StuDocu.com

MIS171 - Summary Business Analytics

Business Analytics (Deakin University)

StuDocu is not sponsored or endorsed by any college or university
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Word	Definition	Notes
Week 1 (Topic 1) - Chapter ONE (Role of Business Analytics in real world context)			
Business Analytics	Business Analytics	Process of **transforming data into actions** through analysis and insights in the context of organisational decision making and problem solving Using data, IT, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers **gain an improved insight** about their business operations and make better fact-based decisions • Report using historical info • Gives info to enable future predictions • End goal: add value through insight + turn data into info • Makes distinction b/w relevant + irrelevant knowledge	
	Purpose of BA	• ID-ing valuable data re business' strategy + objectives • Internal value = revenue growth	
	Importance of BA	• aids decision making • decisions made using analysis are better than those made through gut instinct • use of analytics = profitability and revenue	
	Adds value by its:	• Business relevancy • Actionable insight • Performance + value measurement	
	BA helps understand:	• What will happen • Why happened • Best course of action	
	BA applications	• mgmt. of customer relos • financial + marketing activities • supply chain mgmt. • HR planning • Pricing deicisons	
Scope of BA	Descriptive analysis	**want to know about the past** • Most commonly used + most well understood type of analytics • Use data **to understand past and present performance** to make important decisions • Summarizes data into meaningful charts and reports Focuses on: • descriptive measures • data visualisation • probably distributions / sample + estimation • statistical inference	
Scope of	Predictive	**want to know abc**	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

BA	analysis	• Analyses past performance in an effort to **predict the future** by examining historical data, detecting patterns or relationships in data Techniques include: • regression • forecasting	
Scope of BA	Prescriptive analysis	**making decisions / optimisation** • Uses optimization to **identify the best alternative** to minimize or maximise some objective • Addresses questions such as: How much should we produce to maximize profit? What is the best way of shipping goods from our factory to minimize costs? Techniques include: • optimisation • simulation	
	BA V analytics	**Business Analytics** • BA = making info have contextual relevancy + delivering real value • Insight that is actionable to create value **Analytics** • A = finding interesting things in lrg amounts of data • Focuses on creation of insight • Simply answers question without necessary return 80% accuracy <u>with</u> actionable insight **better than** 98% accuracy <u>without</u> actionable insight	
	Examples of analytics	• **Reporting**: summary of historical data • **Trending**: ID-ing patters • **Segmentation**: ID-ing similarities in data • **Predictive modelling**: using historical data to predict future events	
	Mgmt.	Process which org **achieves goals through use of resources** (ppl, money, materials + info = input, achieving goals = output)	
	Role of manager	• **Interpersonal role**: figurehead, leader, liaison • **Informational role**: monitor, disseminator, analyser, spokesperson • **Decisional role**: entrepreneur, disturbance handler, resource allocator, negotiator	
	Decision	**Choice b/w alternatives** made by individual/group	
	Nature of decisions	• **operational control** = executing specific tasks efficiently + effectively • **mgmt. control** = acquiring + using resources efficiently • **strategic planning** =long term goals + policies for growth + resource allocation	
	Difficulties in	• Number of alternatives	

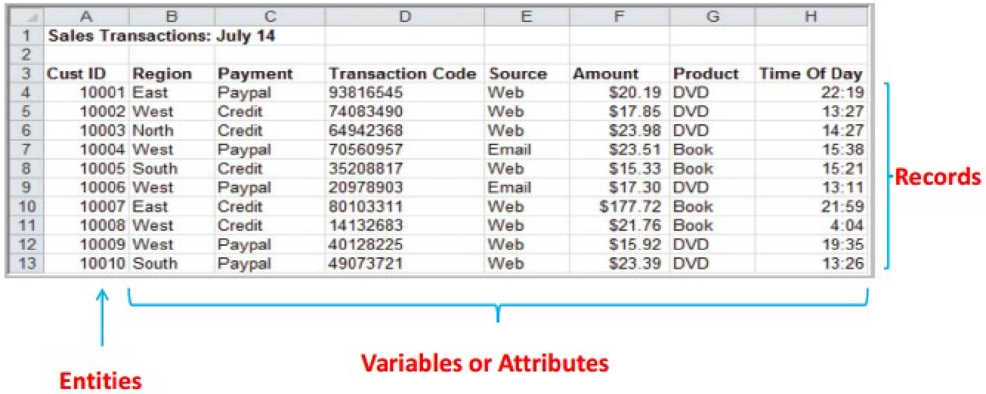
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	decision making	• Time pressures • Conduct analysis • May be necessary to access info/consult with expert = therefore computerised analysis streamlines decision making	
	Process of decision process	1. Intelligence: What is the problem? 2. Design: what are my options? 3. Choice: pick an option + decide how to implement (if choice doesn't work/achieve goal = start process again)	
Type of Decision	Structured decisions	**Std problems** • routine + repetitive problems w std solutions eg. Order entry, accounts receivable	
Type of Decision	Unstructured decisions	**Complex problems** • no cut-dry solution eg. Planning new service offering, hiring executive	
Type of Decision	Semi-structured decisions	**Combination of std solution procedures + individual judgement** eg. Evaluating employees, setting marketing budgets	
Business Intelligence	Business Intelligence	**decision support applications + technologies + processes**	
	BI applications	provide users with **view of what has happened**	
	Data mining	process of **searching for valuable info in lrg database, warehousing or mart** • helps explain why it is happening + predicts what will happen in future • predicts tends + behaviours • identifies prev. unknown patterns • able to provide predictive info eg. For targeted marketing, forecasting bankruptcy/defaults eg. can use prev. promotional info to identify ppl most likely to response to similar offers	
	Drill down	ability to go into more detail	
	Models	simplified representations or abstracts of reality	
Decision support system	Decision support system (DSS)	• **combines models + data to analyse semi-structured + unstructured problems** • enables ppl access to data to manipulate data + conduct appropriate analyses • enhances learning + contribute to all levels of decision making	
Type of DSS	Sensitivity analysis	study of the **impact one or more parts of decision has on other parts** • examines impact of input variable changes on output variables • enables system to adapt to changing conditions + varying requirements of diff situations • provide better understanding of model + problem model describes	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

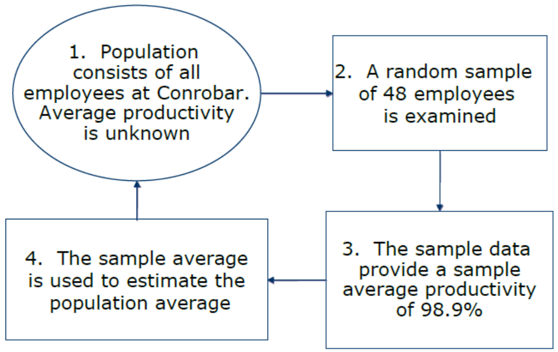
Type of DSS	What-if analysis	attempts to **predict impact of changes** based on assumptions (input data) on proposed solution	
Type of DSS	Goal-seeking analysis	• find the **value of inputs necessary to achieve desired level of output** • finding out **what is needed to achieve certain goal**	
	Dashboards	• evolved from executive info systems = designed for info needs of execs • provides access to timely info + direct access to mgmt. reports • enable managers to examine reports + drill down into detailed info	
	Data visualisation technologies	presenting **data in the form of graphs or tables** to enable users to more easily process and understand data	
	Geographic Info Systems (GIS)	system for capturing, integrating, manipulating + displaying data **using digitised maps**	
	Geo-coding	**identifying geographical location** of every digital record • enables users to generate info for planning, problem solving + decision making • graphical format makes it easy for managers to visualise data	
	Reality mining	able to **extract info from usage patterns** of mobile phones + other wireless devices	
	Real-time BI	**use of data as and when it happens** • better ways to communication and make decisions that affect customers	
	Corp performance mgmt. (CPM)	• monitoring + managing performance in accord with KPIs • BI allows ppl to view info + insights re co. KPIs	
Big Data	Big Data	massive amounts of data / data sets	
	Volume	amount of data produce	
	Variety	different forms EG. msgs, updates + images posted to social media, readings sensors, GPS signals etc	
	Velocity	how fast data is produced + how fast data needs to be processed to meet demand	
	Data	facts + figures collected	
	Information	from analysing data	
	Data matrix	dataset that is stored digitally in spreadsheet or similar	
Dataset	Dataset	data collected in particular study	
	Entities	people, places or things which we store + maintain info	
	variable / attribute	characteristic of interest of entity	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><td></td><td></td><td>
<table>
<tr><td></td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td><td>G</td><td>H</td></tr>
<tr><td>1</td><td>Sales Transactions: July 14</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>Cust ID</td><td>Region</td><td>Payment</td><td>Transaction Code</td><td>Source</td><td>Amount</td><td>Product</td><td>Time Of Day</td></tr>
<tr><td>4</td><td>10001</td><td>East</td><td>Paypal</td><td>93816545</td><td>Web</td><td>$20.19</td><td>DVD</td><td>22:19</td></tr>
<tr><td>5</td><td>10002</td><td>West</td><td>Credit</td><td>74083490</td><td>Web</td><td>$17.85</td><td>DVD</td><td>13:27</td></tr>
<tr><td>6</td><td>10003</td><td>North</td><td>Credit</td><td>64942368</td><td>Web</td><td>$23.98</td><td>DVD</td><td>14:27</td></tr>
<tr><td>7</td><td>10004</td><td>West</td><td>Paypal</td><td>70560957</td><td>Email</td><td>$23.51</td><td>Book</td><td>15:38</td></tr>
<tr><td>8</td><td>10005</td><td>South</td><td>Credit</td><td>35208817</td><td>Web</td><td>$15.33</td><td>Book</td><td>15:21</td></tr>
<tr><td>9</td><td>10006</td><td>West</td><td>Paypal</td><td>20978903</td><td>Email</td><td>$17.30</td><td>DVD</td><td>13:11</td></tr>
<tr><td>10</td><td>10007</td><td>East</td><td>Credit</td><td>80103311</td><td>Web</td><td>$177.72</td><td>Book</td><td>21:59</td></tr>
<tr><td>11</td><td>10008</td><td>West</td><td>Credit</td><td>14132683</td><td>Web</td><td>$21.76</td><td>Book</td><td>4:04</td></tr>
<tr><td>12</td><td>10009</td><td>West</td><td>Paypal</td><td>40128225</td><td>Web</td><td>$15.92</td><td>DVD</td><td>19:35</td></tr>
<tr><td>13</td><td>10010</td><td>South</td><td>Paypal</td><td>49073721</td><td>Web</td><td>$23.39</td><td>DVD</td><td>13:26</td></tr>
</table>
</td><td></td></tr>
<tr><td></td><td>Record</td><td>set of measurements collected for particular entity is a record / observation</td><td></td></tr>
<tr><td></td><td>Data deluge</td><td>prevalence of:
• automatic data collection
• electronic instrumentation
• online transaction processing

• growing recognition of untapped value
• recognition is driving development of BA</td><td></td></tr>
</table>

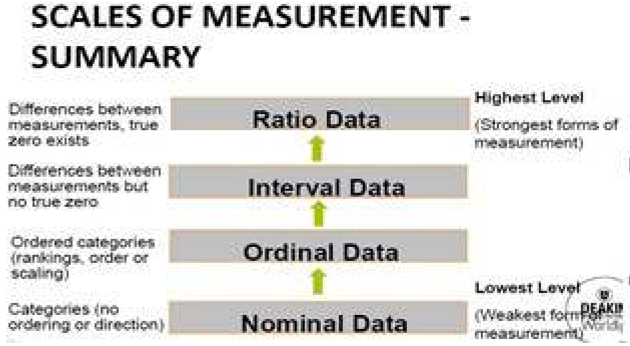
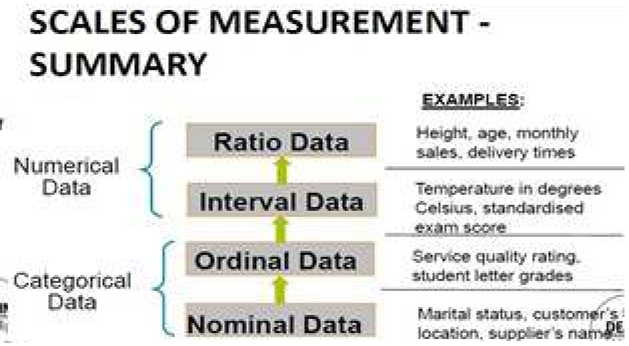
This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Week 2 (Topic 2) - Chapter TWO (What is Data?)			
Statistics	Statistics	**study of variation in data** relates to collection, analysis, interpretation & presentation of data statistical methods: • summarise collection of data • draw inferences about entire population • make predictions or forecasts	
statistical studies	experimental studies	the variables of interest are first identified → one or more factors are controlled so data can be obtained re how it affects variables	
	observational studies	no attempt is made to control or influence variables	
	Descriptive statistics	tabular, graphical + numerical **methods used to summarise data**	
	inference statistics	process of using data to obtain **from a sample to make estimates + test claims** re characteristics of population uses **sample data to reach conclusions re population** from which sample drawn Process of → Inference Statistics  	
Population	Population	**Entire collection** of objects (called units or subjects) of interest	
	Census	Collection of **data on population**	
	Sample	**Subset of units** in population • Representative of whole population • Sometimes this is only way to get info (eg. crash data on cars)	
Types of data	Exploratory data analysis (EDA)	• first step • precursor to more formal + extensive analysis • numerical, tabular + graphical summaries are produced to summarise + highlight key aspects or special	

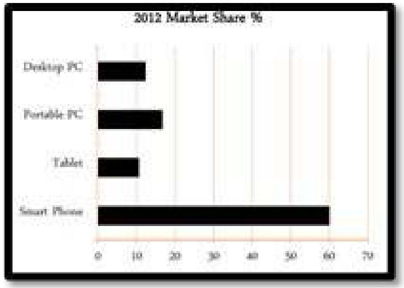
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		features of data • analysis sufficient for purpose of study	
	Parameter	**descriptive measure of <u>population</u>** • denoted by Greek letters μ = population mean σ^2 = population variance σ = population std deviation	
	Statistic	**descriptive measure of <u>sample</u>** • denoted by Roman letters $\bar{x}$ = sample mean s^2 = sample variance s = sample standard deviation	
	Parameter v Statistic	• estimate value of parameter is desirable but impossible/not feasible b/c of time + $$ required for consensus • instead use representative sample of pop + corresponding sample statistic to estimate parameter EG. F&P wants to determine avg. no. of loads its 8kg washing machine can wash before repairs. Population = **<u>all</u>** 8kg washing machines Parameter = population mean = avg no. of washers per machine before repair Population mean = no. of washes for type of machine estimated from sample mean Statistician takes representative sample of machines, conducts trials + records no. of washes before repair for each machine then computes sample avg. no. of washes before repair.	Research – don't quite understand
2 types of data	Qualitative data **aka** Categorical data	**descriptive data** subclass: nominal + ordinal • Labels or names used to identify attributes of each entity • Can be recorded in either numeric or nonnumeric formats • EG. 'Yes or no', 'male or female' answers • Usually counted or expressed as a portion or a percentage	
	Quantitative data **aka** Numerical data	**data that is expressed as number** subclass: discrete + continuous • Take numbers as their observed responses • Numerical data can be converted to categorical data. EG Salary can be converted into low/medium/high. • Cannot convert categorical data back to numerical data	
	Discrete	measuring how many (whole numbers + not able to be fraction)	
	Continuous	measuring how much (decimal/fraction)	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Scale of Measurement		 	
	Nominal (categorical)	classification of categorical data that implies no ranking EG. male or female? green or blue?	
	Ordinal (categorical)	scale of measurement where values are assigned by ranking EG. rating customer service: worst - excellent	
	Interval (numerical)	ranking numerical data where differences are meaningful but there is no true zero point EG. shoe size, amount of liquid	
	Ratio (numerical)	ranking numerical data where differences involve a true zero point EG. length, weight, age salary	
2 broad types of data	Cross-sectional data	relates to **group of items or individuals at given point in time**	
	Time ordered (time series) data	relates to particular **entity or situation at different points of time**	
Data Sources	Primary	data collector is one using data EG. survey, experiment	
	Secondary	another org. or individual has collected data used for analysis by org or individual	
	Web data	enhanced with social media data	
	Data in Business	used in: • annual reports • accounting audits • profitability analysis • eco trends • market research	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><th colspan="4">Week 3 (Topic 3) - Chapter THREE (Exploring Data - Data Visualisation)</th></tr>
<tr><td>Purpose of Data Visualisatio n</td><td>Analysis</td><td>purpose = insight - discovery, decision making + explanation
• able to comprehend huge amounts of data
• important info from millions of data immediately visible and more comprehendible
• provides insight into pattersn + relos not readily visible
• enables problems with data to become apparent</td><td></td></tr>
<tr><td>Categorical Data</td><td>Frequency table</td><td>gives frequency, proportion or % of value in each category / class
<table>
<tr><th>Type of device</th><th>2012 Shipments (in millions)</th><th>2012 Market Share</th></tr>
<tr><td>Smart Phone</td><td>722.4</td><td>60.1%</td></tr>
<tr><td>Tablet</td><td>128.3</td><td>10.7%</td></tr>
<tr><td>Portable PC</td><td>202</td><td>16.8%</td></tr>
<tr><td>Desktop PC</td><td>148.4</td><td>12.4%</td></tr>
<tr><td>Total</td><td>1201.1</td><td>100%</td></tr>
</table></td><td></td></tr>
<tr><td>Categorical Data</td><td>Bar chart / Column chart</td><td>each category represented by a bar - the length indicates frequency, proportion or % of value in each category


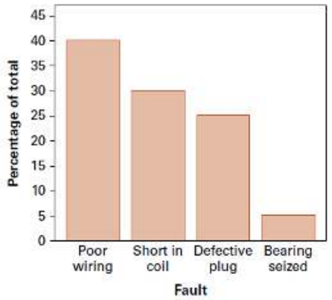
**use if <u>comparing categories</u> is most important</td><td></td></tr>
<tr><td>Categorical Data</td><td>Pie graph</td><td>circle used to represent total - is divided into "slices", each representing category
(used to observe proportion / market share)


**use if <u>observing portion of whole</u> that lies in each category is most important</td><td></td></tr>
</table>

This document is available free of charge on StuDocu.com

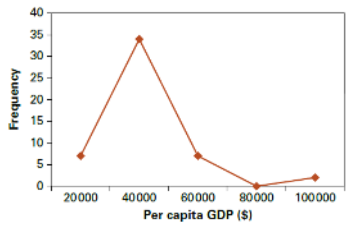
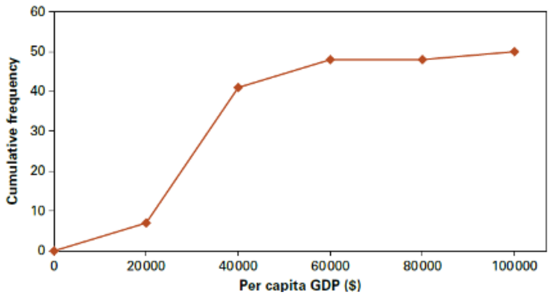
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><td>Categorical Data</td><td>Pareto chart</td><td>verticle bar chart where no. + type are in order of magnitute (greatest to least)

**used to <u>display causes of problems in products + processes</u>

</td><td></td></tr>
<tr><td>Numerical Data</td><td></td><td>Common tabular + graphical techniques for organising + previewing include:
• arrays
• frequency table / summary table
• pie chart
• bar chart
• dot plots

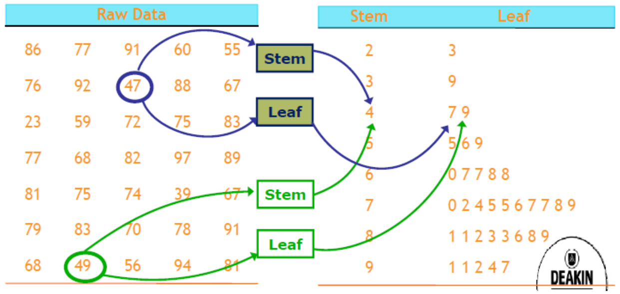
useful tool to summarise:
• small dataset (<20 observations)
• numrical data that is discrete + repeats frequently</td><td></td></tr>
<tr><td>Numerical Data</td><td>Arrays</td><td>putting data into ascending order</td><td></td></tr>
<tr><td>Numerical Data</td><td>Frequency distribution</td><td>summary table of data arranged into classes
• rules:
  - use 5 - 15 calsses
  - class width = range/no. of classes
  - centre = mid point (used for graphing)
• no. of observations in each class = frequency of class
• allows quick visual interpretation of data
• allows first look at shape of data
<table>
<tr><th>Weekly Sales</th><th>Count</th><th>Percentage</th><th>Cum. Percentage</th></tr>
<tr><td>0 kg < 200 kg</td><td>3</td><td>5.8%</td><td>5.8%</td></tr>
<tr><td>200 kg to < 400 kg</td><td>10</td><td>19.2%</td><td>25%</td></tr>
<tr><td>400 kg < 600 kg</td><td>16</td><td>30.8%</td><td>55.8%</td></tr>
<tr><td>600 kg < 800 kg</td><td>16</td><td>30.8%</td><td>86.6%</td></tr>
<tr><td>800 kg < 1000 kg</td><td>6</td><td>11.5%</td><td>98.1%</td></tr>
<tr><td>1000 kg < 1200 kg</td><td>1</td><td>1.9%</td><td>100%</td></tr>
<tr><td>Total</td><td></td><td>100%</td><td></td></tr>
</table>
**use for large datasets + non-repeating values to summarise data</td><td></td></tr>
<tr><td></td><td>relative freq distribution</td><td>Equation: $\frac{frequency \in each\, cla\,s\,s}{total\, no.\, of\ value}$</td><td></td></tr>
<tr><td></td><td>% distribution</td><td>Equation: $each\, relative\ frequency \times 100$</td><td></td></tr>
<tr><td></td><td>cumulative % distribution</td><td>% of values that are less than certain value</td><td></td></tr>
</table>

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Numerical Data	Histogram	Graphical representation of frequency, relative frequency, % distribution tables • Allows for a representation of the shape of the data set (skewness)  **use **for large datasets + non-repeating values** to summarise data	
Numerical Data	Frequency polygon	• graph constructed by plotting dot **for frquency at class mid points** + connecting the dots  	
Numerical Data	Ogive	• cumulative frequency polygon • plotted by graphing dot at each class **endpoint for cumulative frequncy** value + connecting dots  	
Numerical Data	Stem + leaf plots	• another way to display **continuous data** • major advantage is original data preserved • displays info similar to histogram • plot numbers constructed by separating each no. into 2 groups = stem + leaf • leftmost digits are stems • rightmost are leaves	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr>
<td></td>
<td></td>
<td>


<table>
<tr><th>Stem</th><th>Leaf</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>2</td><td>3</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>9</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>4</td><td>7</td><td>9</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>5</td><td>6</td><td>9</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>6</td><td>0</td><td>7</td><td>7</td><td>8</td><td>8</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>7</td><td>0</td><td>2</td><td>4</td><td>5</td><td>5</td><td>6</td><td>7</td><td>7</td><td>8</td><td>9</td></tr>
<tr><td>8</td><td>1</td><td>1</td><td>2</td><td>3</td><td>3</td><td>6</td><td>8</td><td>9</td><td></td><td></td></tr>
<tr><td>9</td><td>1</td><td>1</td><td>2</td><td>4</td><td>7</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>
</td>
<td></td>
</tr>
<tr>
<td>Purpose of Data Visualisation</td>
<td>Communication</td>
<td>

tell story or show pattern that has already been discovered in data
focuses on the msg - clear + easy to understand
focus is on design of visualisation = should be honest, unambiguous + effective presentation of data

</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tufte's Principles of graphical display</td>
<td>
Graphical displays should:

show the data
tell the truth (avoid distorting what the data has to say)
focus on the content (help the viewer think about the information rather than design)
encourage the eye to compare the data
make large data sets coherent
revel data at several levels of detail
Closely integrate statistical and verbal descriptions

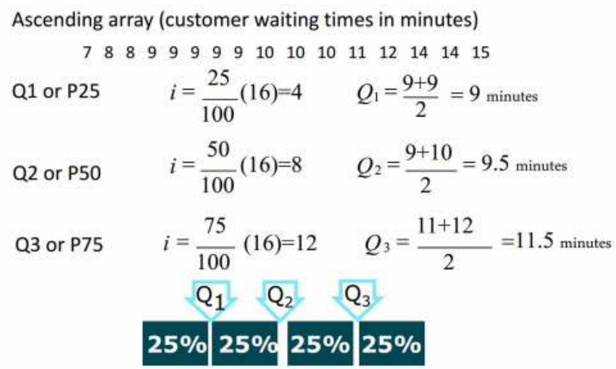
</td>
<td></td>
</tr>
</table>

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

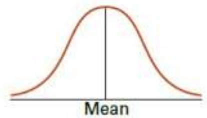
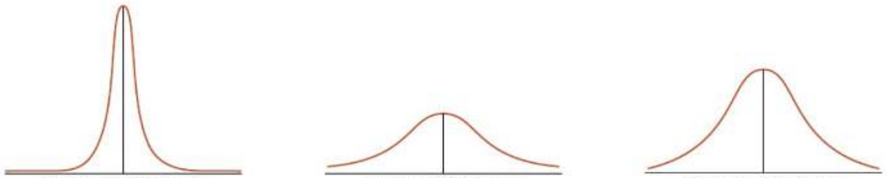
Week 4 (Topic 4) - Chapter FOUR (Numerical Summaries - Exploring Relationships)			
Summary measures	summary measures	**Single figure which attempts to sumarise particular feature of set of data** • if measures are computed for data from **population = parameters** • if meausres are computed for data from **sample = statistics** • sample statisitcs is a point estimator of population parameter	
What looking for in summary measures?	Measure of Central tendency	Are averages: • mean: common average (arithmetic mean) • median: middle value • mode: most common value	
	Median	**middle value** of data set arranged in ascending order • is a measure of location – often reported for annual income + property value data • if data set has any extreme values, median is preferred measure of central tendency • extreme values can inflate the mean making it less typical	
	Mode	value that occurs with **greatest frequency** • greatest frequency can occur at 2+ different values • if data has exactly 2 modes = data is bimodal • if data has 2+ modes = multimodal • if data has 1 mode = nominal	
What looking for in summary measures?	Measure of Location	• Averages are measures of location – indicating middle or centre of data set • More specific measures of location can be found from ascending array: min, max, quartiles, percentiles	
	Percentiles	• Tell us location of certain percentages of data EG. Top 10%, smallest 1% etc Process: 1. Organise data in ascending array 2. Calculate percentile location: $i=\frac{P}{100}(n)$ - If i is a whole number, percentile is average of values at the i and (next to i) positions - If i is the whole number, the percentile value is found by rounding i up to the whole no. + reporting value at this position	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Quartiles	• Specific percentiles that are commonly used: - 1st quartile: 25th percentile - 2nd quartile: 50th percentile = median - 3rd quartile: 75th percentile 	
	Minimum	Lowest value	
	Maximum	Highest value	
What looking for in summary measures?	Measure of Variability	Describes the **spread or dispersion of data set** • Distance measures: range, interquartile range • Average variability: Standard deviation, and variance • Relative variability: coefficient of variation	
Distance measure	Range	• Range of data set is **difference b/w largest & smallest values** • Ignores all data points except 2 extreme ends of data set • Very sensitive to smallest & largest data values	
Distance measure	Inter Quartile Range (IQR)	• Difference b/w 3rd quartile & 1st quartile $Q_1 - Q_3$ • Range of middle 50% of data • NOT sensitive to extreme data values	
Average variation	Variation	Variation is expressed in squared units	
	Coefficient of Variation	Indicates how large the std dev is in relation to the mean • Calculated as $\frac{std\ dev}{mean}$ • Expressed at % • Relative measure of variation **Useful for comparing variability b/w data sets in different magnitudes or diff units	
	Std Dev	estimate of **average deviation of value away from mean** • Maintains original unit → preferred • Popular measure of risk (esp. financial analysis)	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

What looking for in summary measures?	Measure of Shape	• Distribution of data where shape is **symmetrical** (bell curve)  • Distribution of data where shape is **skewed / asymmetric** (lacks symmetry) - Higher negative value = negatively skewed - Higher positive value = positively skewed  	
	Kurtosis	Peakness of distribution:  	
Relative location	Z scores	Relative measure of distance is an observation from the mean (re std dev) • If z score is +3 or -3 = outlier For example: – A dataset is normally distributed with a mean of 60 and standard deviation of 5. Determine the z score for a value of 70 $z = \frac{x-\mu}{\sigma}$ $z = \frac{70-60}{5} = 2.0$ – a Z score of 2 means that 70 is 2 standard deviations above the mean	
Relative location	Chebyshev's theorem	• Applies to <u>**all**</u> distributions: - At least 75% of the data values must be within Z=2 Standard deviations of the mean - At least 89% of the data values must be within Z=3 Standard deviations of the mean - At least 94% of the data values must be within Z=4 Standard deviations of the mean	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><td></td><td></td><td>

- At least $1 - 1/k^2$ values fall within + or $-k$
- Std dev of mean, regardless of shape of distribution

<table>
<tr><th>Number of Standard Deviations</th><th>Distance from the Mean</th><th>Minimum Proportion of Values Falling Within Distance</th></tr>
<tr><td>K = 2</td><td>$\mu \pm 2\sigma$</td><td>$1-1/2^2 = 0.75$</td></tr>
<tr><td>K = 3</td><td>$\mu \pm 3\sigma$</td><td>$1-1/3^2 = 0.89$</td></tr>
<tr><td>K = 4</td><td>$\mu \pm 4\sigma$</td><td>$1-1/4^2 = 0.94$</td></tr>
</table>

</td><td></td></tr>
<tr><td>Relative location</td><td>Empirical rule</td><td>

- Guideline that states the approx. % of values that fall within given no. of std devs of mean = normally distributed (bell shaped)
- For data that is **symmetrical** / bell shaped:
  - Approx 68% of the data values lie within Z=1 Standard deviations of the mean
  - Approx 95% of the data values lie within Z=2 Standard deviations of the mean
  - Approx 99.7% of the data values lie within Z=3 Standard deviations of the mean

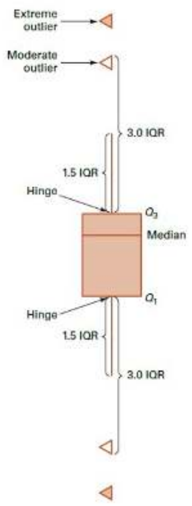
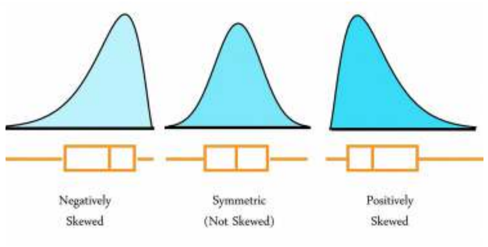

</td><td></td></tr>
</table>

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

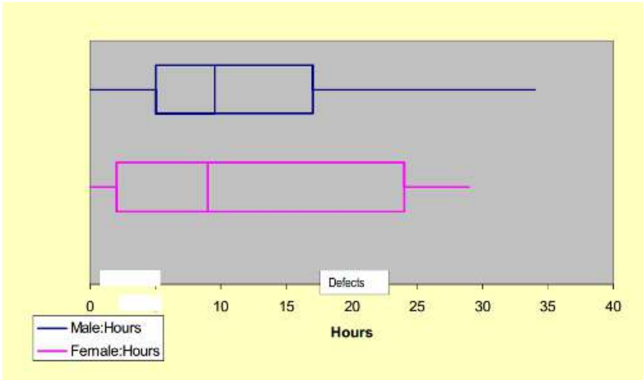
What looking for in summary measures?	Outliers	Data values that fall so far from average they are considered unusual • Extreme extreme values • Tend to be separated from rest of data • Detecting outliers is important = need to be investigated more • May indicate: - Incorrectly recorded data value - Data value that was incorrectly included in data set - Potential problem - Potential opportunity • Detecting Outliers Graphically - Displaying data graphically = good way to spot potential outliers - Calculation rules are <u>not</u> the “be all and end all” of outlier detection - Various graphs can be used. Dot plots and box plots are often useful	
Outliers	Empirical rule	• For **symmetrical bell-shaped data** • Further than 3 std dev from mean = potential outlier • Any data with z-score 3+ or -3 = potential outlier	
Outliers	Tukey’s 1.5 Step rule	• For non-bell shaped distributions • Works by calculating limits (fences) by determining quartiles + IQR - Lower fence = 1.5(IQR) below Q1 - Upper fence = 1.5(IQR) above Q3 • Value outside fences = potential outlier	
	EXAMPLE: Tukey's rule	• Variable: Number of Daily iPhone Sales for 12 months for a Telco Company. • Summary Measures: - Min 13; - Q1 24; - Q3 46; - Max 97 (Positively Skewed Distribution) • Potential Outliers? - Lower fence: Q1-1.5(IQR) = 24-1.5 (22) = 24-33 = -9 - Upper fence: Q3+1.5(IQR) = 46+1.5(22)= 46 + 33 = 79 • ∴ at least one potential outlier as 97 sales <u>does not</u> fall within the limits.	

This document is available free of charge on StuDocu.com

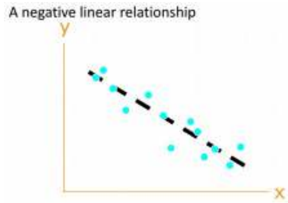
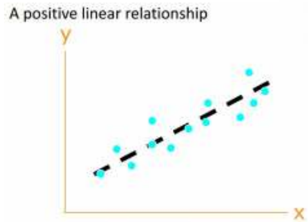
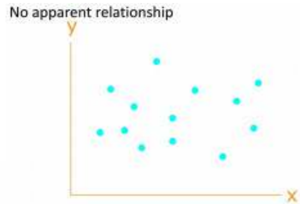
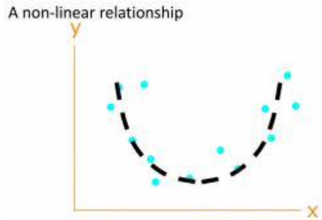
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Box + Whisker Plot	• Five specific values are used: 1. Min value 2. 1st quartile 3. 2nd quartile / median 4. 3rd quartile 5. Max value • Inner fences - IQR $Q_1 - Q_3$ - Lower inner fence = Q1 – 1.5IQR - Upper inner fence = Q3 + 1.5IQR • Outer fence - Lower outer fence = Q1 – 3IQR - Upper outer fence = Q3 + 3.0 IQR   	
What summary measure to use	Symmetrical distribution	Use: • Mean • Std dev	
	Skewed distribution	Use: • Median (mean = too distorted) • IQR	
Process: Relationship b/w 2 variables	Relationship b/ w **TWO** variables	1. Classify data (numerical / categorical) 2. Decide which variable is dependent variable & independent / explanatory variable 3. Select technique 4. Look for **differences**	
	Techniques to use	• Numerical Dependent & Categorical Independent - Comparative summary measures - Multiple box plots • Categorical Dependent & Categorical Dependent - Cross tabulations (contingency tables) • Numerical Dependent & Numerical Independent - Scatter diagrams • Categorical Dependent & Numerical Independent - Numerical data can be converted to categorical data – cross tabulation	
	Comparative summary measures	• **Substantial** differences = relationship • **No** or **small** differences = **no** relationship	
	Multiple Box	• Useful way of displaying single numerical variable using 5 no. summary	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Plots	• Places more than 1 box plot in single graph • Enables us to quickly compare features + patterns across subgroups  	
	Cross tabulations	• Displays results of 2 categorical variables together in 1 table • Depend on type of cross tabulation, cells may contain: - Absolute frequencies - Relative frequencies • Can be misleading if sample size of categories not similar • Must calculate % • % chosen to analyse should be where independent variable is located: - If located in row = row % to compare - If column = column % to compare If % across row / column similar = no relationship b/w 2 categorical variables If % across rows dissimilar to columns = relationship b/w 2 categorical variables	
	Relative frequencies constructed in 3 ways:	• % of overall total • % of row total • % of column total	
	Scatter diagram	• Graphical presentation of relo b/w 2 numerical variables • Independent variable shown on horizontal axis (x) • Dependent variable shown on vertical axis (y) • Pattern of plotted points suggest overall relo b/w variables • Trend line = approximation of relationship	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Linear relationship	• closer points are to trend line = stronger relo • measure direction + strength of relo = **correlation coefficient (r)** • measure absolute strength of relo = **coefficient of determination (r^2)**    	
	Causality	• Independent (explanatory) variable **does not have** causal effect on dependent variable • Cannot be demonstrated by data analysis techniques alone – must be throguh experiments where environment is controlled	
SUMMARY		• More spread out the data: the larger the range + IQR + SD • The more concentrated or similar the data: the smaller the range+ IQR + SD • If the value are the same: the range + IQR + SD will be zero • No measure of variation can ever be negative	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><th colspan="4">Week 5 (Topic 5) - Chapter FIVE (Discrete Probability Distributions)</th></tr>
<tr><td></td><td>Probability & Probability distributions</td><td>Enable us to develop models that take into account uncertainty
• Decision maker has evidence to base outcome on</td><td></td></tr>
<tr><td></td><td>Probability distributions</td><td>distribution of all possible values of random varaibela nd correpsonding probabilities</td><td></td></tr>
<tr><td></td><td>Uncertainty & risk</td><td>• A key aspect of solving business problems is dealing with uncertainty
EG. Not knowing how much stock to produce

• Problem solving also involves risk, which depends on the position of the business/decision maker
-EG. Interest rates have less effect on businesses that sell essential items to those that sell luxury items</td><td></td></tr>
<tr><td></td><td>Probability properties & rules</td><td>• Event X $\text{¿}0 \leq P(X) \leq 1$
• Probabilities of all e vents must = 1 ∑P(X) = 1</td><td></td></tr>
<tr><td>Assigning probabilities</td><td>Classical method

(priori classical probability)</td><td>Occurs with games of chance or where all outcomes are known + probabilities are fixed
$\text{¿}\frac{no.of\ ways\ events\ can\ occur}{total\ no.of\ possible\ outcomes}$
• Example: Rolling a die. 
• P(rolling a 4) = 1/6 
• P(rolling an odd number) = 3/6 
• P(rolling more than 4) = 2/6 </td><td></td></tr>
<tr><td>Assigning probabilities</td><td>Relative frequency method

(empirical classical probability)</td><td>Observed (histroical) data
• past surveys or observations can provide insight to what may occur in the future
• it is done thorugh a <u>probability distribution</u>

Formula: P(exactly 2) = P(x = 2) = 0.45 (see table →)

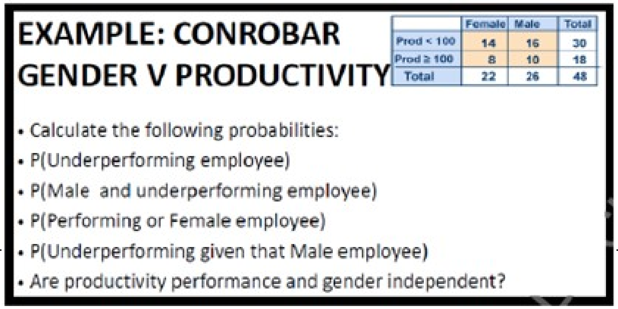
Number of insurance claims per day<table><tr><th>Num of Claims</th><th>Number of Days</th><th>Probability P(X)</th></tr><tr><td>0</td><td>4</td><td>0.10 = 4/40</td></tr><tr><td>1</td><td>6</td><td>0.15 = 6/40</td></tr><tr><td>2</td><td>18</td><td>0.45 = 18/40</td></tr><tr><td>3</td><td>10</td><td>0.25 = 10/40</td></tr><tr><td>4</td><td>2</td><td>0.05 = 2/40</td></tr><tr><td></td><td>40</td><td>1.00 = 40/40</td></tr></table>P(Exactly 2) = P(X = 2) = 0.45
P(3 or more in a day) = P(X ≥ 3) = 0.25+0.05 = 0.30</td><td></td></tr>
<tr><td>Assigning probabilities</td><td>Subjective method

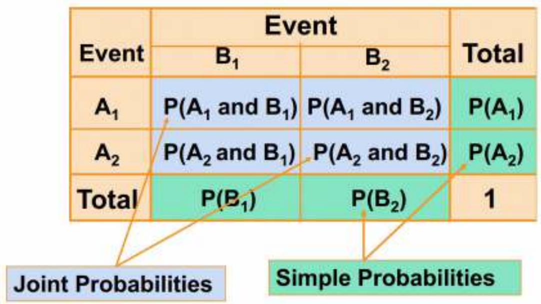
(guess)</td><td>Use this method if there is no method of obtaining probabilities from past experience or mathematical distributions

EG. Economic conditions + company’s circumstances change rapidly → may be inappropriate to assign probabilties based on historical data</td><td></td></tr>
</table>

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		• may have to adjust probabilities if past figures are no longer suitable in current conditions EG. production manager 'feels' that typically 1 of 500 produced as manufacturing fault • Estimate probabilities using data, experience + intuition available • This subjective probability value express our degree of belief of what will occur	
Probability rules + properties	Complement of an event A	All outcomes are not part of event A	
Probability rules + properties	Joint event (AND)	Involves 2+ characteristics occurrung sumultaneously	
Probability rules + properties	Mutually exclusive events	Events that cannot occur together	
Probability rules + properties	Collectively exhaustive events	Set of events such that one of the events must occur	
Probability rules + properties	General addition rule (OR)	P (A or B) = P(A) + P(B) – P(A and B)	
Probability rules + properties	Conditional probabilities	P(A | B) = $\frac{P(A \wedge B)}{P(B)}$	
	If probability of event A is not changed by the existence of event B …	… then A and B are independent P(A | B) = P(A)	
	Cross-tabs (*contingency tables*)	**sample space for joint events** classified by 2 characteristics • Summarise past data • Can be used to provide variety of probability estimates	





Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Discrete random variable	• Mean tells us what we expect will happen on average • SD tells us the expected dev around this average	

Expected values — Expected **value** of X (mean):

tells us what we expect will **happen on avg**

formula:

$$E(x)=\mu=\sum XP(x)$$

Number of insurance claims per day

X	P(X)	P(X).X
0	0.10	0.00
1	0.15	0.15
2	0.45	0.90
3	0.25	0.75
4	0.05	0.20
	1.00	**2.00**

Expected number of claims/day is: E(X) = 2.00

Expected values — Expected **variation** of X (std dev):

tells us the expected **variation around the avg**

$$SD(x)=\sigma=\sqrt{(\sum (X-\mu)^2P(X))}$$

Number of insurance claims per day

X	X - μ	(X - μ)²	P(X)	(X - μ)²P(X)
0	-2	4	0.10	0.40
1	-1	1	0.15	0.15
2	0	0	0.45	0.00
3	1	1	0.25	0.25
4	2	4	0.05	0.20
				1.00 = σ^2

The SD of claims is √1 = 1 claim

Pay off tables — ***EXAMPLE:*** Pay off table

- Makes 10c per egg sold on 1st round market
- Otherwise **loses** 15 c per egg sold on 2nd round market

Q4.

	Demand				
Supply	10	11	12	13	14
10	1000	1000	1000	1000	1000
11	1000-150= 850	$1,100	$1,100	$1,100	$1,100
12	1000-300= 700	1100-150= 950	1200	1200	1200
13	1000-450= 550	1100-300= 800	1200-150= 1050	1300	1300
14	1000-600= 400	1100-450= 650	1200-300= 900	1300-150= 1150	1400

1. Expected Profitability values = (Probability x Profit)

Supply 10'000 eggs	10	11	12	13	14
Profit	1000	1000	1000	1000	1000
Probability	0.2	0.2	0.3	0.2	0.1
Profitability x Probability	200	20[illegible]			

Supply 11'000 eggs	10	11	12	13	14
Profit	850	1100	1100	1100	1100
Probability	0.2	0.2	0.3	0.2	0.1
				220	110

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Supply 12'000 eggs	10	11	12	13	14
Profit	700	950	1200	1200	1200
Probability	0.2	0.2	0.3	0.2	0.1
Profitability x Probability	140	190	360	240	120

Supply 14'000 eggs	10	11	12	13	14
Profit	400	650	900	1150	1400
Probability	0.2	0.2	0.3	0.2	0.1
Profitability x Probability	80	130	270	230	140

Supply 13'000 eggs	10	11	12	13	14
Profit	550	800	1050	1300	1300
Probability	0.2	0.2	0.3	0.2	0.1
Profitability x Probability	110	160	315	260	130

Expected Profit:	
10	$1,000
11	$1,050
12	$1,050
13	$975
14	$850

-The manager of the Waverly store is not correct in saying that ordering the maximum possible quantity will return a higher profit.

-This is because the stores who ordered 11000 and 12000 eggs experienced the highest profit at $1050. Whereas the store who ordered the most eggs (14000) experienced the lowest profit of $850.

-Probability provides (and data) provides evidence for a decision maker to justify the chosen outcome

Another Example:

Q4 The Colchester Garden Centre purchases and sells Christmas trees during the holiday season. It purchases the trees for $10 each and sells them for $20 each. Any trees not sold by Christmas day are sold for $2 each to a company that makes wood chips. Suppose that the probabilities of the demand for the different number of trees are as follows:

Demand (Number of Trees)	Probability
100	0.20
200	0.50
500	0.20
1000	0.10

Following are payoffs for purchasing 100, 200, 500 or 1000 trees.

a) Complete the missing payoffs;

	Probability	Purchase 100	Purchase 200	Purchase 500	Purchase 1000
Demand 100	0.2	1000	200	-2200	-6200
Demand 200	0.5	1000	2000	-400	-4400
Demand 500	0.2	1000	2000	5000	1000
Demand 1000	0.1	1000	2000	5000	10000

b) Calculate the expected values and suggest how many trees to purchase.

Expected Value

Purchase 100: $1000
Purchase 200: $1640 <<purchase 200 trees>>
Purchase 500: $860
Purchase 1000: -$2240

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Mathematical probability **distributions**	• Important source of probabilities • Probabilities can be derived <u>w/o</u> need for data • Each distribution has own unique + important characteristics & properties	
Type of mathematical probability distribution	Binominal distribution properties	<u>4 types of properties</u>: 1. Fixed no. of identical trials/observations (n) EG. 15 tosses of a coin; 10 light bulbs taken from a warehouse 2. Each trial has only two possible outcomes • called 'success' and 'failure' • probability of success is p, probability of failure is q = 1- EG. head/tail in each toss of a coin; defective/not defective light bulb 3. Constant probability for each trial EG. Probability of a tail is the same each time we toss the coin 4. observations are independent • means 1 outcome does not affect another outcome • to ensure independence - observations are randomly sampled	
Binominal distribution properties	Applications	• sampling with replacement • sampling w/o replacement -n < 5% N	
how to calculate binominal distributions :	Mean	$\mu = E(x) = np$ n = no. of people in sample p = probability	
Binominal distribution characteristics	Variance & Std Dev	$\sigma^2 = np(1-p)$ $\sigma = \sqrt{np(1-p)}$ n = no. of people in sample p = probability	
Binominal formula	Compute probabilities	$P(x) = \frac{n!}{x!(n-x)!} p^x . q^{n-x}$ for $o \leq x \leq n$ Where: • *n* = sample size • *p* = probability	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

- $P(x)$ = probability of x success given n and p

Binominal table

Compute probabilities

Demonstration Problem 5.4

find binominal probabilty for:

- n=20
- p=0.4
- x=10

	$n = 20$								
	Probability								
x	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	.122	.012	.001	.000	.000	.000	.000	.000	.000
1	.270	.058	.007	.000	.000	.000	.000	.000	.000
2	.285	.137	.028	.003	.000	.000	.000	.000	.000
3	.190	.205	.072	.012	.001	.000	.000	.000	.000
4	.090	.218	.130	.035	.005	.000	.000	.000	.000
5	.032	.175	.179	.075	.015	.001	.000	.000	.000
6	.009	.109	.192	.124	.037	.005	.000	.000	.000
7	.002	.055	.164	.166	.074	.015	.001	.000	.000
8	.000	.022	.114	.180	.120	.035	.004	.000	.000
9	.000	.007	.065	.160	.160	.071	.012	.000	.000
10	.000	.002	.031	.117	.176	.117	.031	.002	.000
11	.000	.000	.012	.071	.160	.160	.065	.007	.000
12	.000	.000	.004	.035	.120	.180	.114	.022	.000
13	.000	.000	.001	.015	.074	.166	.164	.055	.002
14	.000	.000	.000	.005	.037	.124	.192	.109	.009
15	.000	.000	.000	.001	.015	.075	.179	.175	.032
16	.000	.000	.000	.000	.005	.035	.130	.218	.090
17	.000	.000	.000	.000	.001	.012	.072	.205	.190
18	.000	.000	.000	.000	.000	.003	.028	.137	.285
19	.000	.000	.000	.000	.000	.000	.007	.058	.270
20	.000	.000	.000	.000	.000	.000	.001	.012	.122

$$n = 20$$

$$p = 0.40$$

$$P(X = 10) = 0.117$$

Binominal distribution

EXAMPLE: binominal distribution

A components manufacturer samples 20 products randomly from each shift to monitor quality. Over an extended period they have found that that the number of defectives is approximately 10% of all products made.

Explain why the Binomial distribution is appropriate in this case.

n= 20 number of trials (products)

Two possible outcomes for each trial (defective/not defective)

P(defective) = 0.10; P(not defective) = 0.90

Products independent (randomly selected)

What would be the expected number of defectives in each shift.

$E(X) = n \times p = 20 \times 0.10 = 2$ products

In a particular shift, five (5) defectives were found. Find the probability of this occurring (or worse). Comment on your answer.

Want $P(X \geq 5)$ $= 1 - P(X < 5)$

$= 1 - 0.957$

$= 0.043$

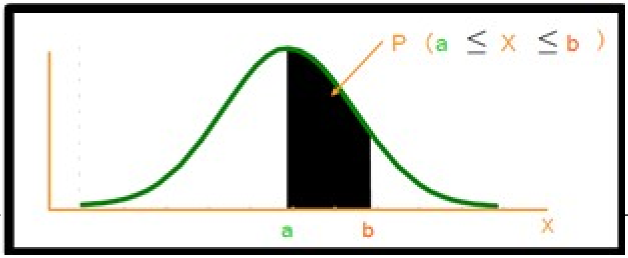
Small chance (4.3%) to get five or more defective products.

There may be a problem with this particular shift.

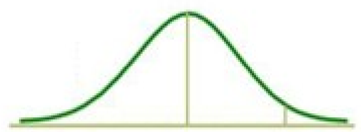
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

how to recognise Poisson distributions	Poisson distribution properties	<u>4 essential properties:</u> 1. You wish to count no. of occurrences over an interval: - Each occurrence is independent of other occurrences - No. of occurrences in each interval can range from 0 to ∞ - Average (expected) no. of occurrences in interval is λ (lambda) EG. ▪ seeing how many pixel burnouts there are in TV surface area ▪ No. of telephone calls per minute at a small business ▪ No. of customer arrivals at a bank in an hour ▪ No. of paint spots per new vehicle ▪ No. of units of product demanded per week 2. probability that an event occurs in 1 area of opportunity = same for all areas of opportunity EG. Pixel burnout probability for TV screen is same for middle of screen, top + bottom 3. no. of events that occur in 1 area of opportunity independent of no. of events that occur in other areas of opp EG. 1 computer crash will not affect another computer crash 4. probability that 2+ events occur in an area of opportunity approaches zero as the area of opportunity becomes smaller EG. when you focus whole TV there is greater probability of pixel burnout - if focus only on bottom of TV = less likely to get burnout	
Calculate poisson distributions	Poisson distribution characteristics	Mean: $\mu = ¿$ Variable + Std Dev: $\sigma^2 = \lambda$ $\sigma = \sqrt{\square}$ Where λ = expected no. of occurrences in the interval	
Poisson formula	Calculate probabilities	$P(x) = \square^x \frac{e^{-¿}}{x!} ¿$ For x = 0, 1, 2, 3, … Where: λ = long – run average *E* = 2.718282 … (base of natural logarithms)	

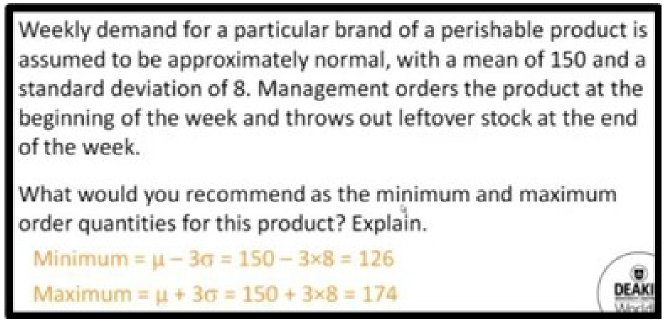
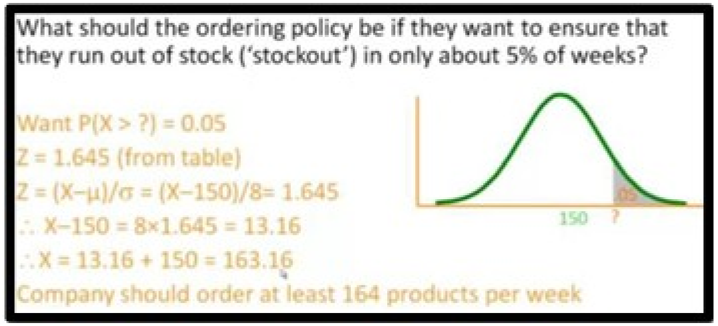
This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Poisson distribution	***EXAMPLE***: Poisson distribution	The number of times a company's HR system "crashes" is, on average, 3.3 per month. Explain why the Poisson distribution would apply in this case. Interested in number of times a crash occurs over an interval (one month period). Assume that a crash in a one month interval has no effect on any other crash during any one month interval (independent). Number of crashes in a month has no upper limit. Expected number of crashes (lambda) per month is known. Calculate the probability that the system does not crash in a given month. Want $P(X = 0) = 0.0369$ What is the probability it does crash three or more times in a given month. Want $P(X \geq 3) = \cancel{P(X=3)+P(X=4)+P(X=5)+\dots}$ $= 1 - P(X < 3)$ $= 1 - 0.3594 = 0.6406$	
	Probability & decision making	• Probabilities can be used to measure lvl of uncertainty involved • Use of probabilities in decision making does not necessarily mean we get a desired outcome • Common mistake is to associate good (poor) decisions with good (poor) outcomes • May be the case that a persons a sound decision (w high probability) but due to uncertainty of situation = unlucky + diff outcome subsequently occurred	
Normal distributions	**characteristics** of normal distributions	• bell shaped • symmetrical • mean, median + mode are equal • location is determined by mean • spread is determined by std dev • depends on 2 parameters: Mean + std dev	
Normal distributions	calculate normal distribution	• probability measured by area under curve: 	

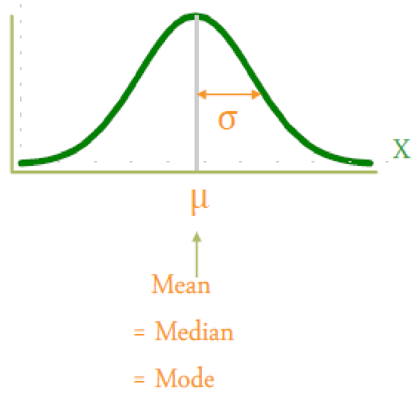
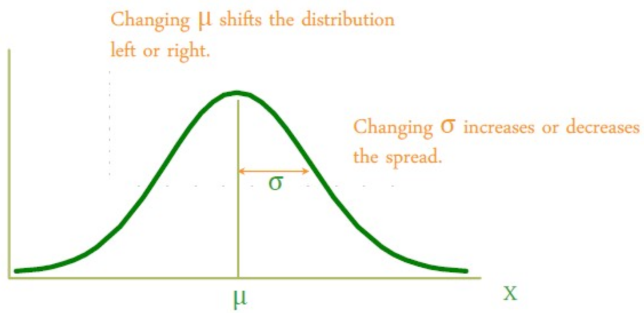
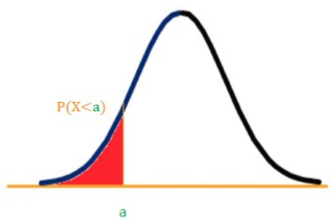
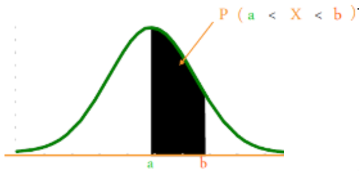
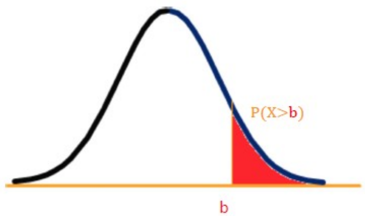
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		-Need to transform X units into Z units: $$Z = \frac{X - \mu}{\sigma}$$ The Z distribution has mean = 0 and standard deviation = 1 **COMPARING X AND Z UNITS**  Note that the distribution is the same, only the scale has changed. We can express the problem in original units (X) or in standardised units (Z)	
	EXAMPLE: normal distribution	If X is distributed normally with mean of 100 and standard deviation of 50, then: $$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$ This says that X = 200 is two standard deviations above the mean of 100 **GENERAL PROCEDURE FOR FINDING PROBABILITIES** To find P(a < X < b) when X is distributed normally: 1. Draw the normal curve for the problem in terms of X 2. Translate X-values to Z-values 3. Use the Standardised Normal Table or software to find the area **FINDING THE X VALUE FOR A KNOWN PROBABILITY** Steps to find the X value for a known probability: 1. Draw the normal curve for the problem showing the known probability 2. Find the Z value for the known probability 3. Convert to X units using the formula	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

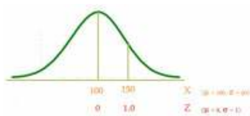
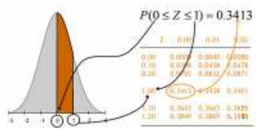
<table>
<tr><td></td><td>EXAMPLE:
finding probabilities</td><td>
</td><td></td></tr>
<tr><td></td><td>EXAMPLE:
finding X value for known probability</td><td>
</td><td></td></tr>
</table>

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Week 6 (Topic 6) - Chapter FIVE (Continuous Probability Distribution)			
	Continuous distributions	• assoicated with **random variables that take values at any point over given interval** • uniform + exponential distributions are important beause normal distributionis most widely encountered continuous distribution	
	continuous probability distributions	• no longer talk of probability of random variable being equal to specific value • talk about probability of random variable having value within given range • area under probability curve b/w 2 given values = the probabilitiy that random variable lies b/w the 2 values	
	characteristics of normal distribution	1. Bell-shaped 2. Symmetrical distribution 3. Continuous distribution 4. Mean = median = mode (all equal) 5. Location (centre) is determined by the mean (μ) 6. Spread is determined by the standard deviation (σ) 7. Area under the curve is equal to 1  	
finding normal probabilities	Normal probability	8. measured by area 9. **under the curve**   	
	Standardised normal distribution	10. any normal distribution (with any mean + std dev combination) can be transformed into **standardised normal distribution (Z)** 11. need to transer X units into **Z** units	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		Formula: $Z=\frac{X-\mu}{\sigma}$ * standardised normal (z) distribution has mean = 0 std dev = 1	
Standardised normal distribution	***EXAMPLE:*** stdardised normal distribution	2. If X is **distributed normally** with mean of 100 and std dev of 50, z-score for X = 150 is: $z=\frac{X-\mu}{\sigma}=\frac{150-100}{50}=1.0$ 3. Z-score = no. of std dev particular value (X) is away from mean This says that X = 150 is 1 std dev above mean of 100	
	Comparing X & Z	disribution is the same • scale has changed • express problem in standardised units (Z)  	
Procedure for finding probabilities	Find **P(a< X <b)** when X is distributed normally	Find **P(a< X <b)** when X is distributed normally: 1. draw normal curve for problem in terms of X 2. translate X-values to Z-values 3. Use standardised Normal Table or software to find the area (probability)	
	Table look up of std normal probability	 	
Finding X for known probability (Reverse	Finding X for known probability (**Reverse**	Steps to find X value for known probability: 1. draw the normal curve for probalm showing area of interest 2. find probability (area) among values in body of table 3. determine value of Z (in left tail Z must have (neg) -sign, right tail Z has (pos) +sign)	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

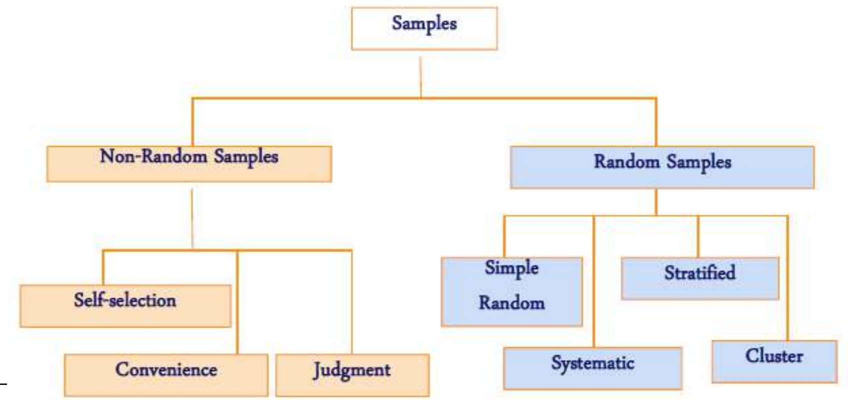
Questions)	**Question**)	4. calculate $X = Z\sigma + \mu$ Calculate $X = Z\sigma + \mu$	
Rules	Reading the table	• If exact probability is not in bdoy of table - find closest probability + determine Z value to 2 decimal places • <u>DO NOT</u> interpolate except when value is exactly HALF way EG. 0.45 is exactly halfway between 0.4495 (for Z=-1.64) & 0.4505 (for Z=-1.65) hence use Z= -1.645	
	EXAMPLE	• Weekly demand for a particular brand of perishable product =normal • Mean = 150 • Std Dev = 8. Management orders the product at the beginning of the week and throws out leftover stock at the end of the week. What would you recommend as the **min + max order quantities** for this product? Explain. Min: $\mu - 3\sigma$ $150 - 3 \times 8$ ¿126 Max: $\mu + 3\sigma$ $150 + 3 \times 8$ ¿174 The ordering policy has been to order 160 units of the brand. What is the **probability that some product be thrown out** at the end of the week? $P(Wastage) = P(Demand < 160)$  $z = \frac{X - \mu}{\sigma}$ $\frac{160 - 150}{8}$ $\frac{10}{8}$ ¿1.25 P(z<1.25) = 0.8944 (from table) = there is almost 90% chance that there will be at least some wastage	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		What should ordeirng policy be if they want to ensure that they run out of stock in only 5% of week?  Want: $P(X>?)=0.05$ $Z=1.645($¿$table)$ $Z=\frac{X-\mu}{\sigma}$ ¿ $\frac{X-150}{8}$ ¿ 1.645 $\therefore X-150$ ¿ 8×1.645 ¿ 13.16 $\therefore X=13.16+150$ ¿ 163.16 L	
Summary		• normal distribution most important distribution in statistics due to wide application in solving many practical & business related problems • probabilities of continuous random variables - calculated as areas under curve & b/w specified values of random variable • normal distribution with mean μ & std devσ can be transformed to std normal distribution involving z-scores • use z-scores allow probabilities to calculate from stdised normal distribution table	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

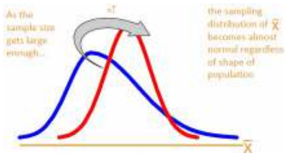
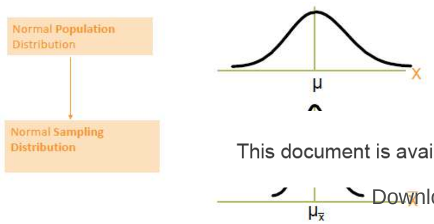
Week 7 (Topic 7) - Chapter SEVEN (Sampling & Sampling Distributions)			
5 key words in Statistics	**5 key words in Statistics**	1. Population 2. Census 3. Sample 4. Error 5. Probability ****Population** is the "goal" – our **purpose** – even though we spend most of the time working with the sample.	
	Types of Samples used	 	
	Non-Random (Non-Probability) Sampling	• **Probability** of particular element of population entering sample is unknown • some individuals / items in population have greater chance of selection than others = statistically valid statements or inferences <u>cannot</u> be made about precision of estimates	
	Advantages of non-random sampling:	• sampling costs are lower + implementation is easier • sometimes there is no alternative (random sample cannot be taken) However this means: • inferential statisical tehcniques <u>cannot</u> be applied • no statistical statement should be made re precision of result	
	Self-selection	people are invited to submmit questionnaire EG. online	
	Convenience	elements included in sample are chose b/c of accessibility or willingness EG. students are interviewed as they enter library	
	Judgment	• knowledgeable person selects sampling units that he/she feels are most representiative of population • quality of result dependent on judgement of person selecting sample • person may be well-qualified - judgement may be highly regarded = does not give results statisitical validity	
	random	• each element of population has known (non-zero) chance of beng included in sample chosen	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

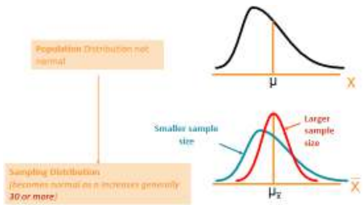
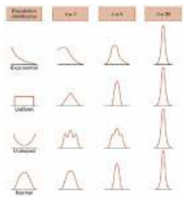
	NB: assume all samples are **random** for this unit	• should be used where possible • inferential statistics requires random samples.	
	simple random sampling	• evey individual/item from sampling fram (list of eligible elements from population) has equal chance of being selected • every sample of fixed size has same chance of selection as every other sample of that size • most elementary random sampling technique • samples normally obtained from computer random number generators	
	Systematic sampling	• decide on sample size • divid frame of N individuals into groups of K individuals: k = N/n • Randomly select 1 individual from 1st group • select every k^{th} individual thereafter • some situations it is more convenient/faster than simple random sample	
	stratified sampling	• divid population in 2+ subgroups ("strata") accoridng to common characteristics EG. gender, age groups, states etc • simple random sample is selcted from each sub group ("stratum") with sample sizes proportional to strata sizes • samples frome ach stratum are combined into one	
	cluster sampling	• population is divided into several "clusters" each rep of population • simple random sample of clusters is selected - all items in cluster can be used or items can be chosen from cluster using another poabbility sampling techniques • clusters are often based on geographical groupings where cost of accessing individual households can be reduced EG. postcodes, electorates, city blocks	
survey errors	**non- sampling error**	range of errors either for census or sample include: • response + non-response bias • interview bias • self-selection bias • measurement error • coverage error • processing error (typos)	
	sampling error	• almost certain to involve error • a (relatively small) sample is unlikely to have exactly same features as population which it was drawn • RECALL: - parameter is numerical characteristic of pop - statistic is numberical characterisitic of sample	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

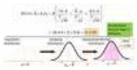
	point estimation	• from sample, value of relevant sample statistic could be used as point estimates of equivalent pop parameters $\acute{x}$ = point estimator of population **mean** (μ) s = point esitmator of population **std dev** (σ) • $\hat{p}$= point estimator of population **proporiton** (p)	
	sampling error	• diff b/w sample statistic estimate + corresponding population parameter • sampling errors are: $¿\acute{x}-\mu\vee¿$ = sample **mean** $¿s-\mu\vee¿$ = sample **std dev** $¿\acute{\hat{p}}-p\vee¿$= sample **proportion** • manage error using probability	
	Manage estimate sample error	• using random sampling combined with probability = control + preduct magnitude of error likely to occur • statistical theory based on concept known as sampling distribution of sample stat show relatively small samples (at great savings of time + money) can provide remarkably high degrees of accuracy in estimating features of population	
	sampling distribution	• distribution of all possible values of statistic for given size sample selected from population	
	Central limit theorem	 • if population distribution is normal = sample distribution of $\acute{x}$ will be normal <u>regardless of</u> size of *n* • if populatio distribution is <u>not</u> normal = sample distribution of $\acute{x}$ will be normal or approximately normal when *n* is sufficiently large enough (30+) in both cases: $\mu\acute{x}=\mu$ 	
	EXAMPLE:		

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Normal population		
	EXAMPLE: **Not** normal population		
	Distribution of sample means from various sample sizes		
	Z-formula for	• z-score is used to calculate probabilities from sampling distribution of $\acute{x}$	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	sampling distribution of $\acute{x}$	• z-formula for the sampling distribution of $\acute{x}$ $Z=\frac{\acute{X}-\mu\acute{X}}{\sigma\acute{x}}$ ¿ $\frac{\acute{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ Where: $\acute{X}$ = sample mean μ = population mean σ = population std dev n = sample size	
	Std error of sample mean	• measure of variability in the mean from sample to sample given by std error of the sample mean (std dev of all possible sample mean). • **NB**: std error of sample mean <u>decreases</u> as the sample size *n* <u>increases</u> • Std error = avg error expected to make in using sample mean as point esitmate of population mean Formula: $\sigma\acute{x}=\frac{\sigma}{\sqrt{n}}$	
	EXAMPLE	• A population has mean μ = 8 • Std Dev σ = 3 • A random sample of size n = 36 is selected. What is the probability that the sample mean is between 7.8 and 8.2? Even if the population is not normally distributed, the central limit theorem can be used as n > 30 • sampling distribution of $\acute{X}$ = approximately normal • mean $\mu\acute{x}$= 8  • std dev: $\sigma\acute{x}$ ¿ $\frac{\sigma}{\sqrt{n}}$ ¿ $\frac{3}{\sqrt{36}}$ ¿ 0.5(std error)	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

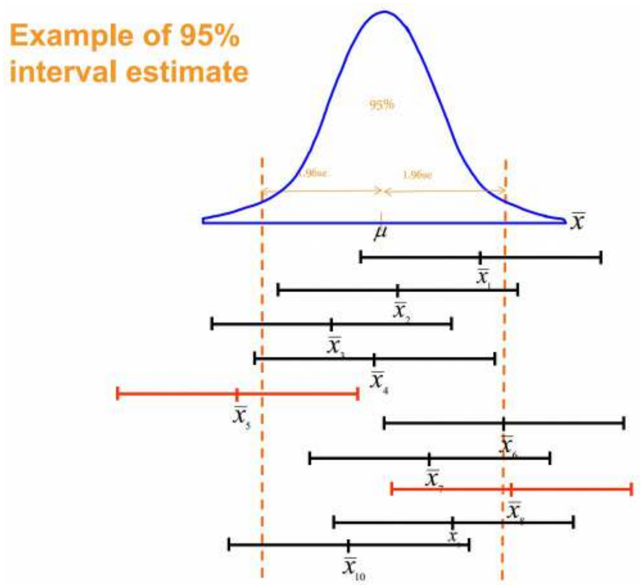
	Sampling distribution of $\hat{p}$	• is the probability distribution of all possible values of the sample proportion $\hat{p}$ for given sample size (*n*) sample proportion: $\hat{p} = \frac{x}{n}$ where: *x* = no. of items in sample with characteristics *n* = no. of items in sample sample distribution: • approx. normal = *np* > 5 & *np* > 5 • *p* = population proportion & *q* = 1 – *p* • std dev of distribution = $\sqrt{\frac{pq}{n}}$	
	Std error of sample proportion ($\hat{p}$)	• measure of variability in proportion from sample to sample given by std error of sample proportion $\sigma\,\hat{p} = \sqrt{\frac{pq}{n}}$ • NB: std error of sample proportion decreases as sample size (*n*) increases • Std error is the avg error we expect to make using sample proportion as point estimate of population proportion	
	Z-formular for sampling distribution of $\hat{p}$	• When *np* > 5 & *nq* > 5 $z = \frac{\hat{p} - p}{SE\,\hat{p}}$ ¿ $\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ Where: $\hat{p}$ = sample proportion *n* = sample size *p* = population proportion *q* = 1 – *p*	
	EXAMPLE	• If true proportion of voter who support Proposition A is p = 0.4 What is possibility that sample size of 200 yields sample proportion b/w 0.40 & 0.45? If *p* = 0.4 & *n* = 200 what is: $P(0.40 \leq \hat{p} \leq 0.45)$?	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		Find $\sigma\ \hat{p}$: $\sigma\ \hat{p}=\sqrt{\dfrac{pq}{n}}=\sqrt{\dfrac{0.4(1-0.4)}{200}}=0.03464$ Convert to standardised normal: $P(0.40\le\hat{p}\le0.45)=P\left(\dfrac{0.40-0.40}{0.03464}\le Z\le\dfrac{0.45-0.40}{0.03464}\right)=P(0\le Z\le1.44)$ Use standardised normal table: $P(0\le Z\le1.44)=0.4251$ 	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Week 8 (Topic 8) - Chapter EIGHT (Confidence Intervals)			
sampling distributions	***EXAMPLE*** male v female productivity	• sample mean for productivity: - males: $\acute{x}_M = 99.9\%$ - females: $\acute{x}_F = 97.5\%$ • are 26 males & 22 females enough to draw conclusions re **all** 3,000 employees at Conrobar? • what is mean (avg) productivity: - for <u>all</u> males? $\mu_M = ?$ - for <u>all</u> females? $\mu_F = ?$	
	Point estimation	• $\acute{x}$ = point estimator of population mean (μ) • s= point estimator of population std dev (σ) • $\hat{p}$ = point estimator of population proportion (p)	
	Why don't use point estimate	• Should <u>not</u> use value of $\acute{x}$ by itself as estimate of μ as: - Almost certain to be wrong, and - We won't know how wrong. • Point estimator <u>does</u> not provide info re how close estimate is to population parameter • We would have <u>no confidence</u> in using it – instead we use confience interval	
	Margin of error	• Interval estimate is constructed by subtracting + adding margin of error (ME) to point estimate: $Sample\,statistic \pm ME$	
	Interval estimate	• Interval estimate of population mean is: $\acute{x} \pm ME$	
			

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Confidence interval of μ (σ known¿

- Use *z* (stdarised normal) distribution
- General form of our confidence interval would be:
  - use the z (stdised normal) distribution
  - general form of confidence interval would be:

- Most commonly used confidence intervals + corresponding z values:
  - 90% of values within ± 1.645 std error of μ
  - 95% of values within ± 1.96 std error of μ
  - 98% of values within ± 2.33 std error of μ
  - 99% of values within ± 2.575std error of μ
- if you construct a 95% confidence interval, 95% is called confidence coefficient other intervals can be used (eg. 96%)

EXAMPLE:
no. of defects

- calculate 95% confidence interval estimate for mean no. of defects per shift across <u>all</u> shifts
- assume population & std dev is known:
  - σ = 10
  - $\acute{x}$ = 11.6
  - *n* = 40
  - 95% confidence = 1.96 std errors (z)

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \Rightarrow 11.6 \pm 1.96\frac{10}{\sqrt{40}} \Rightarrow 8.5 \text{ to } 14.7$$

Interpretation: we are 95% confident that the mean no. of defects per shift across all shits is somewhere in the range 8.5 to 14.7 defects.

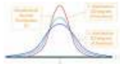
- calculate 90% & 99% confidence interval estimate for true mean no. of defects per shift

90% Interval: μ		99% Interval: μ	
- From:	8.999	- From:	7.527
- To:	14.201	- To:	15.673

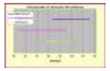
- comparing intervals to 95% CI, can see that there is trade-off b/w confidence & margin of error
  ****higher the confidence = less precise (or wider) the interval is**

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

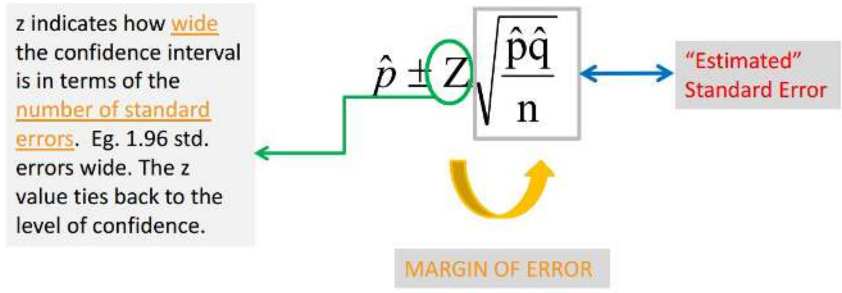
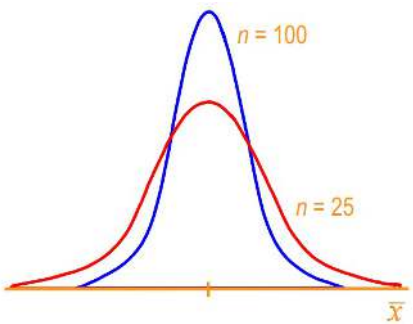
	T Distribution	$t=\frac{\acute{x}-\mu\acute{x}}{s\acute{x}}$ $z=\frac{\acute{x}-\mu_{\acute{x}}}{\sigma_{\acute{x}}}$ • Shape of distribution similar to normal distribution <u>but</u> varies re sample size	
T-distribution	Properties	• **T-scores are bigger than equivalent z-scores** • Family of distributions • T-score **depends** on degree of freedom (d.f.) • Bell shaped like normal distribution – <u>but</u> flatter + wider • T-distribute provides **more conservative confidence interval estimate**, since intervals (slightly) wider for given sample size **NB**: T-distribution should be used if σ is unknown T → Z as *n* increases 	
	Confidence interval for μ ($\sigma\ unknown$ ¿	• Use T-distribution • General form of confidence interval would be: - use t-distribution - general form of our confidence interval would be: 	
	EXAMPLE:	• Assume population std dev (σ) is unknwon (more realistic)	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	No. of Defects	• Use value of sample std dev (s) (9.77) as point estimate of σ • Use T-distribution instead of normal distribution • Calculate 95% confidence interval estimate for true mean no. of defects per shift 	
Conrobar Productivity	***EXAMPLE:*** Confidence Interval	• 95% confidence intervals for the mean productivity per group (male/female) as follows: • All males:  - Sample mean: $\acute{x}$ = 99.981% - 95% confidence interval: 98.376% to 101.585% • All females: - Sample mean: $\acute{x}$ = 97.53% - 95% confidence interval: 95.499% to 99.574% • Males: 95% confidence interval: 98.376% to 101.585% All males could be achieving 100% on average • Females: 95% confidence interval: 95.499% to 99.574% All females not achieving 100% on avg. • At 95% confidence, we <u>cannot</u> conclude males are doing better on avg than females • 98% or 99% confidence intervals are wider: same conclusions except females could now be achieving 100% avg • 90% confidence intervals are narrower: different conclusions	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Categorical variables **(assume ME either side of sample proportion)**	Confidence interval for *p*	general form of interval is:  	
	EXAMPLE: Shifts with defects	What is true proportion of shifts which defects occurred (use 95% confidence) • Of sample of 40 shifts – defects founds across 33 shifts: $\hat{p}=\frac{33}{40}=0.825$ • Sample size lrg enough for sampling distribution of $\hat{p}$ to be approximated by normal distribution • 95% confidence = 1.96 std errors (*z*) • 95% confidence interval: $\hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}} \quad =0.825\pm1.96\sqrt{\frac{0.825\times0.175}{40}}$ $=0.825\pm1.96\times0.06$ $=0.825\pm0.118$ $=0.707 \text{ to } 0.943$ • 95% confident true proportion of shifts where there are defects somewhere in range: 70.7% to 94.3%	
	Different sample sizes	• Logic tells us the bigger the sample = more accurate • If *n* is larger ∴ SE must be smaller: $SE=\frac{s}{\sqrt{n}}$ • Sample distribution would be narrower • Confidence interval ∴ be narrower  	

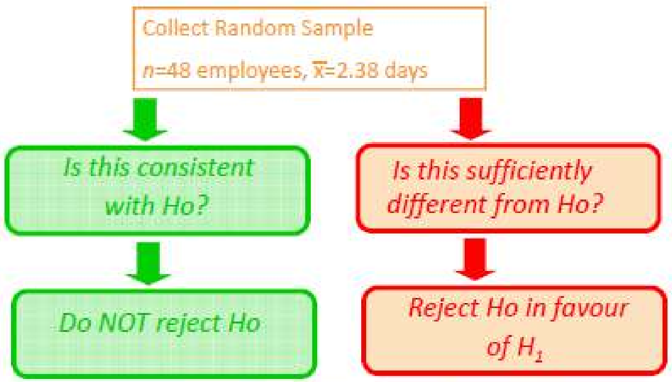
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Calculating sample size (n)	• Before taking sample to estimate μ or p – first esitmate what size sample required • To do this, specify: - Margin of error (ME) - Degree of confidence required (eg. 95%) - Obtain estiate of σ and p/q $n=\frac{z^2\sigma^2}{ME^2}$ For numerical data $n=\frac{z^2 pq}{ME^2}$ For categorical data	
	EXAMPLE: No. of defects	• If we know mean no. of defects per shift (3 defects) + 95% confidence Use σ = 10 $n=\frac{z^2\sigma^2}{ME^2}=\frac{1.96^2 10^2}{3^2}=42.7$ • N = 43: is min sample size needed for specifications	**NB:** always round up
	EXAMPLE: Shifts with defects	• if want to know true proportion of shifts in which defects occur to within 5% & 95% confidence • use previous study's $\hat{p}$ = 0.825 to approx. p $n=\frac{z^2 pq}{ME^2}=\frac{1.96^2(0.825)(0.175)}{0.05^2}=221.84$ **NB:** n = 222 is min sample size needed for our specifications	**NB:** always round up
	Calculating sample size	• if no idea re likely value of p then use p = 50% in sample size calculation • if some idea re likely value of p then: - pilot study - similar / previous study ∴ use that in calculation of n	
	EXAMPLE: Newspaper polls	• election poll is estimate with 95% confience, proportion of voters who will vote for labor party within 3% accuracy • use p = 0.5 (no better estimate of p is available) $n=\frac{z^2 pq}{ME^2}=\frac{1.96^2(0.5)(0.5)}{0.03^2}=1067.1$ **NB:** n = 1068 is min sample size needed for specifications	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Week 9 (Topic 9) - Chapter NINE (Hypothesis Testing)			
	EXAMPLE: Conrobar	• Sample of 48 employees showed: - Mean productivity = 98.9% - Mean days absent = 2.4 days • Do these mean **ALL** employees, on average, are not meeting the 100% standard for productivity and the 1.5 days target for days absent? • Need to allow for **sampling error** • Confidence intervals (last lecture) provide one approach • Hypothesis tests provide another	
2 key inferential tools	Confidence interval estimation	use if **no idea** re value of population parameter being investigated	
2 key inferential tools	Hypothesis test	Use if **some idea** re value of population parameter being investigated **or** if some hypothesised value against can **compare** sample results	
2 hypotheses	Null Hypothesis (H_0)	**assumed population parameter is <u>correct</u>**	
	Alternative hypothesis (H_1)	**assumed population parameter is <u>incorrect</u>**	
	EXAMPLE: Conrobar	• Days Absent – number of days employees are absent from work on sick/family leave • Conrobar management claim that the averageabsenteeism rate is excessively high at over 2 days per employee. • Set the null and alternative hypothesis **NB:** Remember that the analogy is: - **null** = **innocent** - **alternative** = **guilty**	
	EXAMPLE: null hypothesis	• Assume employees are <u>not</u> taking excessive leave until evidence demonstrates the contrary • assume the avg employee absenteeism rate is <2 days • Thus the null hypothesis can be written: $H_0: \mu \leq 2$ **NB**: null hypothesis must **always** contain an **equal sign**	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	EXAMPLE: alternative hypothesis	• statistical evidence is required to contradict the assumption contained in the null hypothesis • Management's suspicion or contention is that employees overall have a mean that is > 2 days. ∴ the alternative hypothesis is: $H_0 : \mu \leq 2$  Hypothesis test procedure enables us to choose which outcome	
Step 1: Setting up hypotheses	when are hypothesis tests used?	• in situations where we have: - prior knowledge - prior experience - a standard - a claim • in these situations we have: - some idea of value of population parameter being investigated, **<u>or</u>** - some hypothesised value against which we can compare sample results	
Setting up (H_0) and (H_1) prior knowledge	***EXAMPLE:*** assume prior knowledge is still current: set as H_0	Detailed survey in 2004 showed avg km travelled yearly per car was 14,500km - has there been any changes in 2015? • take random sample of cars • want to know whether the mean is diff (has increase/decreased) $H_0 : \mu$ = 14,500 no change: **avg still** 14,500km in 2015 $H_1 : \mu \neq 14{,}500$ change: **avg not** 14500km in 2015	
Setting up (H_0) and (H_1) prior experience	***EXAMPLE:*** assume prior knowledge is still reliable: set as H_0	Past experience has shown no more than 3% of components from particular supplier are defective - new batch has arrived, test random sample of components from batch • could proportion (p >3%?) $H_0 : \mu \leq 3\$$ new batch has **no more than** 3% defective $H_1 : \mu > 3\$$ new batch has **more than** 3% defective	
Setting up (	***EXAMPLE:***	Wish to test whet'	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

H_0) and (H_1) standard	normally set H_0 equal to std	• take random sample of 48 • we would write 2 hypotheses as: $H_0: \mu \geq 100\%$ employees on avg **are meeting** 100% std $H_1: \mu < 100\%$ employees on avg are **failing** to meet 100% std	
Setting up (H_0) and (H_1) claim		**Who** is making claim and the **seriousness** of claim determines whether we: • treat claim as null hypothesis (accept claim is true) until have evidence to contrary, or • treat claim as alternative hypothesis (claim is false)	
Setting up (H_0) and (H_1) claim	***EXAMPLE:***	Newspaper claims at least 40% of readers will see particular type of ad in its paper. Association of National Ads wants to check claim. • adopt conservative approach + assume claim is true • test claim with random sample of readers $H_0: \mu \geq 40\%$ at least 40% of readers ad $H_1: \mu < 40\%$ fewer than 40% see ad	
	Common mistakes	• equality part of hypotheses always appears in null hypothesis $\cancel{H_0: p > 100}$ $\cancel{H_1: p \leq 100}$ • hypotheses are statements re population parameters <u>not</u> sample statistics $\cancel{H_0: x = 100}$ $\cancel{H_1: x \neq 100}$	
Step 2: Type of test	different types of tests	• a hypothesis test re value of population mean (μ) or proportion (p) must take one of 3 forms:	
Type of test	numerical data	$H_0: \mu \geq A$ $H_0: \mu \leq A$ $H_0: \mu = A$ $H_1: \mu < A$ $H_1: \mu > A$ $H_1: \mu \neq A$	
Type of test	categorical data	$H_0: p \geq A$ $H_0: p \leq A$ $H_0: p = A$	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		$H_1: p < A$ $H_1: p > A$ $H_1: p \neq A$	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr>
<td>Type of test</td>
<td>one-tail test</td>
<td>
sometimes, alternative hypotheseis focuses on particular direction
$H_0: \mu \geq 3$
$H_1: \mu < 3$
This is a lower-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

$H_0: p \leq 8\%$
$H_1: p > 8\%$
This is an upper-tail test since the alternative hypothesis is focused on the upper tail above the proportion of 8%
</td>
<td></td>
</tr>
<tr>
<td>Type of test</td>
<td>two-tail test</td>
<td>
other times, alternative hypotheses focuses on either direction
$H_0: \mu = 25$
$H_1: \mu \neq 25$
This is a two-tail test since the alternative hypothesis is focused on either values below the mean of 25 or values above 25
</td>
<td></td>
</tr>
<tr>
<td>Step 3: Level of significant & critical region</td>
<td>errors in hypothesis tests</td>
<td>

correct outcomes:

H_0 is true and we do not reject H_0
H_0 is false and we reject H_0

incorrect outcomes (errors):

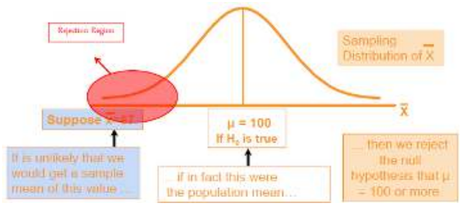
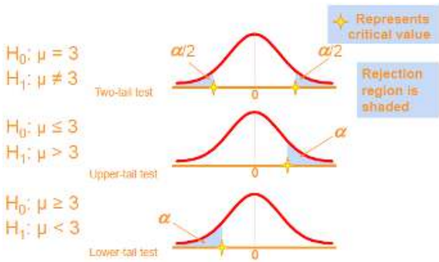
Type I Error: reject H_0 when in fact H_0 is true.
Type II Error: not reject H_0 when in fact H_0 is false

<table>
<tr><td></td><td colspan="2">Population Condition</td></tr>
<tr><td>Conclusion</td><td>H_0 True</td><td>H_0 False</td></tr>
<tr><td>Not Reject H_0</td><td>Correct Decision</td><td>Type II Error</td></tr>
<tr><td>Reject H_0</td><td>Type I Error</td><td>Correct Decision</td></tr>
</table>
</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type I error = more serious</td>
<td>

try to set up H_0 and H_1 to control Type I error.
This is also a conservative approach to decision making
Only conclude the alternative is true if have very strong evidence.

</td>
<td></td>
</tr>
</table>

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

<table>
<tr><td></td><td></td><td>∴ to set up H_0 and H_1 need to think about:
- Which error is the more serious (Type I)
- The risk willing to run of making this error</td><td></td></tr>
<tr><td></td><td>Level of significance (α)</td><td>• this is what hypothesis testing risk is referred to
• specify maximum risk willing to accept in making Type I error
• normally set at 5%, 10%, 2% or 1%
• level of significance is max risk willing to accept in making Type I error (rejecting H_0 when shouldn't)</td><td></td></tr>
<tr><td>critical rejection region</td><td>critical rejection region</td><td>• consider if reject H_0
$H_0: p > 100$
$H_1: p \leq 100$

</td><td></td></tr>
<tr><td>critical rejection region</td><td></td><td>• Need to decide:
- direction of test; and
- specify(α)
• Can next determine the critical (rejection) region = area in sampling distribution where we will reject H_0.
• critical value(s) = border(s) of the critical region</td><td></td></tr>
<tr><td></td><td>level of significance
&
critical (rejection) region</td><td>
</td><td></td></tr>
</table>

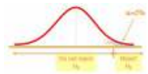
This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	critical value (CV)	• can be determined in terms of Z or t-values EG. suppose α = 5% (2-tail test) using Z  	
Step 4: Decision rule	Decision rule	• if sample result falls beyond critical value = reject H_0 in favour of H_1 • if calculate Z statistic from sample which is: Z > +1.96 or Z < -1.96 ∴ would reject H_0	
Step 5: analyse sample	analyse sample	• collect random sample • calculate sample statistic $\acute{x}$, s for hypothesis tests re means $\hat{p}$ for hypothesis tests re proportions • calculate Z or t-statistic as appropriate	
Step 6: conclusion	Conclusions	• compare calculated test statistic (from step 5) to decision rule (from step 4) • make decision: reject H_0 do not reject H_1 • write conclusion re terms of problem solved	
	EXAMPLE: Defectives	Past experience show that no more than 3% of components from particular supplier are defective. New batch arrives - does this new batch of components have no more than 3% defective? If it has more, we will send it back to the supplier. 1. Take a random sample of components from batch to determine proportion of defective components. 2. Then perform a hypothesis test to see if the population proportion in the batch could be greater than 3%. **Step 1:** step up H_0 and H_1	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

- $H_0: p \leq 3\%$ = new batch has **no more than** 3% defective
- $H_1: p > 3\%$ = new batch has **more than** 3% defective

Step 2: decide on direction of test



- upper tail test

Step 3: Decision on α (lvl of significance)
- set α to be 5%
- critical value (CV) of Z will be +1.645

Step 4: decision rule (using critical values of Z or t)
- if proportion of defecgives in sample result in Z static > 1.645 = reject H_0

Step 5: sample (peform relevant calculations: Z or t-statistic)
- if test on random sample of 1000 components + 43 defective ...:
  - *n* = 1000
  - $\hat{p} = \frac{43}{1000}$ = 0.043 or 4.3%

$$\sigma \hat{p} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.03(0.97)}{1000}} = 0.00539 \vee 0.539\%$$

- find test statistic (z statistic)

$$z = \frac{\hat{p} - p}{\sigma \hat{p}} = \frac{0.043 - 0.03}{0.00539} = 2.41$$

- test statistic = no. of std errors sample result from population parameter (assuming H_0 is true)
- compare Z statistic to CV of Z

Step 6: conclusion (reject / not reject H_0+ answer question)
- Z statistic of 2.41 **exceeds** CV of Z (1.645) = reject H_0
  (means sample proportion (evidence) fell in rejection region)
- conclude at 5% level of significance - new batch has more than 3% defective
- decision: send batch back

EXAMPLE:
Tests on productivity at Conrobar

Are **all** staff on avg meeting 100% productivity std or better? (Use α = 5%)

Step 1: step up H_0 and H_1
- $H_0: \mu \geq 100\%$ = staff **meeting** std
- $H_1: \mu < 100\%$ = staff **not meeting** std

Step 2: decide on

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

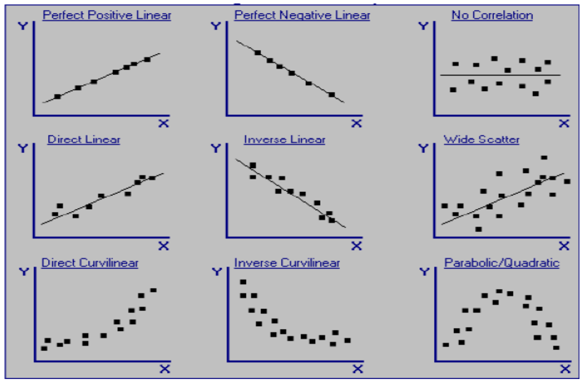
		• lower tail test **Step 3:** Decision on α (lvl of significance) • α = 5% • CV for t = -1.678 (from table) **Step 4:** decision rule (using critical values of Z or t) • if sample results t static < 1.678 = reject H_0 **Step 5:** sample (peform relevant calculations: Z or t-statistic) • n =48 • sample mean: $\acute{x}$ = 98.86 • sample SD: s = 4.399 • Std error: $s\acute{x}=\frac{s}{\sqrt{n}}=\frac{4.399}{\sqrt{48}}=0.635$ • t-statistic:$t\acute{x}=\frac{\acute{x}-\mu}{s\acute{x}}=\frac{98.86-100}{0.635}=1.795$ **Step 6:** conclusion (reject / not reject H_0+ answer question) • t-statistic of -1.795 **less than** CV of t (-1.678) = reject H_0 • sufficient evidence to conclude, at 5% lvl of significance, staff overal on avg **not** meeting productivity std of 100%	
Different approach to hypothesis test	P-value	the probability of getting test statistic more extreme than sample result, given null hypothesis (H_0) is true • derived from (Z or t) test statistic • gives area in tail (or 2 tail if 2-tail test) beyond where sample result falls • if p-value > α = **do not** reject H_0 • if p-value < α = **reject** H_0 • p-value <u>calculated by computer</u>	
	EXAMPLES: Defective	**Step 5** • p-value = 0.0080 or 0.8% **Step 6** • p-value of 0.008 (or 0.8%) = < α of 0.05 (or 5%) ∴ reject H_0 • Decision: the new batch of components has >3% defective	
	EXAMPLES: Conrobar	**Step 5** • p-value = 0.0395 or 3.95%	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		Step 6 • p-value of 3.95% = < α of 5% ∴ reject v • Decision: staff overall, on avg, are **not meeting** productivity std of 100%.	
	Tests on productivity at Conrobar	**Step 6**: Conclusion at differentα's • p-value = 0.0395 or 3.95%, tells us: - would **reject** H_0 at an α of 10% (or < 3.95%) - would **not reject** H_0 at anα of 2% or 1% (or >3.95%)	

This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

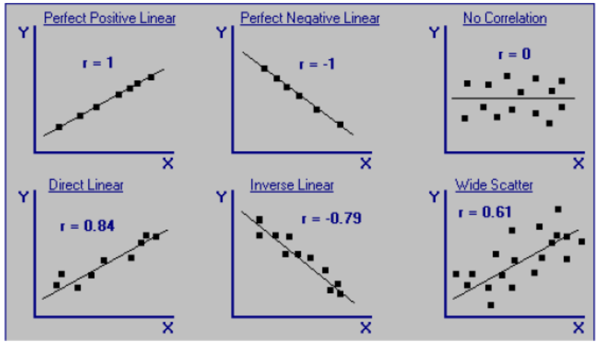
Week 10 (Topic 10) - Chapter TEN (Simple Linear Regression)			
	Regression & Correlation	• relationships b/w variables enable us to explain nature of link b/w variables + how one variable affects the other • build a model b/w variables for purpose of: - **explanation** - **prediction** - **control** - **causality**	
Regression	Explanation	• regression helps explain / understand variation in dependent variable • do this by: **finding other independent variables that relate to dependent variable** • wish to know: - **direction** of the relationship - **strength** of the relationship •	
	Prediction	• make use of explanatory (independent) variable to predict likely outcome of dependent variable EG. knowing no. of customers fast food restaurant has may enable management to forecast sales	
	Control	• if we have some control over value of independent variable, this enables some form of control over dependent variable EG. varying advert expenditure up/down to certain extent = able to control movement in sales	
	Causality	• it is important to note that while regression models may establish association b/w variables - they do not necessarily establish causality	
Concepts in regression & Correlation	Scatter diagram	representation of possible relo b/w 2 variables • plots pairs of variables on scatter diagram to identify possible relationships • verticle (y) axis) always contains dependent variable • look for: - no relationships - linear relationships - non-linear relationships • possible patterns 	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

Correlation coefficient (r)

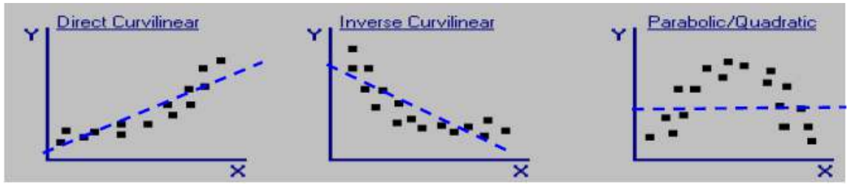
measure strength + direction or linear relationship b/w 2 variables

- $-1 \leq r \leq 1$
- r is relative measure of strength of linear relationship
- closer r is to +1 or -1 = stronger linear relo
- r close to 0 = weaker / no linear relo
- sign of r provides direction of relo



non-linear relationship

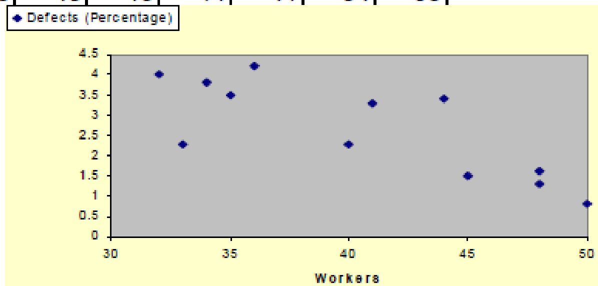
- before interpreting r, scatter plot **must** be drawn
- r associated with regression models would be poor indicators of actual strengths of each



EXAMPLE:
% defects v. no. of workers

random sample of 12 shifts at manufacturer sampled and % of defective components recorded

Defects	4	0.8	1.5	4.2	2.3	2.3	1.3	1.6	3.4	3.3	3.8	3.5
Workers	32	50	45	36	40	33	48	48	44	41	34	35



are defects & workers related?

- negative linear relo

correlation coefficient: r = -0.788

- a moderation to strong negative linear relo b/ w defects & workers

Regression: line of best fit

- estimated simple linear regression linear:

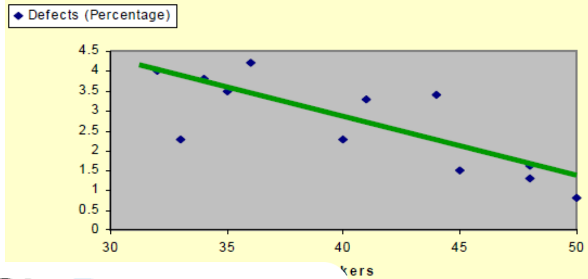
$\hat{y} = b_0 + b_0 x$

$\hat{y}$ = dependent variable

x = independent variable

b_0= y-intercept (line cuts vertical axis)

b_1 = slope of li

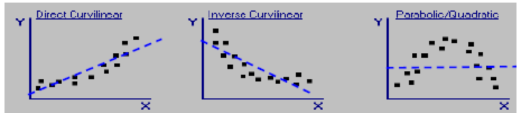
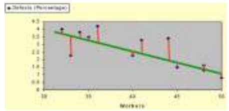


This document is available free of charge on StuDocu.com

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		Regression Coefficients - We explain using the defectives example: $\hat{y} = 8.480 - 0.144\,x$ - regression coefficients b_0 & $b_1$can be interpreted in **3** ways: ▪ **Geometrically** (i.e. graphically) b_0 interpretation**:** on graph b_0 = where line cuts vertical axis EG. line cuts y axis at 8.48% b_{10} interpretation**:** on graph b_1 = slope of line EG. slop = -0.144 ▪ **Algebraically** (i.e. in equation form) b_0 interpretation**:** b_0 is value of y when x = 0 EG. y = 8.48 when x = 0 workers b_1 interpretation**:** b_1 is change in value of y when x changes by 1 EG. if x increases by 1 = y deceases by 0.144% ▪ **Practically** (i.e. practical interpretation) b_0 interpretation**:** b_0 **not always useful interpretation** as x = 0 may be outside range of x values used for regression equation EG. 8.48% of defects produced on avg when 0 workers = nonsensical b_1 interpretation**:** b_1 **indicates impact on y from change in x** EG. for each extra worker employed on a shift, on avg defectives decrease by 0.14%	
Concepts in regression & Correlation	Regression	mathematical model of relationship	
	Simple linear regression	• involves **1 independent** variable	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)

	Multiple regression	• involves **1+ independent** variable to explain variation in dependent variable	
	Non-linear relationship	• before interpreting r, a scatter plot must be drawn • r associated with regression models = poor indicators of actual strengths of each relo 	
how does model fit data?	Residuals	• regression line does not perfectly fit data • there will be **variations (errors) b/w line + actual data points**: $y-\hat{y}$ 	
how does model fit data?		• need to obtain measures of these residuals & how well line fits with data • **measure variation around line** = use std error of estimate (s_e) • for **how well line fits data** = use coefficient of determination (r^2)	
	Std error of estimates (s_e)	• use to measure variation around line • defectives v employees: s_e = 0.758 • interpretation: estimate avg variation around regression line = 0.758% • rough approx.. using empirical rule, could say max deviation from line be: $\pm(3\times 0.758)$ or $\pm 2.27\%$	
	Coefficient of determination (r^2)	• how well line fits data: $0\le r^2\le 1$ • measures proportion of variation in 1 variable (y) dependent, explained by or attributable to variation in 2nd variable (x) independent • provides absolute measure of strength of relo • r^2 close to 1 = strong relo • r^2 close to 0 = weak / non-existent relo • normally expressed as %	
	coefficient determination (r^2)	• Defective v employees: r^2 = 0.6209 • interpretation: - approx.. 62% of variation in "defectives" explained by or attributed to variation in "no. of workers" in shift	

This document is available free of charge on StuDocu.com
Downloaded by Xia Jianfeng (jackxia5@gmail.com)

		- remaining 38% variation = result of other factors not included in model (eg. time of shift, worker's experience etc)	
	Use regression equation for estimation / prediction	• regression equation coefficient (b_0 and b_1) define nature of relationship b/w variations • regression equation also used for estimation / prediction EG. if 50 workers in shift, estimated proportion of defectives would be: $\hat{y} = b_0 + b_1 x$ This is a point estimate $\hat{y}$ = 8.48 - 0.144 * (50) = 1.28%	
	Prediction	when we use regression model with value of x contained **in range** of x values from sample	
	extrapolation	when we use the model with value of x **outside range** • should be used with caution as no guarantee same model holds outside original range of data	
	EXAMPLE:	Collected data over the last 10 years re annual expenditure on advertising & its total sales (all figures scaled for inflation).  Develop a regression model and answer the following questions: • How well does the model predict sales? • Interpret b0 and b1. • What would you estimate sales to be when $1m is spent on advertising?	

Downloaded by Xia Jianfeng (jackxia5@gmail.com)