.
- Some cts rvs include:
    - → thickness, height, weight, volume of cans, packets, boxes etc.
    - → delivery times, project completion times, times between event.
    - → download times, web query times
    - → coffee temperature
    - → hotel occupancy rate
    - → financial return, volume of trades, time between trades
    - → revenue, cost, accounting variables (EPS, ROA, book value etc.)
    - → Churn rate, 'reach', bouncing rate, default rate

Continuous Probability Density Function

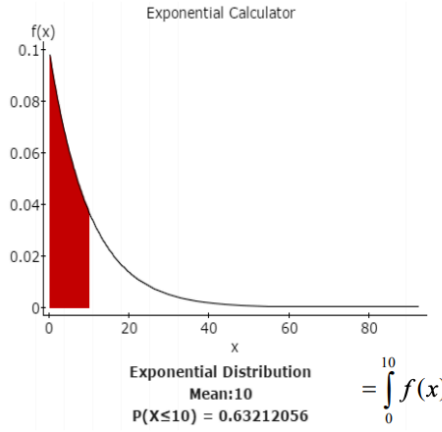




Probability Density Function Interpretation

- Each density curve f(x) represents the relative likelihood of each X value.
- The area of each shaded region is the probability that X is in that region.
- Probabilities for continuous rvs are only considered for regions e.g.

$$P(a < X < b) = \int_a^b f(x)dx$$



$$= \int_0^{10} f(x)dx$$

ersity of Sydney

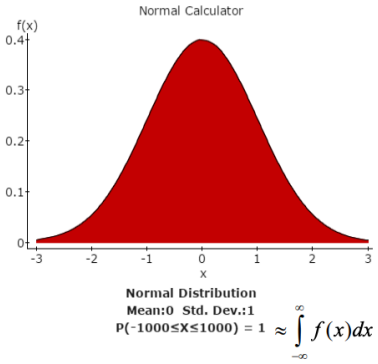
- The area under the entire density curve is always exactly 1.

  i.e. P(a<X<b) = 1 if (a, b) covers all possible values of X

  - The area under a single value for X equals 0, i.e.

$$P(X = a) = P(a \le X \le a)$$

$$= \int_a^a f(x)dx = 0$$

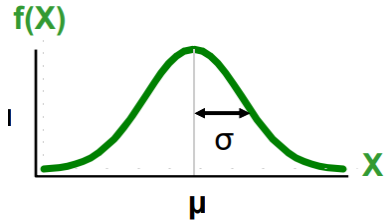


$$\approx \int_{-\infty}^{\infty} f(x)dx$$

The University of Sydney

Normal Distribution (Gaussian)

- Bell shaped density curve → symmetric (skewness = 0), mean = median = mode.
- Location given by the mean parameter, spread given by the standard deviation parameter and the random variable (rv) has an infinite theoretical range.



- Most real data are NOT normally distributed.
- Some exceptions:

  → IQ (constructed to be normal, but very slight right skew).

  → Positions of particles in fluid (Brownian motion).

  → Sums of many rvs (central limit theorem).

- Mostly assumed as an approximation; sometimes disastrously e.g. Black-Scholes option pricing, GFC, CDo pricing.

Normal Density Function

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

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

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

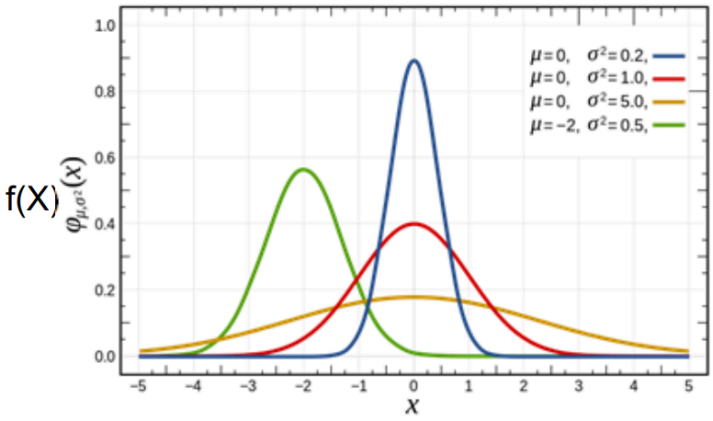
$\pi =$ mathematical constant ≈ 3.14159

$\mu =$ the population mean $E(X)$

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

$X =$ a value of the continuous variable

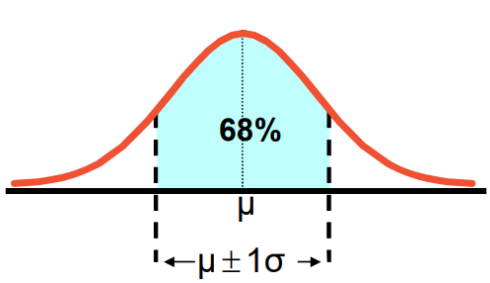
Changing **μ** shifts the distribution left or right.

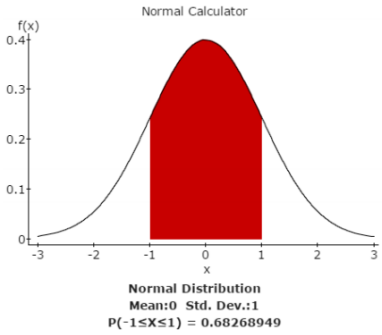


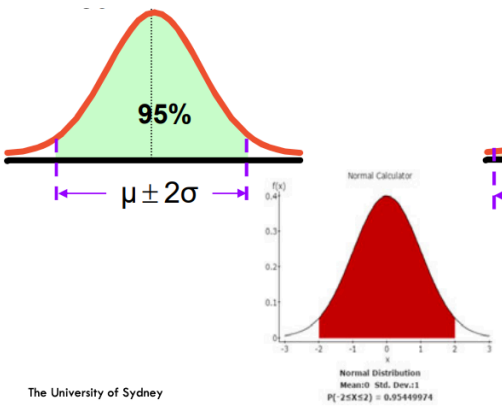
Changing **σ** increases or decreases the spread.

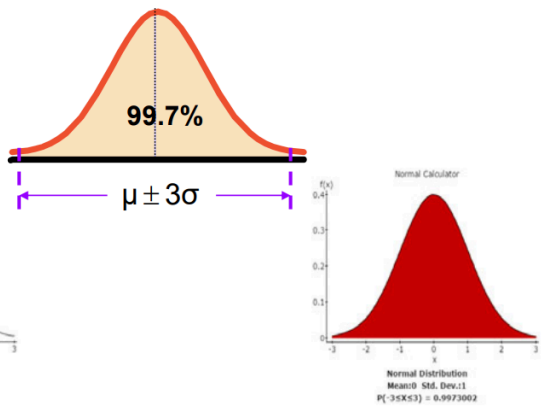
"Normal" Rules

- The normal distribution is exactly bell-shaped → follows the empirical rule.









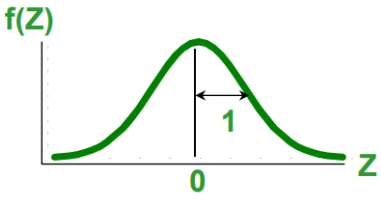
The Standardised Normal

- 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 $\sigma = 1$
  - Also known as the "Z" distribution

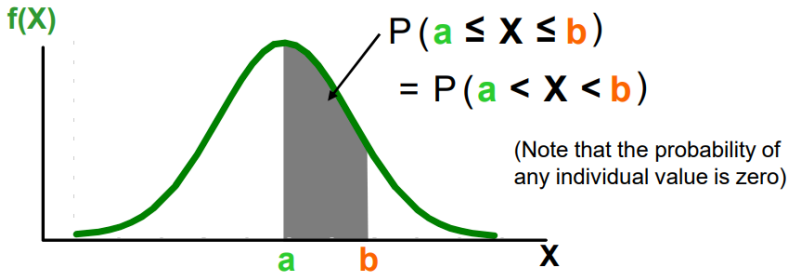


X values above the mean have **positive** Z-values,

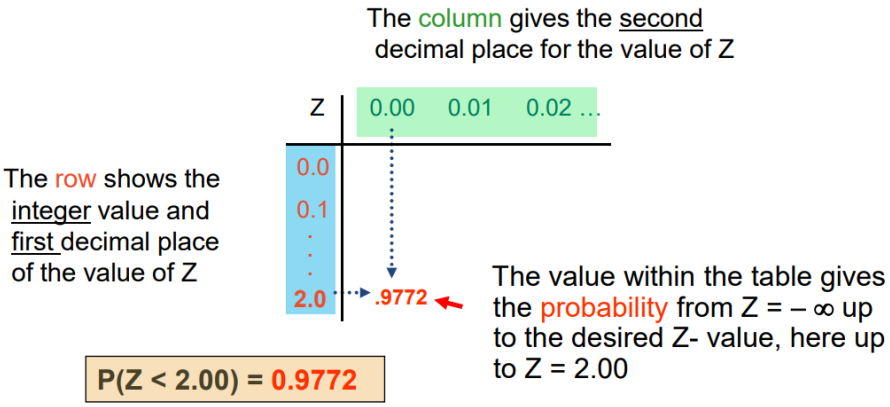
X values below the mean have **negative** Z-values

Finding Normal Probabilities

- Probability is measured by the area under the curve.



- Half the area is above the mean, half the area is below the mean, all summing to 1.
- The Cumulative Standardised Normal table in the textbook gives the probability less than a desired value of Z (i.e. from negative infinity to Z).

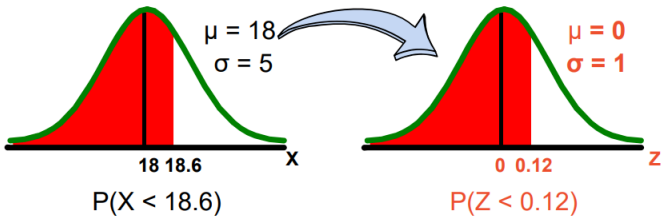


- General procedure is to translate X values to Z-values and then use the standardised normal table.

Example 1a

- Let X represent the time (in seconds) that visitors spend on Telstra's website.
- Suppose X is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds.
- Find P(X < 18.6)

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

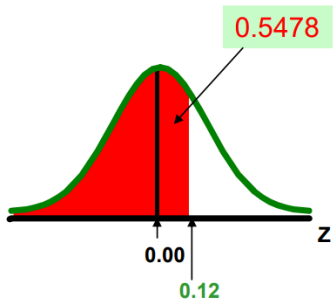


P(X < 18.6) P(Z < 0.12)

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



Given a Normal Probability, Find the X Value

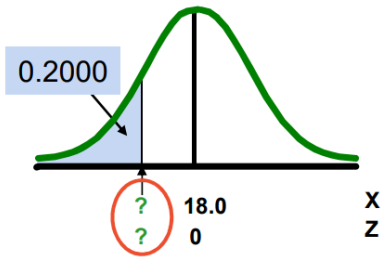
Steps to find the X value for a known probability:

1. Find the Z-value for the known probability
2. Convert to X units using the formula:

$$X = \mu + Z\sigma$$

Example:

- Let X represent the time (in seconds) that visitors spend on Telstra's website.
- Suppose X is normal with mean 18.0 and standard deviation 5.0
- Find X such that 20% of website visit times are less than X.



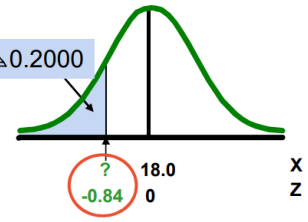
Find the Z-value for 20% in the Lower Tail

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



University of Sydney

Page 34

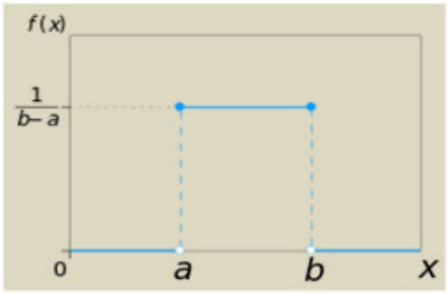
- Using $X = \mu + Z\sigma$, approximately 20% of the visits will be less than 13.80 seconds.

Assessing Normality (Roughly!) or Informally

- Is the sample mean ≈ sample median ?
- Is the empirical rule approximately satisfied? (using $\bar{X}, s$)
- 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 "fat tails"?
- Are the sample skewness and kurtosis statistics ≅ 0?

Uniform Distribution

- The continuous uniform distribution has equal density for all possible outcomes of the random variable.
- Also called a rectangular distribution.



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

where

f(X) = value of the density function at any X value

a = minimum possible value of X

b = maximum possible value of X

– The mean of a uniform distribution is

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

– The standard deviation is

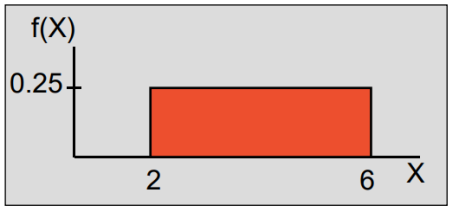
$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Uniform Distribution 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 ~ Uniform, over the range $2 \le X \le 6$:

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



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

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

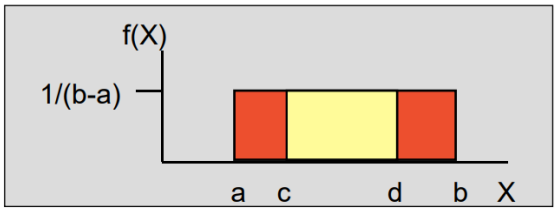
Uniform Distribution Example *(continued)*

Example: Use the Uniform (2,6) distribution to find the probability that the bus is between 3 and 5 minutes late, i.e. $P(3 \le X \le 5)$:

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

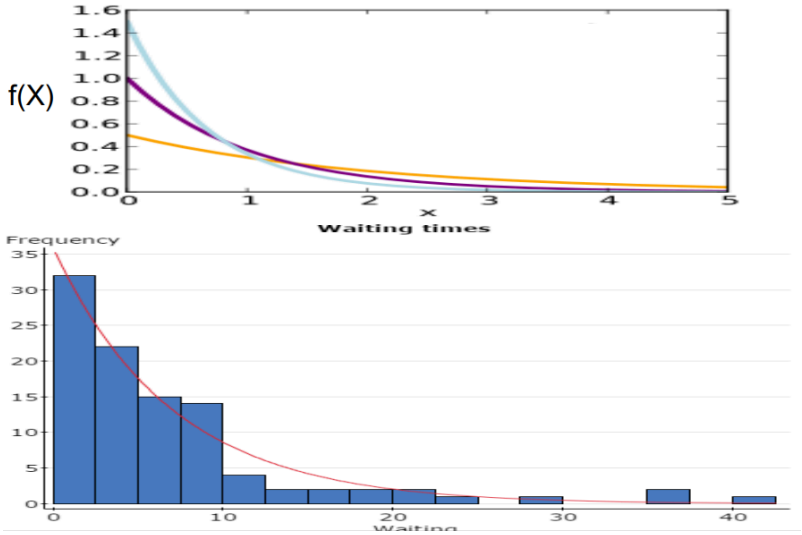
In general, if X ~ Uniform (a, b), then to find $P(c \le X \le d)$:

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



The Exponential Distribution

- A positive valued, right-skewed distribution.



- Often used to model the length of time between two occurrences of an event (i.e. time between events).
- Examples
  - → Time between people arriving to be seated in a restaurant.
  - → Time between trucks arriving at an unloading dock.
  - → Time between transactions at an ATM Machine.
  - → Time between phone calls to the main operator.
  - → Time between financial trades on an asset.
- 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/ λ

  - Has only a single parameter λ (lambda)
  - The probability density function is

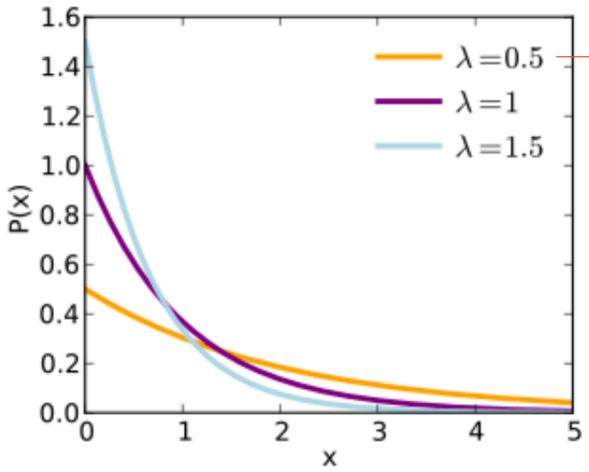
$$f(x) = \lambda e^{-\lambda x} \quad ; \quad x > 0$$

where e = constant, ≈ 2.71828
1/λ = the population mean
x = any value of the continuous variable, 0 < x <∞

- Always right or positively skewed, mode < median < mean (always).

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

c.f. with Poisson where
mean = variance



μ = 2 = σ

μ = 1 = σ

μ = 0.667 = σ

$$\mu = \frac{1}{\lambda} = \sigma$$

– The probability that an exponential rv is less than some specified X=x is

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

where e = constant, ≈ 2.71828

1/λ = the population mean = population standard deviation

x = any value of the continuous variable where $(0 < x < \infty)$

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

Exponential Distribution Example

Example: 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? What is the average time between customers?

- μ = 1/15 hours ; λ = 15 per hour; x = 0.05 hours
- P(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
- The average time between customers is 4 minutes

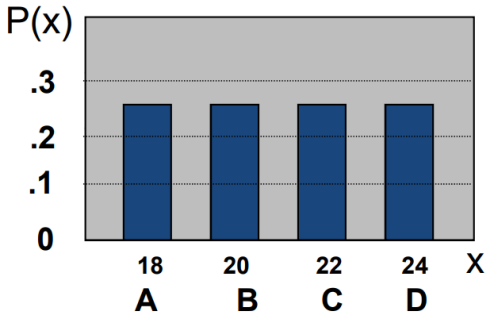
BUSS1020 Week 7 Lecture – Sampling Distributions

Sampling Distributions

- A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population → we are interested in the sampling distribution of the sample proportion/mean.
- E.g. 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.
- Assuming a population of size N=4, Random variable, X, being the number of business meetings this month, and values of X being 18, 20, 22, 24:

$$\mu = \frac{\sum X_i}{N} = \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Uniform Distribution (discrete)

ersity of Sydney

Now consider all possible samples of size n=2:
what is the distribution of possible sample means?

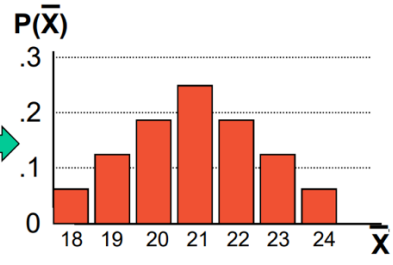
1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

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

Sample Means Distribution



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

Sampling Distribution of the Mean

– 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

Standard Error of the Mean

- Different samples of the same sample size (n) from the same population will yield different sample means.
- A measure of the variability in the mean from sample to sample is given by the Sample Error of the Mean (sampling is either with replacement from a finite population or without replacement from an infinite population).

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

st.error decreases as the sample size increases

- Z- Value for Sampling Distribution of the Mean:

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

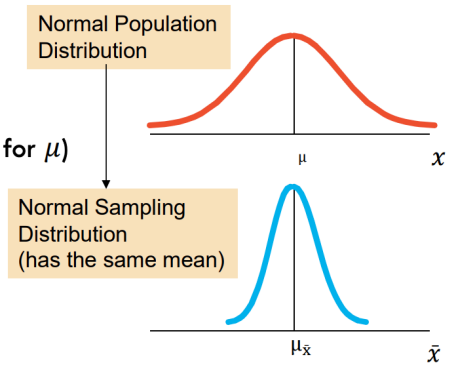
where:
$\bar{X}$= sample mean
μ= population mean
σ= population standard deviation
n = sample size

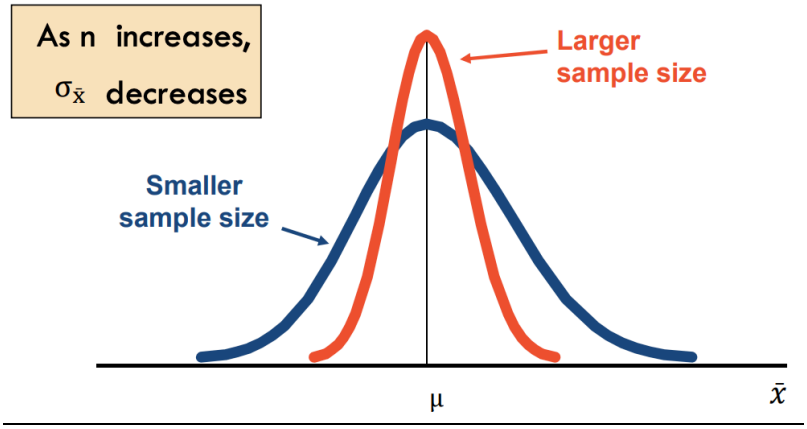
- If a population is normally distributed, the sampling distribution is also normally distributed with the expected value of X equalling the mean, and the standard deviation equal to the standard error.

Sampling Distribution of a Normal Population

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

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





Determining an Interval Including a Fixed Percentage of the Sample Means

Find a symmetrically distributed interval around μ that will include 95% of the sample means when μ = 368, σ = 15, and n = 25 and the 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 standard 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

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

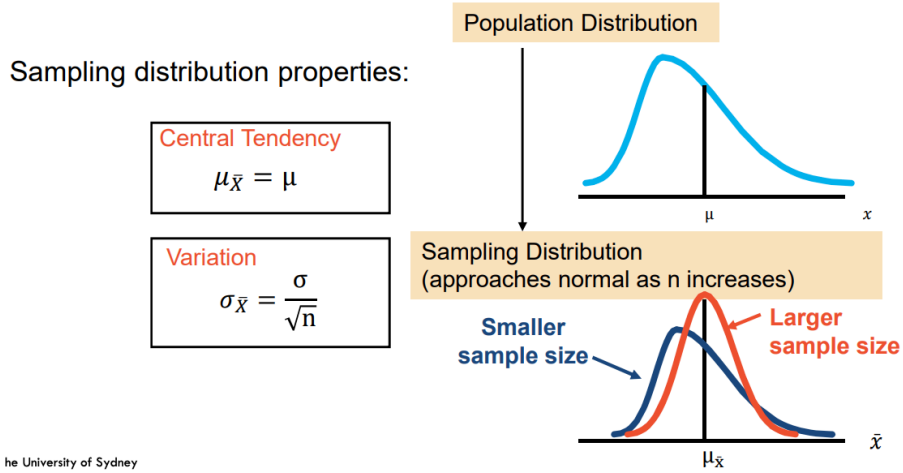
- Calculating the upper limit

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

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

Sampling Distribution of the Mean for Non-Normal Populations

- Central Limit Theorem → if the population is not normally distributed, sample means of random samples from the population will be approximately normally distributed, as long as the sample size n is large enough.



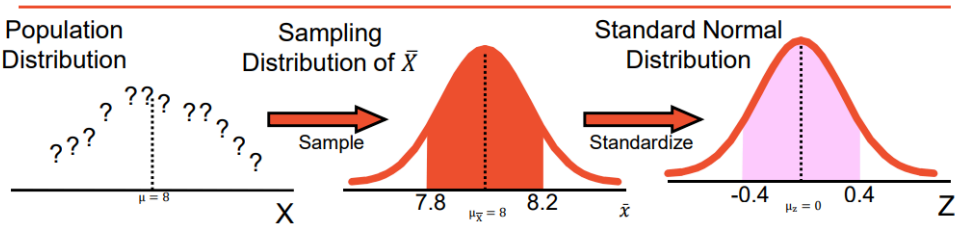
How Large is Large Enough?

- For most population distributions, n >= 30 will give a sampling distribution for the sample mean that is approximately normality.
- For fairly symmetric distributions, n >= 15 is enough for approximate normality.
- For normal population distributions, the sampling distribution of the mean is always exactly normally distributed.

Example

- Supposing a mean of 8, standard deviation of 3 and random sample size of 36, the probability that the sample mean is between 7.8 and 8.2 can be found as such:
- Even if the population is not normally distributed, the central limit theorem can be used (n >= 30).

$$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8-8}{\frac{3}{\sqrt{36}}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.2-8}{\frac{3}{\sqrt{36}}}\right)$$
$$= P(-0.4 < Z < 0.4) = 0.6554 - 0.3446 = 0.3108$$



Population Proportions

π = the population proportion

Sample proportion (p) estimates π:

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

$0 \le p \le 1$

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

σ_p= "The standard error of the proportion"

- The sampling distribution of p follows a binomial distribution. By the central limit theorem, p is approximately distributed as a normal distribution when n is large enough.

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

Z-Values for Proportions

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

Example

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

Find :

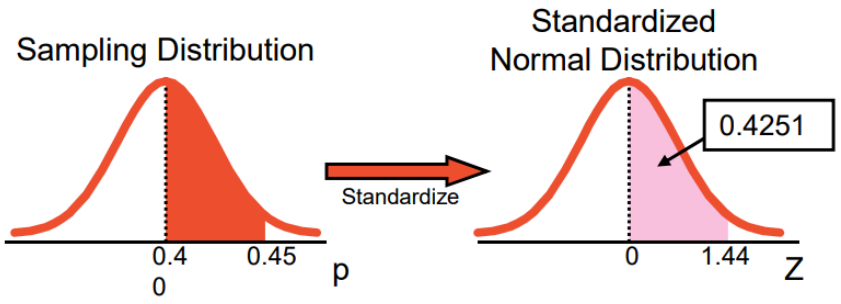
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

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:

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



Sampling Distributions from Finite Populations

- Used to adjust the standard error of both the sample mean and the sample proportion.
- Needed when the sample size, n, is more than 5% of the population size, AND sampling is without replacement.
- The Finite Population Correction (FPC) Factor:

$$\text{fpc} = \sqrt{\frac{N-n}{N-1}}$$

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, resulting in more precise estimates of population parameters.

A random sample of size 100 is drawn without replacement, from a population of size 1000. The sample mean is 9 and standard deviation is 40.

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

So using the fpc we get:

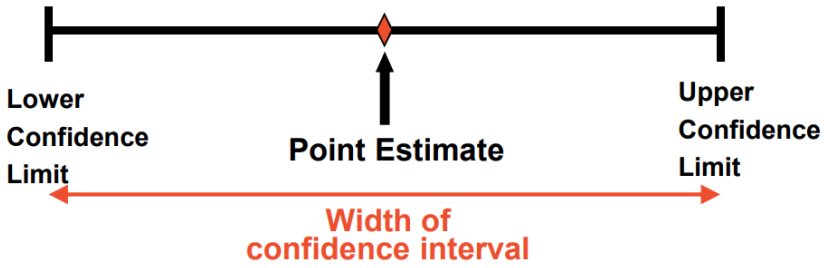
$$\sigma_{\bar{X}} = \frac{40}{\sqrt{100}}\sqrt{\frac{1000-100}{1000-1}} = 3.8$$

E.g. A 95% CI for the true mean would then be

(9-1.96*3.8, 9+1.96*3.8) = (1.552, 16.448)

Confidence Intervals

- A point estimate is a single number, whilst a confidence interval provides additional information using the variability of the estimate.



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

- A confidence interval gives a range of values:
  - → takes into consideration variation in sample statistics e.g. from sample to sample.
  - → based on observations from only one sample
  - → gives information about possible values of unknown population parameters
  - → stated in terms of "level of confidence" e.g. 95%, 99% confident.

- Population has μ = 368 and σ = 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 $n = 25$.
  - When you don't know μ, you use 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)
    - Since 356.42 ≤ μ ≤ 368.18 this inequality, based on this sample, makes a correct statement about μ.

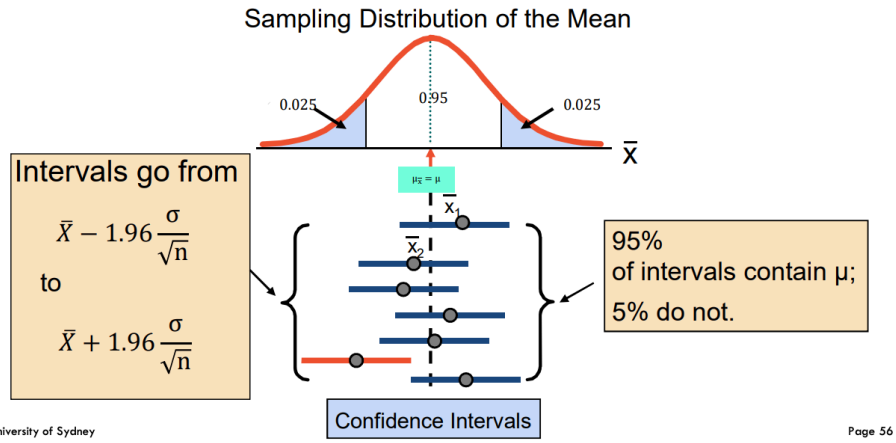
- Intervals from other samples of size 25:

95% intervals

Sample #	$\overline{X}$	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, you only take one sample of size n, you don't know the mean, so you don't know if the calculated interval actually contains the mean.
- However, you do know that 95% of samples will give an interval that contains the mean → form the one sample you actually collect, you can be 95% confident that the interval contains the mean.

95% Confidence Intervals



- Confidence levels are our confidence that the interval will contain the unknown population parameter as a percentage less than 100%.
  - Suppose the confidence level = 95%
    Also written as (1 - α) = 0.95, (so α = 0.05)
  - A relative frequency interpretation:
    - 95% of all the confidence intervals that can be constructed will contain the unknown true parameter

Without knowing μ, there is a 0.95 chance that μ is inside a particular interval, under the frequency definition of chance.

General Formula

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

---

Confidence Interval for Mean (St.Dev. Known)

- Assumptions
  - Population standard deviation σ is known → Realistic? Possible?
  - Population is normally distributed OR
  - If population is not normal, using a "large" sample (CLT)
- Confidence interval estimate: $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

where $\bar{X}$ is the point estimate

$Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail

$\frac{\sigma}{\sqrt{n}}$ is the standard error

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

BUSS1020 Week 8 Lecture – Confidence Intervals

- In the majority of real world business situations, the population standard deviation is not known (barring a very good approximation) → in a situation where the standard deviation is known, the mean also may be known (since it is needed in calculating standard deviation).
- If you truly knew the mean, there would be no need to gather further samples to estimate it.

Confidence Intervals for Mean (St.Dev. Unknown)

- If the population standard deviation 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-t Distribution

- This family of distributions was developed by William Gossett while working for Guinness.

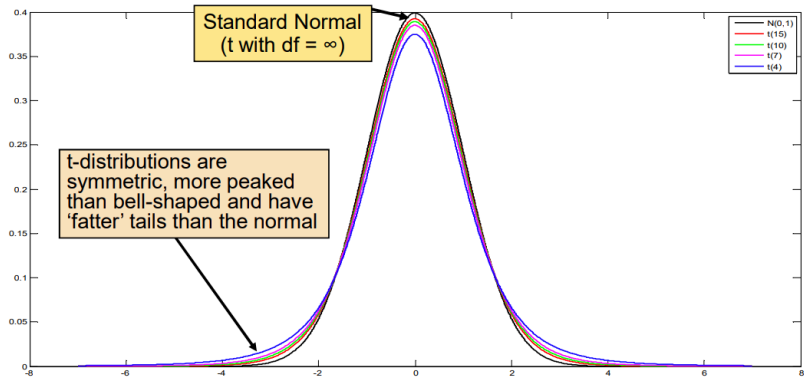
➢ If $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ has a normal distribution with mean 0 and variance 1 (a standard normal),

➢ 'Student' showed that $t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ had a different distribution, later called the "Student-t" distribution.

- It represents the sampling distribution for the standardised sample mean, when we have estimated the standard deviation by S.
- The $t_{\alpha/2}$ values depend on the number of degrees of freedom (d.f.) → number of observations that are free to vary after sample mean has been calculated.

$$d.f. = n - 1$$

Note: t → Z as n increases (Why?)



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

- Assumptions
  - Population standard deviation is unknown
  - Population is
    - normally distributed OR
    - not normally distributed, but a large sample size (CLT works)
- Use Student's t Distribution
- Confidence Interval Estimate: $\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$

(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)

- Degrees of variation are the number of observations are free to vary after the sample mean has been calculated.

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is X_3?

If the mean of these three values is 8.0, then X_3 must be 9 (i.e., X_3 is not free to vary)

Here, n = 3, so degrees of freedom = n – 1 = 3 – 1 = 2

(2 values can be any numbers, but the third is not free to vary for a given mean, when n=3)

- If a random sample of n-25 has a mean of 50 and S=8, the degrees of freedom are n-1=24, so $t=t_{0.0025}$ = 2.0639 → confidence interval can be calculated from there.

Excel

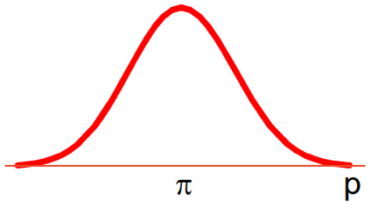
	A	B	C
1	Mean	50	
2	S	8	
3	n	25	
4	T Score	2.0639	= T.INV.2T(0.05,24)
5	SE	1.6	=B2/SQRT(B3)
6	Lower	46.6978	=B1-B4*B5
7	Upper	53.3022	=B1+B4*B5

The Excel function T.INV.2T calculates the two-tailed inverse of the Student-t distribution. For example, to find a 95% interval, i.e. when α=0.05, we need the $\alpha/2$=0.025 percentage point of the Student-t. T.INV.2T gives the $\alpha/2$ percentage point when α is input.

T.INV gives the left, one- tailed percentage point, i.e. T.INV gives the $\alpha/2$ percentage point when $\alpha/2$ is input

Confidence Intervals for Population Proportion

- An interval estimate for the population proportion can be calculated via the estimate of the sampling distribution of the sample proportion.



Confidence Interval Endpoints

- Upper and lower confidence limits for π are calculated via

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

- where
  - $Z_{\alpha/2}$ is the standard normal value at confidence level 1- α
  - p is the sample proportion
  - n is the sample size
  - π is the population proportion

- Note: must assume $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Since we don't know π, we estimate these with p, i.e. $np \geq 5$ and $n(1-p) \geq 5$

- 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 $Z_{\alpha/2} = 1.96$)

$np = 100 * 0.25 = 25 \geq 5$ & $n(1-p) = 100 * 0.75 = 75 \geq 5$ ✓

Make sure the mean of X is away from 0 and from n

$$p \pm Z_{\alpha/2}\sqrt{p(1-p)/n}$$
$$= 25/100 \pm 1.96\sqrt{0.25(0.75)/100}$$
$$= 0.25 \pm 1.96\,(0.0433)$$
$$0.1651 \leq \pi \leq 0.3349$$

- From this, 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.

Determining Sample Size

Sampling Error

- The required sample size can be found that obtains a desired margin of error €, with a specified level of confidence (1 - α).
- The margin of error, e, is also called the sampling error.
  - → The amount of imprecision is 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 Mean Where σ Is Known

$$n = \frac{Z_{\alpha/2}{}^2\,\sigma^2}{e^2}$$

- The desired level of confidence determines the critical value, with the acceptable sampling error and standard deviation also needed.

If $\sigma = 45$, what sample size is needed to estimate the population mean to within $\pm$ 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)

For the Mean Where σ Is Unknown

- 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:

$$e = t_{\alpha/2\,;\,n-1}\frac{S}{\sqrt{n}}$$

For the Proportion

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

- The desired level of confidence (1-σ), which determines the critical value $Z_{\sigma/2}$
- The acceptable sampling error, e.
- The true proportion of events of interest, π.

→ π can be estimated (by p) with a pilot sample.

→ or conservatively use π = 0.5

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

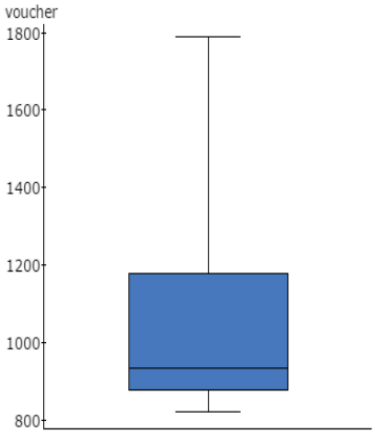
$$n = \frac{Z_{\alpha/2}^2 \, \pi(1-\pi)}{e^2} = \frac{(1.96)^2(0.12)(1-0.12)}{(0.03)^2} = 450.74$$

So use n = 451

Application to Auditing

- Many auditors make extensive use of probability sampling and confidence intervals.
- The population of company "accounts" is often too large or time consuming or too costly to keep up with.
- Dealing with a probability sample of accounts is more feasible.

  - 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



$$\overline{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 1076.39 \pm (2.0096)\frac{273.62}{\sqrt{50}}\sqrt{\frac{950}{999}}$$

$$= (\$1000.56, \$1152.22)$$

  - Thus we are 95% confident that the population average voucher amount is between $1000.56 and $1152.22.
  - Thus, we are also 95% confident that the population TOTAL value of vouchers is between 1000*$1000.56 and 1000*$1152.22. Which is ($1000559, $1152221)

Excel calculations

	A	B	C	D	E
10	$\bar{X}$	1076.39	**N**	1000	
11	**S**	273.62	**FPC**	0.975167	**SQRT(D10-B12) / SQRT(D10-1)**
12	**n**	50			
13	**SE**	38.69571	**B11/SQRT(B12))*D11**		37.73479
14	**t**	2.009575	**T.INV.2T(0.05,B12-1)**		

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.

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, claims or predictions that can tested.
- The most common way to test such hypothesis is via empirical methods i.e. data.

What is a Hypothesis?

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

Example: Telstra's market share proportion in mobile phone customers, π, is greater than 0.5

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

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.

$$H_0: \mu = 30$$

- It is always about a population parameter, not about a sample statistic.

- 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.

Alternative Hypothesis, H_1

- Opposes the null hypothesis in some way e.g. The average diameter of a manufactured bolt is NOT equal to 30mm; the average diameter is less than 30mm.

$$(\ H_1: \mu \neq 30\ )$$

- Challenges the "status quo".
- Is generally the hypothesis that the researcher is trying to find evidence for (or against) and is often formed first → the 'interesting' one.

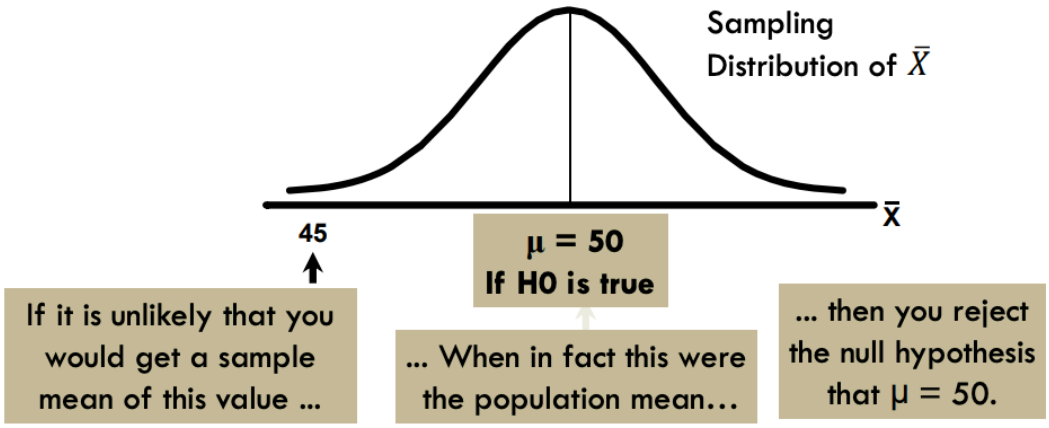
Hypothesis Testing Process

- Usually, we start with the alternative (interesting) hypothesis: Amazon wants a mean of at least $50 for annual customer purchase amount. Is the mean $50.
- Claim: the population mean amount if 50.

$$H_0: \mu = 50$$
$$H_1: \mu \neq 50$$

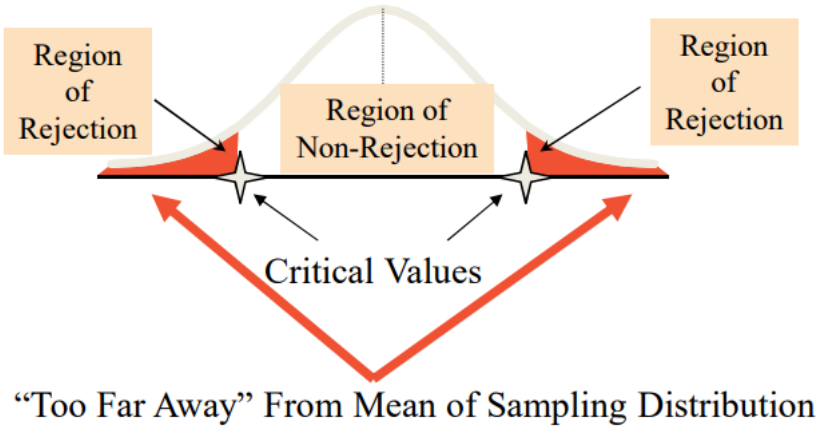
- To test this, we sample the population and find the sample mean.

- Suppose the sample mean amount was $\bar{X}$: 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 can conclude that population mean is very likely not 50.



The Test Statistic and Critical Values

- If the sample mean is 'close' to the stated population mean, the null hypothesis is not rejected.
- If the sample mean is 'far' 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 (answers the question of how far is far enough).



Possible Errors in Hypothesis Test Decision Making

Type I Error

- *Rejecting a true null hypothesis*
- 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 false null hypothesis*
- Probability of Type II Error is **β** = P(not reject null | null false)
- **Determined by the test, sample size, etc. and usually not able to be known precisely.**

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	H_0 True	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, β are conditional probabilities!

- The confidence coefficient (1-α) is the probability of not rejecting H_0 when it is true.
- The confidence level of a hypothesis test is (1-α)*100%.
- The power of a statistical test (1-β) is the probability of rejecting H_0 when it is false.

$P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$: the **size** of a test
$P(\text{reject } H_0 \mid H_0 \text{ false}) = 1 - \beta$: the **power** of a test

Type I and II Error Relationship

- Type I and Type II errors cannot happen at the same time.
- A Type I error can only occur given H_0 is true.
- A Type II error can only occur given H_0 is false.

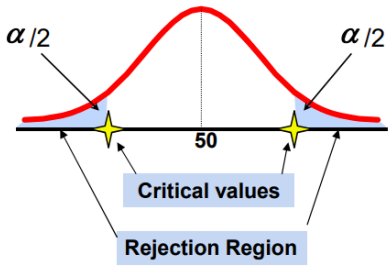
When Type I error probability (α) ⇧
Type II error probability (β) ⇩

Factors Affecting Type II Error

- All else being equal,
  - β ⇧ when the difference between hypothesized parameter and its true value ⇩
  - β ⇧ when α ⇩
  - β ⇧ when σ ⇧
  - β ⇧ when *n* ⇩

H_0: μ = 50
H_1: μ ≠ 50

Level of significance = *α*



This is a two-tailed test because there is a rejection region in both tails.

BUSS1020 Week 9 Lecture – One Sample Hypothesis Testing

Hypothesis Testing for the Mean