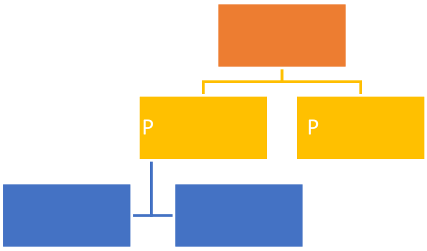
Edexcel GCSE Statistics (9-1) Revision Notes

Chapter 1: Collection of Data

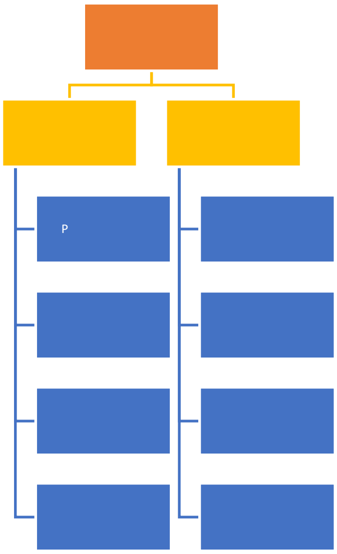
- **Raw Data**
- **Qualitative**
- **Quantitative**
- **Discrete**
- **Continuous**
- **Categorical**
- **Ordinal (rank)**
- **Bivariate**
- **Multivariate**



-
-
-
  - ○
  - ○
-
  - ○ K
  - ○

- **Primary**
- **Secondary**

	Advantages	Disadvantages
Primary Data	• Accurate • Collection method known • Can find answers to specific questions	• Time consuming • Expensive
Secondary Data	• Cheap • Easy • Quick • Data from some organisations can be more reliable than data collected yourself	• Method of collection unknown • Data may be out of date • May contain mistakes • May come from unreliable source • May be difficult to find answers to specific questions



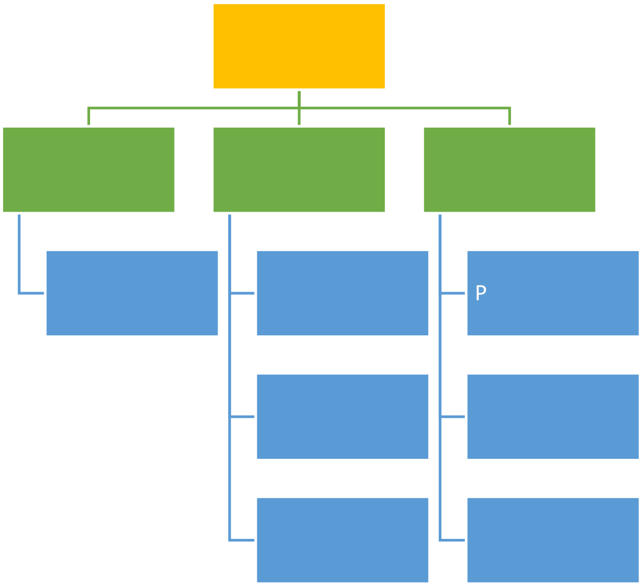
- **Population**

- **Census**
- **Sample**

	Advantages	Disadvantages
Census	• Unbiased • Accurate • Takes into account entire population	• Time consuming • Expensive • Lots of data to manage • Difficult to ensure whole population is used
Sample	• Cheaper • Quicker • Less data to consider	• May be biased • Not completely representative

- **Sampling Frame**

- **Sampling Unit**
- **Biased sample**



- **Random Sample** **equal chance**

- Stratified Sample proportion
    - $$stratified\ sample = \frac{strata}{total} \times sample\ size$$

      each group
- Systematic Sampling intervals
- Cluster Sampling
- Quota Sampling
    - P

- **Opportunity Sampling**
    -
        - P
        -
        -
    -
        -

- **Judgement Sampling**
    -
        -
        - P
    -
        -
        - P

$$\frac{M}{N} = \frac{m}{n} \qquad N = \frac{Mn}{m}$$

N is the population size to be estimated.
M is the number of members of the population that are captured initially and tagged.
n is the number of members of the population that are captured subsequently.
m is the number of members of this subsequent captured population that are tagged.

$$\frac{\textit{First Capture}}{\textit{Total } (N)} = \frac{\textit{Tagged}}{\textit{Second Capture}}$$

Method

they are thoroughly mixed

Assumptions

-
-
-
-

-
    - **Explanatory (Independent) Variable** –
    - **Response (dependent) variable** –
    - **Extraneous Variables** –

- **Laboratory Experiments** full control
  - ○
    - ▪
    - ▪
    - ▪
  - ○
    - ▪
    - ▪
  - ○
    - ▪
- **Field Experiments** some control
  - ○
    - ▪
    - ▪
    - ▪
  - ○
    - ▪
  - ○
    - ▪
    - ▪
- **Natural Experiments** no/very little control
  - ○
    - ▪
    - ▪
    - ▪ P
  - ○
    - ▪
  - ○
    - ▪ K
    - ▪
    - ▪

Steps

Example

P

Questionnaire –

P

Features of a good questionnaire:

-
-
-
- P
-
-
- P
- P

Problems with Questionnaires:

-
    -
    -
    -
    -
-

Random Response Method:

Pilot Study

-
-
-
-
-

Interviews:

	Advantages	Disadvantages
Interview	• Interviewer can explain questions • Interviewer can put people at ease when having to answer personal qs • Respondents can explain their answers • High response rate	• Less likely to answer personal questions and may be less honest • Time consuming • Expensive • Smaller sample size than questionnaire • Interviewer bias – interviewer may interpret answers to suit their opinion • Respondent may try to impress/guess the answer the interviewer wants.
Anonymous Questionnaire	• Respondents more likely to answer personal questions • No interviewer bias • Easy to send questionnaires to large sample size • Quick • Cheap	• Some questions may not be understood • Researchers may not understand some of the responses • Low response rate

Outliers -

Cleaning Data –

-
-
-
-

- **Control Groups**
  - ○
  - ○
  - ○

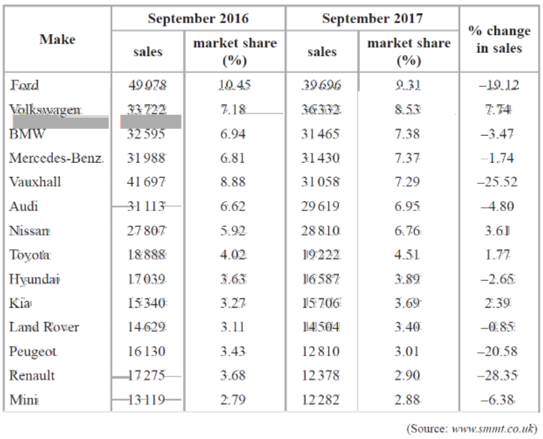
- **Matched pairs**

Hypothesis -

-
-
-
-
-

Chapter 2 – Processing and Representing Data

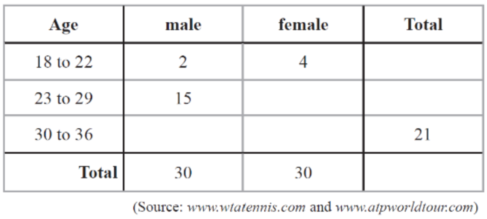
Databases –



Make	September 2016		September 2017		% change in sales
	sales	market share (%)	sales	market share (%)	
Ford	49 078	10.45	39 696	9.31	–19.12
Volkswagen	33 722	7.18	36 332	8.53	7.74
BMW	32 595	6.94	31 465	7.38	–3.47
Mercedes-Benz	31 988	6.81	31 430	7.37	–1.74
Vauxhall	41 697	8.88	31 058	7.29	–25.52
Audi	31 113	6.62	29 619	6.95	–4.80
Nissan	27 807	5.92	28 810	6.76	3.61
Toyota	18 888	4.02	19 222	4.51	1.77
Hyundai	17 039	3.63	16 587	3.89	–2.65
Kia	15 340	3.27	15 706	3.69	2.39
Land Rover	14 629	3.11	14 504	3.40	–0.85
Peugeot	16 130	3.43	12 810	3.01	–20.58
Renault	17 275	3.68	12 378	2.90	–28.35
Mini	13 119	2.79	12 282	2.88	–6.38

(Source: *www.smmt.co.uk*)

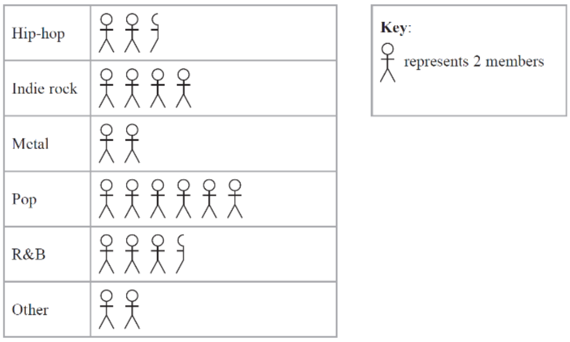
Two-Way Tables



Age	male	female	Total
18 to 22	2	4	
23 to 29	15		
30 to 36			21
Total	30	30	

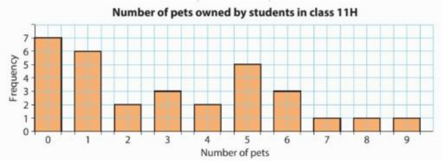
(Source: *www.wtatennis.com* and *www.atpworldtour.com*)

-
-



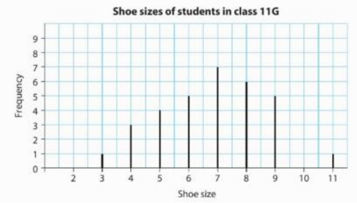
-
-

- **Simple Bar Charts**

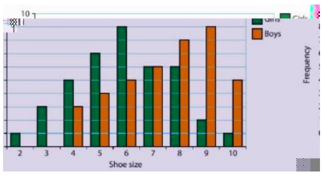


  - ○
  - ○
  - ○

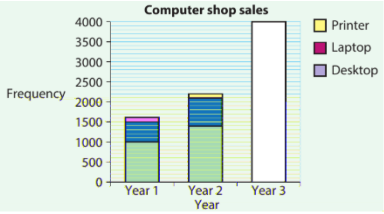
- **Vertical Line Graph**



- **Multiple Bar Charts**



- **Composite Bar Charts**



K

A good way of organising data without losing any of the detail

K

How to draw one:

first digits **numerical order**

correct row.

numerical order

key

- **Back-to-back Stem and Leaf Diagrams**
  - ○
  - ○
  - ○

Area of Pie Chart = Total Frequency

1.
2.
3.
4.
5.
6. K

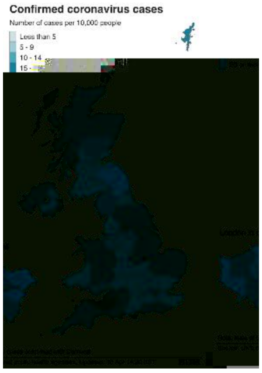
Comparative Pie Charts

Area of Pie Chart = Total Frequency

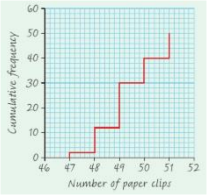
$$r_2 = r_1 \frac{\sqrt{F_2}}{\sqrt{F_1}}$$

-
-

-
-
-

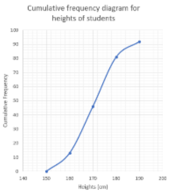

Confirmed coronavirus cases
Number of cases per 10,000 people
Less than 5
5 - 9
10 - 14
15 -

K



CF Step Polygons **discrete**

ch

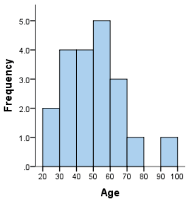


CF Curves **grouped continuous**

ch

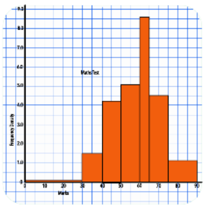
- P

Equal Class Widths



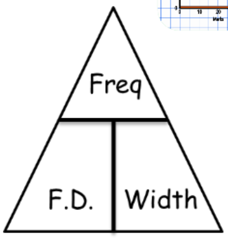
K

Unequal Class Widths



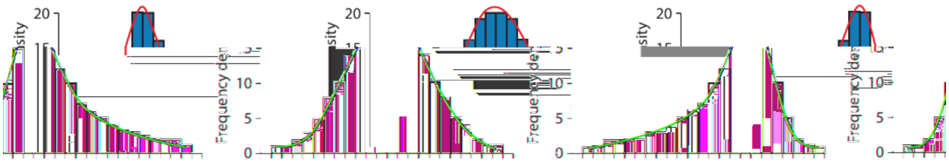
$$Frequency\ Density = \frac{Frequency}{Class\ Width}$$

$$Frequency\ Density \times Class\ Width = Frequency$$



Drawing Histograms:

Estimating frequencies from histograms:



tion has positive skew.
data values are at the
xample: The age at
on learns to write.

tion is stretched out in
direction →.

This distribution is symmetrical.
It has no skew. Example:
The lengths of leaves on a tree.

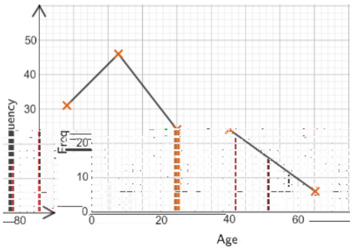
This distribution has negative skew.
Most of the data values are at the
upper end. Example: The age at
which a person dies.

The distribution is stretched out in
the negative direction ←.

This distribu
Most of the c
lower end. Ex
which a pers

The distribut
the positive

-
-
-



Types of Misleading Diagrams:

-
-
-

- K

Axes and Scales that can be misleading:

-
-
-
-
-

Chapter 3 – Summarising Data

Averages

____

most

____

middle

Discrete Data:

median is the $\frac{1}{2}(n+1)th$ value

$\frac{1}{2}(n+1)th$

Grouped Data:

½nth value

Estimate Median using Linear Interpolation:

____

____

____

_______ ___________________

_________ ____________________

____ _______________________________

sum of all the values divided by the number of values

Discrete Data:

Formula for Mean: $\boldsymbol{\overline{x} = \frac{\sum x}{n}}$

$\bar{x}$

x

n

$\sum x$

Frequency Table (not grouped):

$f \times x$

$\sum fx$

$\sum f$

Formula: $\boldsymbol{\frac{\sum fx}{\sum f}}$,

$\sum fx$

$\sum f$

Frequency Table (grouped):

$f \times midpoint$

$f \times midpoint$

$f \times midpoint$

$\sum fx$

$\sum f$

Formula: $\boldsymbol{\frac{\sum(f \times midpoint)}{\sum f}}$

Weighted Mean

different number of values or weights in each group

$$\boldsymbol{Weighted\ Mean} = \frac{\sum(\boldsymbol{weight\ x\ value})}{\sum \boldsymbol{weights}}$$

Geometric Mean

The nth root of the product of all the values

$$\boldsymbol{Geometric\ Mean} = \sqrt[n]{\boldsymbol{value}_1 \times \boldsymbol{value}_2 \times \ldots \times \boldsymbol{value}_n}$$

Linear Transformation

Example

Mode –

Median – ____ ________ _________

____ ________ __________

________ ________ __________

________ ________ _________

___________________ ________ _____ ________

____

Mean – ____ ________ __________

__________ ____ __________

____ ____ __________

__________ ________ __________

	Advantages	Disadvantages
Mode	• • • •	• • •
Median	• • • • P	• •
Mean	• •	• • •

Measures of Dispersion

spread

$$Range = Largest\ Value - Smallest\ Value$$

P

"Between Quartiles"

$$Interquartile\ Range = Upper\ Quartile - Lower\ Quartile$$

K P KP KP
P P P

Discrete Data

KP
P

KP
KP

P P KP

Grouped Data

KP
P

KP P

KP P
P P KP

)

Percentiles

Frequency Table (not grouped)

Formulae: $\sigma = \sqrt{\frac{\sum f(x-\overline{x})^2}{\sum f}}$ **OR** $\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - (\frac{\sum fx}{\sum f})^2}$

$$\sum f = n$$

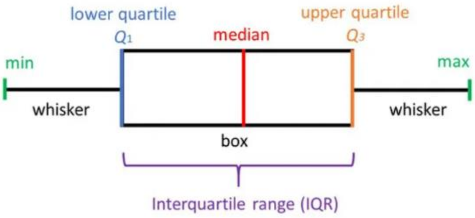
$$\frac{\sum fx}{\sum f} = mean$$

$$\boldsymbol{x - \overline{x}}$$

Grouped

K P KP

P P



P

Drawing Box Plots:

KP P

Outliers

far from the rest of your data

distort the data

P P KP

$$Outliers\ are\ values > UQ + (1.5 \times IQR)$$
$$or < LQ - (1.5 \times IQR)$$

P

P

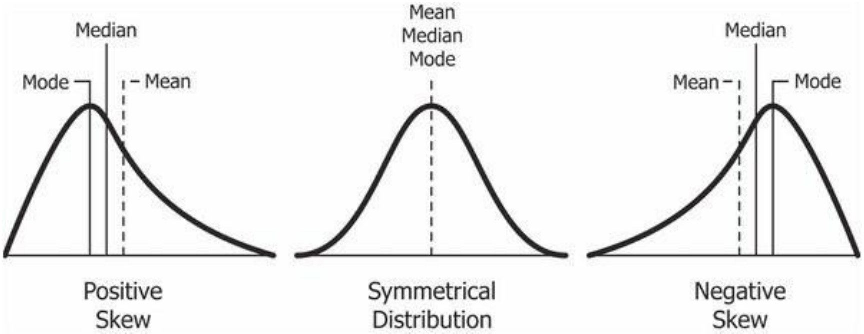
KP P

$$Outliers = Values\ outside\ \bar{x} \pm 3\sigma$$

Interpreting box plots

P

Types of Skew:

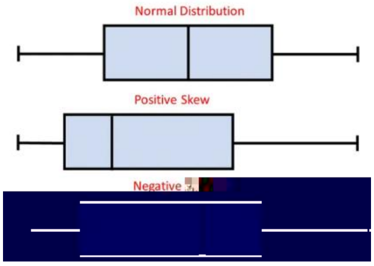


- ______________

- _______________

- ____________

Skewness on Box Plots:



- _________________________________
  KP P
- _____________ KP
- ______________ P

Skewness using the Formula:

Formula: $$Skewness = \frac{3(mean - median)}{standard\ deviation}$$

- ______________
- _______________
- __________

a measure of average (mean/median/mode) and spread (range/IQR/SD)

Example Comparisons and Interpretations of Data

- ____________________

- ________________
  P

P

ch ch ch ch ch ch ch ch ch ch chch ch

- _______________

	P

Chapter 4 – Scatter Diagrams and Correlation

bivariate

Explanatory variable

Response Variable

- **Positive Correlation**
- **Negative Correlation**
- **Zero Correlation**

- **Linear Correlation**
- **Non-Linear Correlation**

Causation

P

K K

K

$$\boldsymbol{Mean\ Point\ (\bar{x}, \bar{y}) = (Mean\ of\ x\ values, Mean\ of\ y\ values)}$$

K

Interpolation K **within the range of data**

K

K

Extrapolation K **outside of the range of values**

K

K

K K

$$\boldsymbol{Eqn\ of\ LOBF:\quad y = ax + b}$$

Drawing Regression Line:

K

Finding Equation of LOBF/Regression Line:

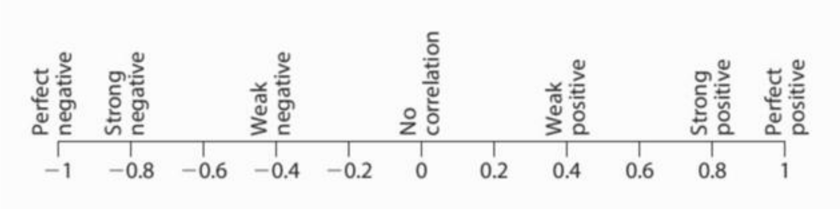
(x_1, y_1) (x_2, y_2)

$$\boldsymbol{a = \frac{y_2 - y_1}{x_2 - x_1}}$$

K

$b = y_1 - ax_1$

$\boldsymbol{y = ax + b}$



-
-
-

$$SRCC, r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Calculating SRCC:

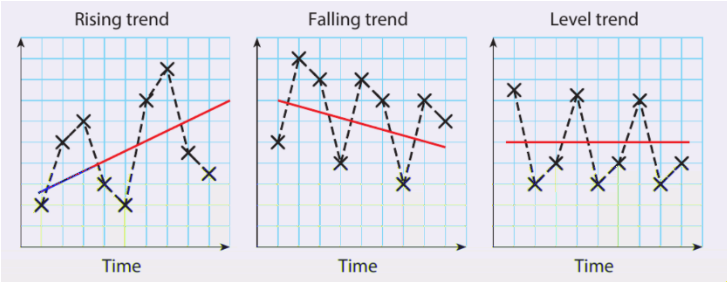
-
-
-

Chapter 5 - Time Series

time plotted on the x-axis

K

general trend



K

Term	Autumn 2000	Spring 2001	Summer 2001	Autumn 2001	Spring 2002	Summer 2002
Number of people	520	300	380	640	540	500

K

$$\mathbf{Seasonal\ Variation = Actual\ Value - Trend\ Value}$$

Estimated Mean Seasonal Variation (EMSV)

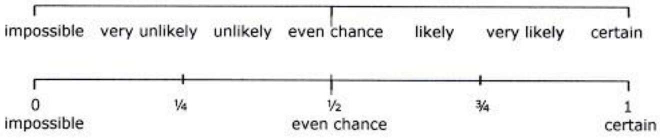
$$\begin{aligned}&\mathbf{Estimated\ Mean\ Seasonal\ Variation}\\&\quad= \mathbf{Mean\ of\ all\ the\ seasonal\ variations\ for\ that\ season}\end{aligned}$$

Predicting Values

$$\mathbf{Predicted\ Value = Trend\ Line\ Value\ (from\ graph) + EMSV}$$

Chapter 6 – Probability

how likely



outcome

event

$$P(event) = \frac{Number\ of\ successful\ outcomes}{Total\ number\ of\ outcomes}$$

The probabilities of all outcomes add up to 1.

expected frequency

$$Expected\ Frequency\ of\ Event\ A = P(A) \times number\ of\ trials$$

Trial

$$Estimated\ Probability = \frac{Number\ of\ trials\ with\ successful\ outcomes}{Total\ number\ of\ trials}$$

Estimated Probability is also called Relative Frequency.

Probability negative events

For Bias:

$$Risk = \frac{Number\ of\ trials\ in\ which\ event\ happens}{Total\ number\ of\ trials}$$

2 types of risk:

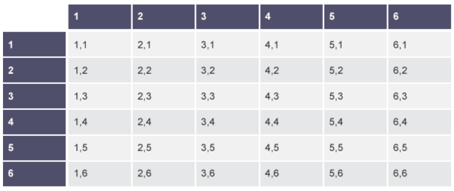
Absolute Risk

Relative Risk

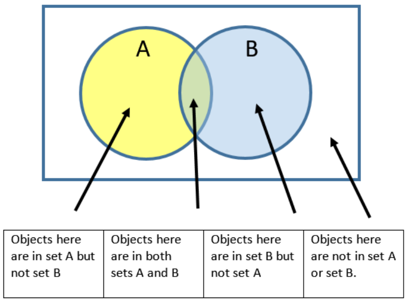
$$Relative\ Risk = \frac{Risk\ for\ those\ in\ the\ group}{Risk\ for\ those\ not\ in\ the\ group}$$

Sample Space list of all the possible outcomes

Sample Space Diagram table outcomes of two events



	1	2	3	4	5	6
1	1,1	2,1	3,1	4,1	5,1	6,1
2	1,2	2,2	3,2	4,2	5,2	6,2
3	1,3	2,3	3,3	4,3	5,3	6,3
4	1,4	2,4	3,4	4,4	5,4	6,4
5	1,5	2,5	3,5	4,5	5,5	6,5
6	1,6	2,6	3,6	4,6	5,6	6,6


A
B
Objects here are in set A but not set B
Objects here are in both sets A and B
Objects here are in set B but not set A
Objects here are not in set A or set B.

Completing Venn Diagrams:

Mutually Exclusive Events CANNOT happen at the same time

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

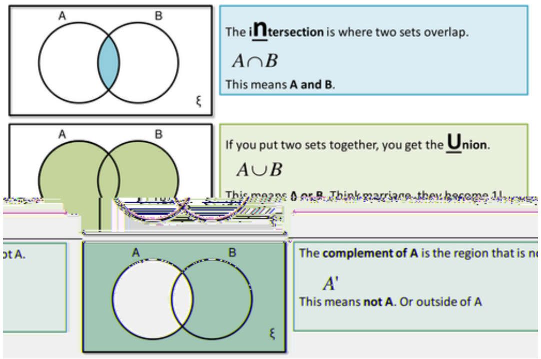
Exhaustive Events contains ALL the possible outcomes

$$P(A) + P(not\ A) = 1$$
$$P(not\ A) = 1 - P(A)$$

K

General Addition Law.

not mutually exclusive



The **i<u>n</u>tersection** is where two sets overlap.

$A \cap B$

This means **A and B**.

If you put two sets together, you get the **<u>U</u>nion**.

$A \cup B$

This means **A or B**. Think marriage, they become 1!

ot A.

The **complement of A** is the region that is n

A'

This means **not A**. Or outside of A

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

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

$P(A \cap B)$
$P(A \cup B)$

Unconnected Events

Multiplication Law for Independent Events:

$$P(A\ and\ B) = p(A) \times P(B)$$

$$P(A\ and\ B\ and\ C) = P(A) \times P(B) \times P(C)$$

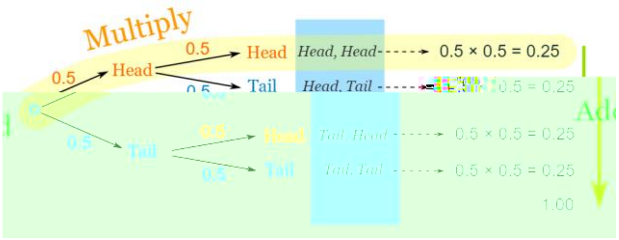
$$P(at\ least\ 1) = 1 - P(none)$$

multiply along the branches

Add probabilities down columns

Replacement

Without replacement



When one event affects the chances of another event happening

Notation:

$P(B|A) = P(B \text{ given that } A \text{ happens})$

How to know it is conditional probability?

'given that' 'if'

from that' 'this'

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

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

For two independent events, A and B P(A) = P(A|B).

Chapter 7 - Index Numbers

Simple Index numbers

$$\text{Index Number} = \frac{\text{Price}}{\text{Base Year Price}} \times 100$$

•

•

Retail Price Index (RPI

Consumer Price Index (CPI)

J

J

Gross Domestic Product (GDP)

Weighted Index Numbers

$$\text{Weighted Index Number} = \frac{\sum(\text{index number } \times \text{weight})}{\sum \text{weights}}$$

$$Chain\ Base\ Index\ Numbers = \frac{price}{last\ year's\ price} \times 100$$

Crude Rate

Crude Birth Rate

Crude Death Rate

$$Crude\ Rate = \frac{number\ of\ births/deaths}{total\ population} \times 1000$$

Standard Populations

$$Standard\ Popualtion = \frac{number\ in\ age\ group}{total\ population} \times 1000$$

Standardised Rate

$$Standardised\ Rate = \frac{Crude\ Rate}{1000} \times Standard\ Population$$

Chapter 8 - Probability Distributions

with

Notation **B (n, p)**

Conditions for Binomial Distribution:

Finding Probabilities using the Binomial Distribution: Use $(p+q)^n$ to find the probabilities

$(p+q)^n$ K

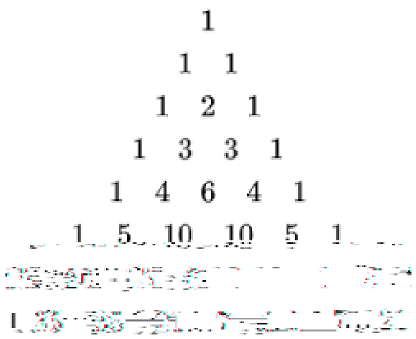
$$10 \times (\frac{1}{6})^3 \times (\frac{5}{6})^2$$

Finding the Probabilities/Coefficients:

$$(p+q)^n$$

$$(p+q)^4$$

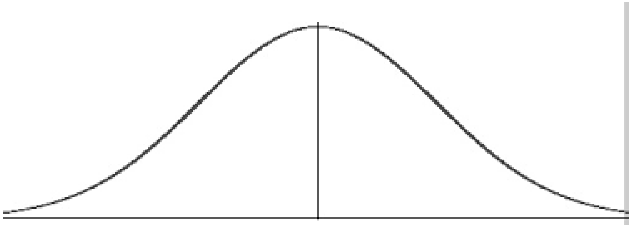
$$1p^4 + 4p^3q^2 + 6p^2q^2 + 4p^1q^3 + 1q^4$$



1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

The mean (or expected value) **B *(n, p)* is *np*.**

smooth, bell-shaped curve.

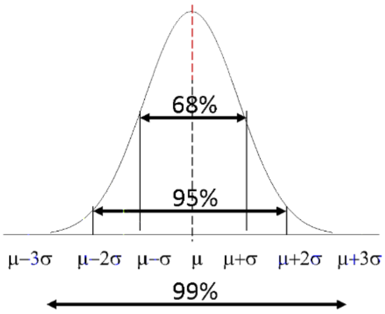


Notation: $N\ (\boldsymbol{\mu},\ \boldsymbol{\sigma}^2)$ μ σ^2 σ

Conditions for Normal Distribution:

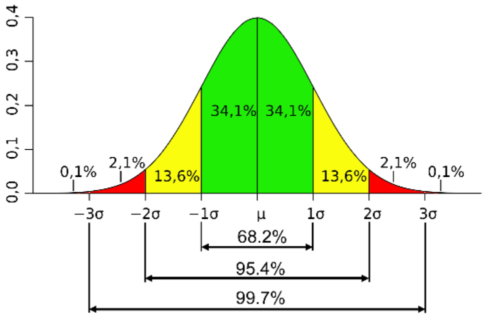
Important properties of a Normal Distribution:

- $\mu \pm \sigma)$
- $\mu \pm 2\sigma)$
- $\mu \pm 3\sigma)$



For each property half the area lies either side of the mean.

- $\mu + \sigma$ $\mu - \sigma$
- $\mu + 2\sigma$ $\mu - 2\sigma$
- $\mu + 3\sigma$ $\mu - 3\sigma$



Sketching a Normal Distribution:

Calculating number of SDs $Number\ of\ SD\ from\ mean = \frac{value - mean}{standard\ deviation}$

$\frac{960-1000}{15} = -2$ $\frac{1030-1000}{15} = 2$

$$Standardised\ Score = \frac{Score - Mean}{Standard\ Deviation}$$

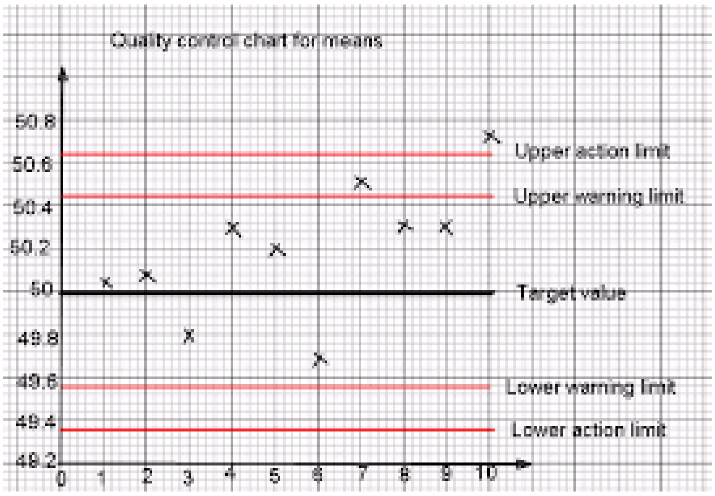
P

Involves checking samples to make sure products are all of the same quality and standard

How it works:

Control Chart

-
- K K
- K K


Quality control chart for means
Upper action limit
Upper warning limit
Target value
Lower warning limit
Lower action limit