

FIT1006 Lecture 2 Pre-reading

Surveys and Data Collection

Populations and samples, Collecting data,
Sources of data, Designing surveys,
Survey errors, Choosing a sample.

Reference:

Selvanathan 7th Ed. Chapter 2.

A Motivating Problem

A major bank currently offers Private Banking services to its clients in the CBD. The bank is considering extending its Private Banking business to include several major country towns: Geelong, Ballarat, Hamilton and Bendigo.

The bank wants to survey potential clients in these districts to determine whether likely demand would make this business viable.

Suggest a sampling/survey design for the bank to use, outlining the issues you need to consider and any problems you anticipate.

10 Survey questions

Consider how the questions below could be addressed using a survey:

1. Who is settling into university better, country or city students?
2. Are students happy with Campus Centre food?
3. What is the average savings of first year students?
4. Do savings differ between students of different faculties?
5. What is the average age of glider pilots?
6. Are OZ Lotto Division 1 winners happy one year after winning?
7. Where do the majority of visitors to my website click through from: Google? Yahoo? Or?
8. Is Coles cheaper than Woolworths?
9. What proportion of Australians support immigration?
10. Would the Liberal Party win an election held today?

Populations and Samples

For statistical purposes, a population refers to the whole collection, or set of things (people, companies, university students etc) that we wish to observe some aspect of.

A sample refers to the subset that we select to observe.

Generally speaking, we want to make an assessment of a large population by surveying a smaller, but representative sample.

Parameter/Sample

Population Parameter

A measure descriptive of the whole population. E.g.,
Average weekly wage.

Sample Statistic

Provides an estimate of the population parameter. E.g.,
The average weekly wage of our sample.

Thus, the sample statistic is used to estimate the
population parameter.

Variable/Observation

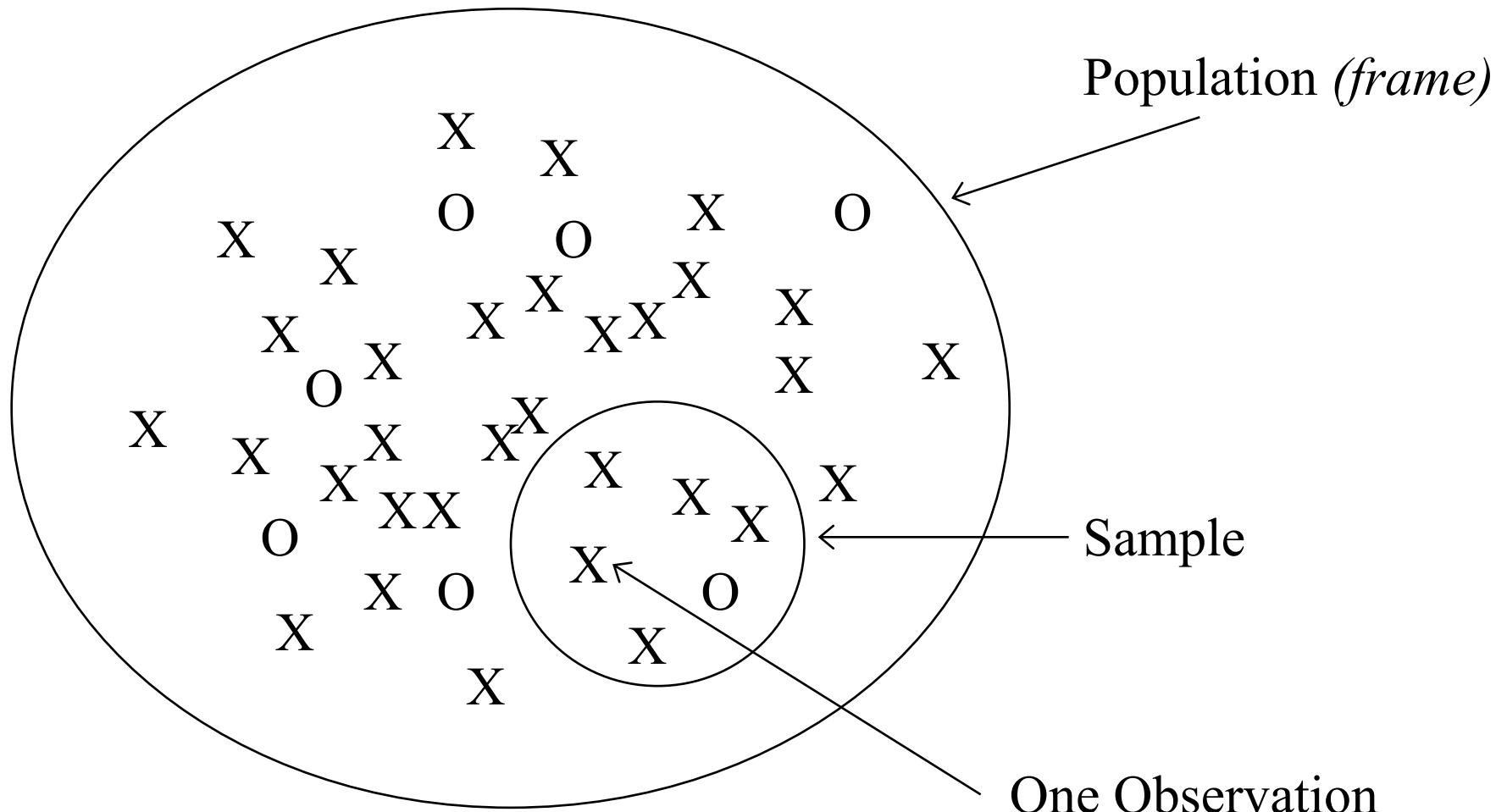
Variable

The characteristic of the population of interest. E.g.,
Weekly wage.

Observation

One data point in our sample. E.g., One person's
weekly wage.

The General Picture



Reasons for Using Samples

High cost of sampling a whole population.

Length of time required to sample whole population.

May have too much data.

May not have access to whole population.

Sampling can be destructive

Testing seat belts for breaking strain.

A census is a survey of a whole population.

The Australian population census is conducted once every five years.

Examples of surveys

Television ratings

Market research

Pre-election polls

Product registration information

Product Testing

Sources of (Primary) Data

Australian Bureau of Statistics

Surveys, Census, Routine recording – where companies are required by law to notify: Taxation, Administrative Processes – new houses, births etc.

Private Research

Financial Databases; Property sales monitoring;
Market Research companies; Internet traffic monitoring: **Google Analytics for example...**



Turn insights into action.

Get stronger results across all your sites, apps, and offline channels. Google Analytics Solutions offer marketing analytics products for businesses of all sizes to better understand your customers.

"Google's analytics products helped us improve engagement by 33% and click-throughs by 21% for content promotions on our homepage."

Types of Surveys

There are three main types of survey methods:

Personal (face to face) Interview

Telephone Survey

Self-administered survey (postal/internet)

Two main issues are at stake:

How to encourage people to respond truthfully and accurately?

How to conduct the survey with the minimum cost?

Personal Interview

Trained interviewers question people at home, at work, or on the street.

Advantages

Response rate is high, quality of data obtained is high
- responses are likely to be truthful and consistent.

Disadvantages

High cost, if travel is involved, then this method is also time consuming.

Telephone Interview

Trained interviewers question people over the telephone.

Advantages

Response rate is fairly high, quality of data obtained is high. *Computer Aided Telephone Interviewing* can be used to directly key responses and to detect inconsistencies.

Disadvantages

Complexity and length of questionnaire is limited.

Self-administered Questionnaire

Questionnaires distributed by mail and returned by mail. Internet survey participants are usually advised by email with the URL of the questionnaire (SurveyMonkey for example).

Advantages

Low cost, a large sample size is possible.

Disadvantages

The complexity of questions/quality of responses low.

Low rate of response.



Introduction

Thank you for taking this survey. The survey is sent to you by CrowdFundRES, a Horizon 2020 research project funded by the European Commission, and is carried out by a team of renewable energy developers, crowdfunding platforms, academics and crowdfunding experts. Full details of the project and the team can be found on our project website [CrowdFundRES](#).

* 1. You can complete this survey in English, French, German or Dutch. Please choose how to continue:

- I would like to complete this survey in English.
- Je voudrais répondre au questionnaire en Français.
- Ich möchte die Umfrage auf Deutsch beantworten.
- Ik wil graag antwoorden op deze enquête in het Nederlands.

Next

Stages of a Survey

- Exploratory interview with focus group
 - Identify issues, form hypothesis and questions
- Questionnaire designed and tested
- Experimental design and population *frame* is determined
- Samples are selected
- Questionnaire is administered
- Analysis and reporting

Survey Errors

Sampling Errors

The characteristics of the sample don't match the population. Usually addressed by taking a larger sample. Could the samples in the following slide have come from the population?

Non-sampling Errors

Errors in response or in recording data

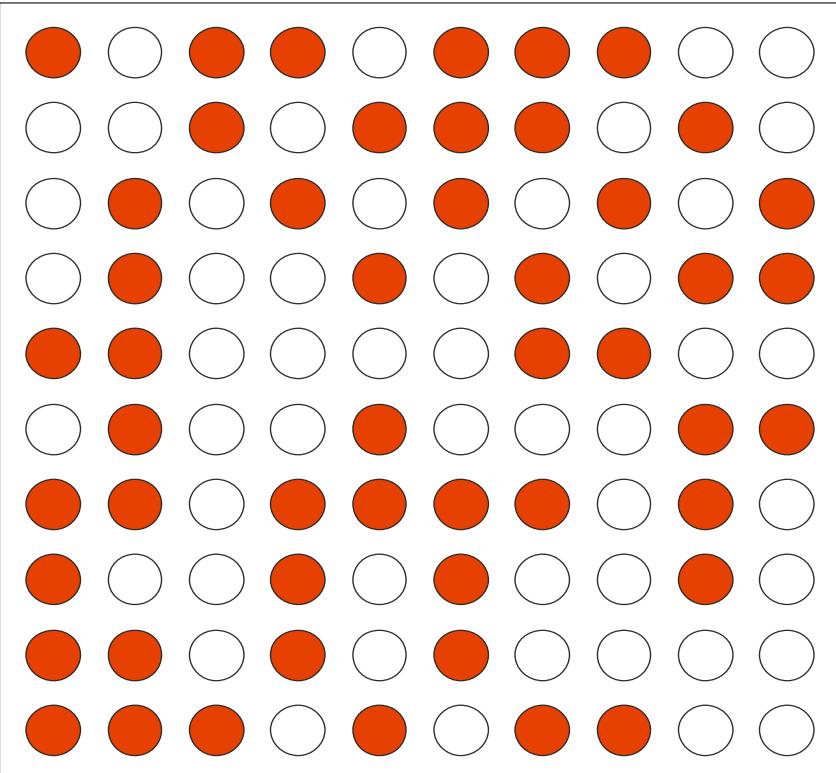
Misclassification or inaccurate response

Bias in the selection of the sample

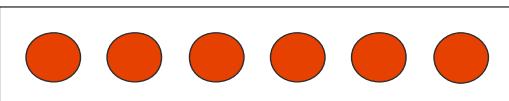
Non-response bias, Self-selection of respondents.

Survey Errors

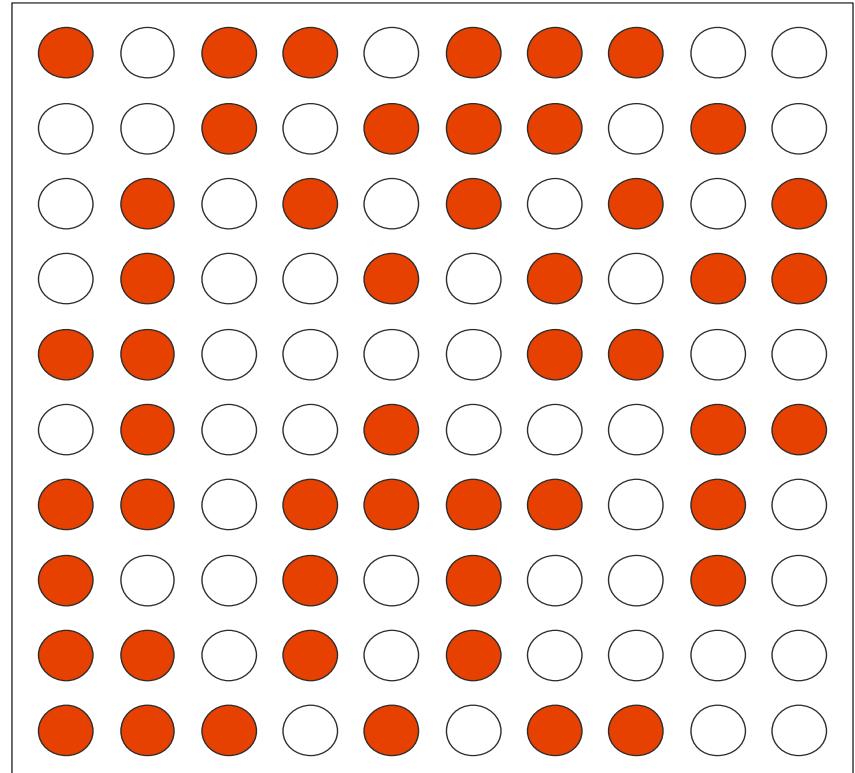
Population



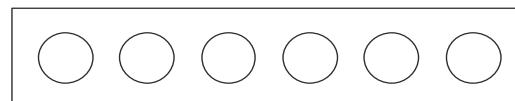
Sample



Population



Sample



Choosing a Sample

Probability Designs

Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling

Non-Probability Designs

Convenience Sampling

Quota Sampling

Judgemental Sampling

Snowball Sampling

Non-probability designs are more prone to bias

Random Sampling

Method

A frame of all sampling units is constructed and samples are selected randomly. *This is the ‘gold-standard’ for an unbiased sample.*

- Pulling names out of a hat
- Using random numbers

Considerations

- Typically unbiased selection
- May require a large sample to detect events with a low probability of occurrence. (Eg. OZ Lotto Division 1 winners).

Systematic Sampling

Method

A frame of all sampling units is constructed and every n^{th} sample is selected.

- Interviewing every 20th person

Considerations

- Typically unbiased selection
- Periodicity in data (if every n^{th} sample has some characteristic for example monthly temperatures)

Stratified Sampling

Method

A population is divided into distinct groups or strata.

Each strata is treated as a frame and sampled.

- A university may have 25,000 city students and 5,000 country students. A surveyor may interview 100 city students and 100 country students.

Considerations

- Enables small groups to be represented. This may be required to obtain conclusions with suitable statistical power.

Cluster Sampling

Method

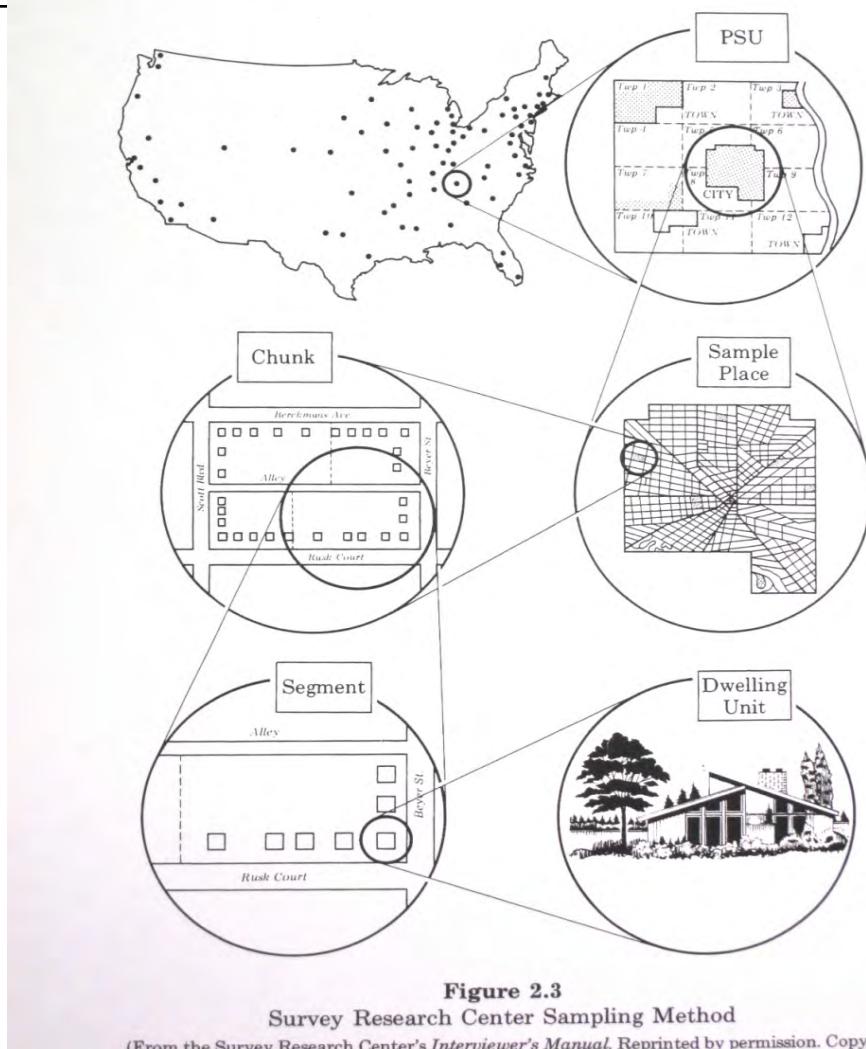
Clusters in a population are identified. These become the frames and are sampled.

- Employees of fast food restaurants may be surveyed by first choosing restaurants randomly (clusters) and then choosing staff from the restaurant at random.

Considerations

- Is an efficient way to sample a group that may be geographically wide spread.
- Clusters should be similar to each other.

Choosing Whom to Interview. Rather than let the interviewer make a subjective choice of respondents after all this care in drawing the sample, objective procedures have been developed to choose the



Source: An Introduction to Survey Research and Data Analysis, Weisberg & Bowen, W. H. Freeman and Company, San Francisco (1970s)

Convenience Sampling

Method

A surveyor interviews an easily accessible group.

- Your lecturer surveying you.
- Passers by on a street.

Considerations

- May be biased, and certainly non-random.
- Conclusions are limited.

The Age online newspaper used to have pop-polls like the one below many years ago!

Poll Booth

Should job seekers who refuse to travel more than 90 minutes to a job be denied unemployment benefits?

Yes



22%

No



78%

Total votes: 502. [CAST YOUR VOTE](#) | Poll closes in 6 hours.

Disclaimer:

These polls are not scientific and reflect the opinion only of visitors who have chosen to participate.

Judgemental Sampling

Method

An expert identifies a representative sample.

- Passers by on a street which seem to be unemployed.
- Credit card users who maintain a debt of \$10,000 or more.

Considerations

- The sample is deliberately biased.
- Enables a researcher to focus on a group with special qualities.
- Those prominent in a sample tend to be selected.

Quota Sampling

Method

People are surveyed until enough responses are obtained.

- Samples may be chosen by any of the other methods.

Considerations

- The sample may be biased by considering people who speak English well enough to be interviewed, or people who are at home when the telephone interviewer rings.
- Self-selection by those prominent in the sample may arise.

Snowball Sampling

Method

People surveyed recommend others to be surveyed.
This method utilizes the power of social networks!

- Allows for hard to identify groups to be surveyed.
- For example Hot-Air Balloon Pilots.

Considerations

- Maybe biased.
- Allows hard to detect groups in the population to be identified.
- The name comes from the ‘snowball effect’.

Motivating Problem

A major bank currently offers Private Banking services to its clients in the CBD. The bank is considering extending its Private Banking business to include several major country towns: Geelong, Ballarat, Hamilton and Bendigo.

The bank wants to survey potential clients in these districts to determine whether likely demand would make this business viable.

Suggest a sampling/survey design for the bank to use, outlining the issues you need to consider and any problems you anticipate.

Key ideas

You should be familiar with the following:

- Parameter, Sample Statistic, Variable, Observation;
- The reasons for sampling;
- The 3 main methods of surveying and the advantages and disadvantages of each;
- Causes of error in surveys;
- Methods for choosing samples and the advantages and disadvantages of each.

FIT1006 Lecture 3 Pre-reading

Graphical Presentation of (Quantitative) Data

Types of Data

Visualisation of Data

Tally, Frequency Table; Stem and Leaf Plot, Histogram.

*You can read about creating pie charts and column graphs in Excel on your own.

*We will cover multivariate data and time series later in the course.

Textbook: 7th Ed. Sections 2.1, 3.1, 4.1.

Motivating problem...

A grocery store wants you to analyse the amount spent by their customers. They have given you the sales history of 10 randomly sampled customers.

Oh, you can't use a calculator or computer...

We'll work on this in Lectures 4 & 5 also...

Introduction

Different types of data allow us to perform different methods of analysis.

The first step in analysing quantitative data should be to ‘see’ the data in order to observe the underlying pattern, or distribution, of the data.

It is the distribution of the data which determines which statistical measures are appropriate.

Types of Data

Quantitative

Discrete eg. 1, 2, 3...

Continuous eg. 3, 4.256, 3.999, 11.2, ...

We can calculate summary numerical statistics

Ordinal

Qualitative eg. agree, neutral, disagree

Qualitative eg. Lecturer, Senior Lecturer, Professor

We can calculate statistics based on rank

Nominal

Categorical eg, red, green, silver, ...

We can calculate statistics based on counts

Data

The following marks were obtained by students of
FIT1234 Basic Business Applications:

81	59	53	58	52	73	55	37	58	52
48	69	65	58	42	57	63	37	53	54
52	61	38	42	24	67	68	49	76	55
71	73	48	62	38	53	47	60	51	51
63	75	27	57	17	37	76	91	26	53

Note that we can't say much about the students' results just by looking at the data.

Tally

The simplest form of summary is to construct a tally which groups the data into class intervals.

Class	Tally
0 - 10	
10 - 20	
20 - 30	
30 - 40	++++
40 - 50	++++
50 - 60	++++ +--+ +--+
60 - 70	++++
70 - 80	++++
80 - 90	
90 - 100	
100+	

Frequency Distribution

A count of the tally in each of the class intervals gives a frequency distribution.

Class	Frequency
0 - 10	0
10 - 20	1
20 - 30	3
30 - 40	5
40 - 50	6
50 - 60	18
60 - 70	9
70 - 80	6
80 - 90	1
90 - 100	1
100+	0

Determining class intervals

Some guidelines:

- Class intervals should lie between possible observations.
(Previous example breaks this rule)
- Express cut-offs at one more decimal point accuracy than data.
- For n data, \sqrt{n} intervals is usually a good starting point.
- Sturge's formula: classes = $1 + 3.3\log_{10}n$
- Choose 'intuitive' intervals: 1, 2, 5, 10, 20, 50, 100 etc.
- It is more usual to use equal sized intervals.

Class Intervals in Practice

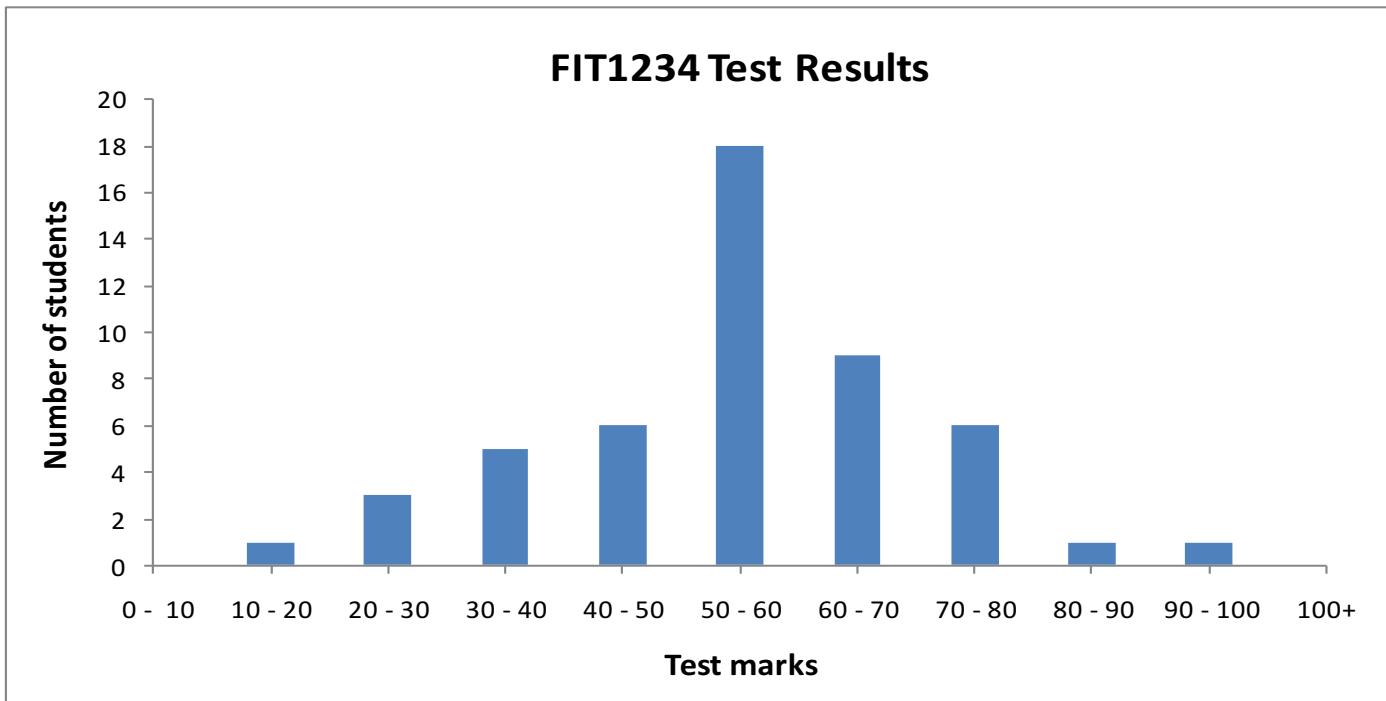
81	59	53	58	52	73	55	37	58	52
48	69	65	58	42	57	63	37	53	54
52	61	38	42	24	67	68	49	76	55
71	73	48	62	38	53	47	60	51	51
63	75	27	57	17	37	76	91	26	53

We have 50 observations:

- Now, $\sqrt{50}$ is approximately 7; $1 + 3.3\log_{10}50 = 6.6$.
- The lowest value is 17, the greatest value is 91.
- $(91-17)/7$ is approximately 10 suggesting class intervals 10 units wide, starting at 10.

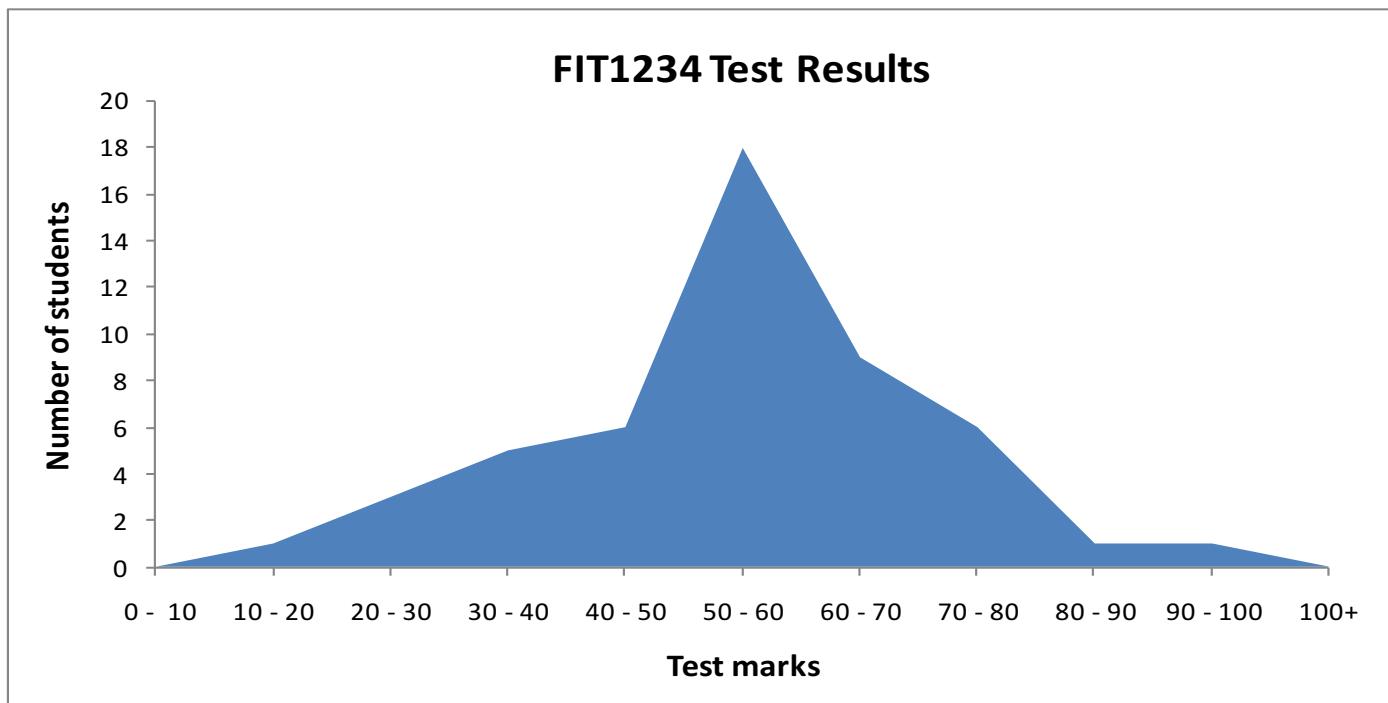
Histogram

We refer to a column graph of the frequency count as a histogram. Note: it is conventional, but not strictly necessary, that all intervals are the same width.



Frequency Polygon

Connect the frequency values of the histogram to make a frequency polygon. The polygon starts and ends at zero to form a closed shape.



Cumulative Frequency

We can sum the frequencies, from lowest to highest or highest to lowest and thus create ‘greater than’ or ‘less than’ cumulative frequency counts.

Class Limit	Less Than	Greater Than
0 - 10	0	50
10 - 20	1	50
20 - 30	4	49
30 - 40	9	46
40 - 50	15	41
50 - 60	33	35
60 - 70	42	17
70 - 80	48	8
80 - 90	49	2
90 - 100	50	1
100+	50	0

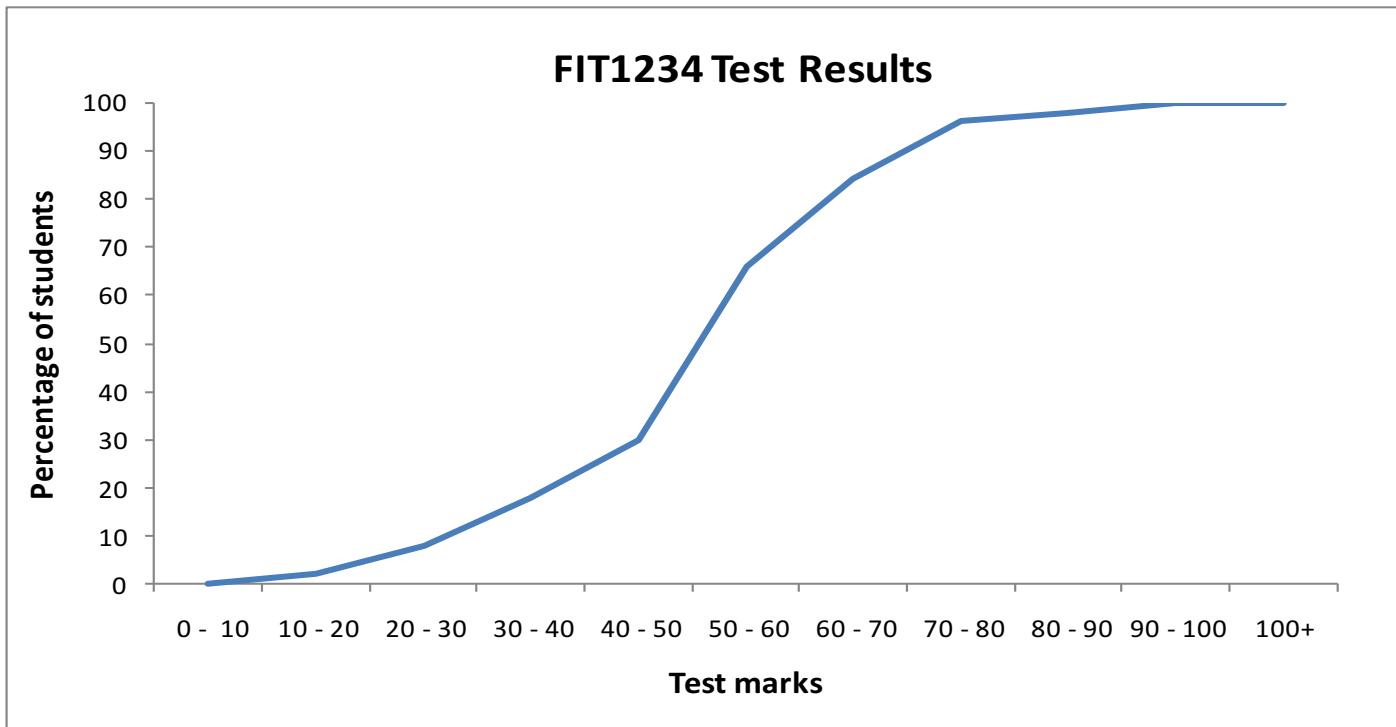
Cumulative Relative Frequency

Each of the cumulative frequencies is expressed as a percentage of the total (with a maximum of 100%).

Class Limit	Less Than (%)	Greater Than (%)
0 - 10	0	100
10 - 20	2	100
20 - 30	8	98
30 - 40	18	92
40 - 50	30	82
50 - 60	66	70
60 - 70	84	34
70 - 80	96	16
80 - 90	98	4
90 - 100	100	2
100+	100	0

Cumulative Frequency Plot

A graph of the cumulative relative frequencies. The resulting plot is called an *ogive*.



Stem and Leaf Plot 1

So far, we have used the tally and frequency count approaches to grouping data. These methods are widely used, but suffer from the problem that the subsequent tables contain no information about individual data values.

The stem and leaf plot overcomes this weakness by displaying the data values as leaves on stems

We will start by treating 10s as stems, with units as leaves. That is, $17 = 10 + 7$. This model works well for data values less than 100.

Stem and Leaf Plot 2

This plot has 3 columns showing cumulative count (from top and bottom), stem and leaves. The middle stem contains the median indicated by ().

Leaf Unit = 1.0		
1	1	7
4	2	467
9	3	77788
15	4	227889
(18)	5	11222333345577889
17	6	012335789
8	7	133566
2	8	1
1	9	1

Stem and Leaf Plot 3

We can construct plots with smaller intervals. In this case, intervals of 5 units are used. No stem contains the median in this case.

Note that the count is not always shown in the plot.

Leaf Unit = 1.0		
1	1	7
2	2	4
4	2	67
4	3	
9	3	77788
11	4	22
15	4	7889
25	5	1122233334
25	5	55778889
17	6	01233
12	6	5789
8	7	133
5	7	566
2	8	1
1	8	
1	9	1

Stem and Leaf Plot 4 - SYSTAT

1	7
2	4
* * * Outside Values * * *	
2	67
3	77788
4 H	227889
5 M	112223333455778889
6 H	012335789
7	133566
8	1
* * * Outside Values * * *	
9	1

(We'll discuss outside values next lecture)

1	7
2	4
* * * Outside Values * * *	
2	67
3	
3	77788
4	22
4 H	7889
5 M	1122233334
5	55778889
6 H	01233
6	5789
7	133
7	566
8	1
* * * Outside Values * * *	
9	1

Back-to-Back Stem and Leaf Plot

Group 1

31	10	26	33	29	14
19	27	16	13	25	26
27	25	21	17	6	

Group 2

27	28	23	22	33	15
23	31	21	20	23	26
20	20	28	29	27	27
29	25	26	26	19	30
27	18	30	26	26	22

Group 1		Group 2	
Leaf = 1		Leaf = 1	
	6	0	
		0	
	0	1	
		3	1
	4	1	5
	67	1	
		9	1 89
		1	2 1
		2	22333
	55	2	5
	6677	2	666667777
		9	2 8899
		1	3 1
		3	3 3

Visualising Data

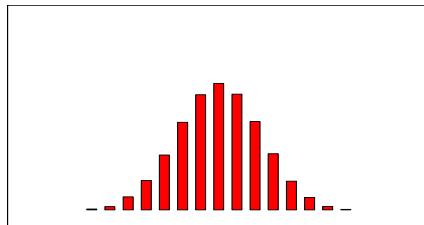
Why do we want to ‘see’ our data?

- Visual inspection is the fastest way to get an overview of data.
- Visual inspection enables a description of the distribution of the data to be made.
- The distribution of data determines which statistics are appropriate.
- To make comparisons between data from different groups.

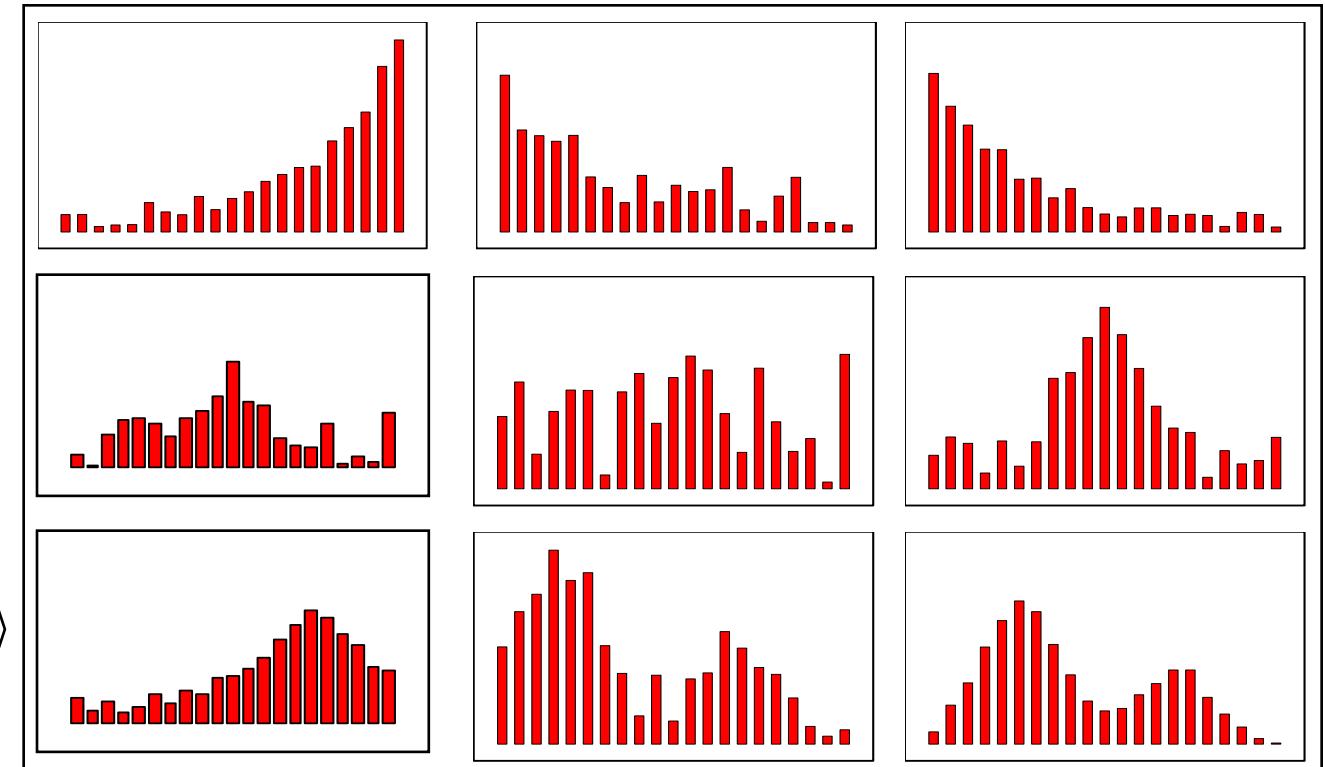
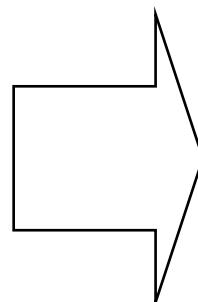
Comparing Distributions

Histograms of some other distributions

Normal Dist.



Some *skewed*,
bimodal, and
uniform
distributions



Example 1

A class of 20 students obtained the following results in a class test:

46	43	58	59	48	67	42	50	55	62
50	44	47	91	54	60	48	57	42	50

We will construct a tally, stem and leaf plot and histogram of the data.

Solutions A

Stem and leaf plot from SYSTAT.

Stem and Leaf Plot of variable: TEST_SCORES, N = 20

Minimum: 42.000

Lower hinge: 46.500

Median: 50.000

Upper hinge: 58.500

Maximum: 91.000

4 2234

4 H 6788

5 M 0004

5 H 5789

6 02

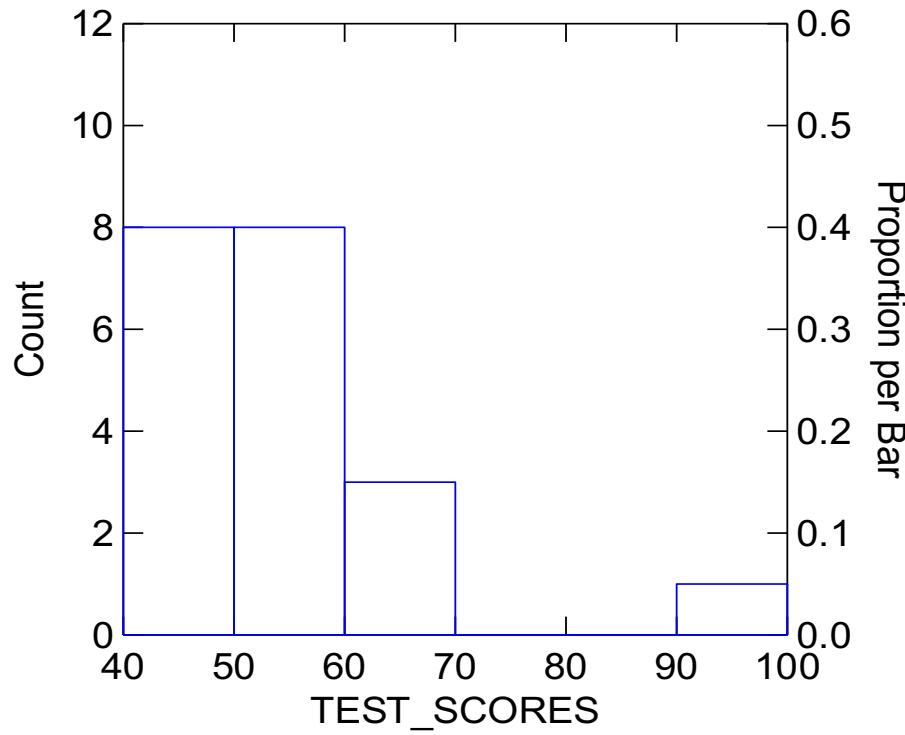
6 7

* * * Outside Values * * *

9 1

Solutions B

SYSTAT gives the following histogram by default.



Key Ideas

You should be able to construct a tally, histogram, cumulative frequency plot and stem and leaf plot of a data set.

You should be able to calculate an appropriate class interval, determine cut-offs and mid-points.

Get into the habit of always visualising a data set before you start your analysis.

FIT1006 Lecture 4 – Pre-reading

Descriptive Statistics

Measures of Centre

Mean, Median, Mode, Trimmed Mean, Robust Statistics

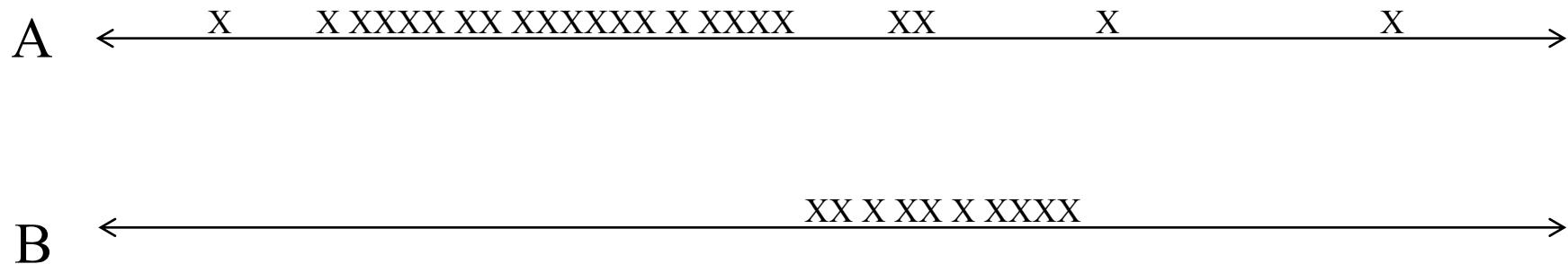
Measures of Spread

Variance and Standard Deviation; Quartiles and
Percentiles; Boxplots; Quick Quartiles.

Textbook: 7th Ed. Sections 5.1 – 5.3.

Learning Objectives

This lecture is about how we characterise a data set using summary statistics. A typical problem that could be answered with the techniques covered is: describe the differences between the two data sets A and B below?



Data

The following marks were obtained by students of
FIT1234 Basic Business Applications:

81	59	53	58	52	73	55	37	58	52
48	69	65	58	42	57	63	37	53	54
52	61	38	42	24	67	68	49	76	55
71	73	48	62	38	53	47	60	51	51
63	75	27	57	17	37	76	91	26	53

Mean: \bar{x}

The mean, or sample mean, is the most well known measure of centre. It is also known commonly as the average. It is the sum of data divided by the number of data.

The formula for the mean is:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The calculation of the mean of the FIT1234 data is:

$$\bar{x} = \frac{81 + 59 + 53 + \dots + 53}{50} = 54.6$$

Median (Me)

The sample median is the central observation in a data set.

To calculate the median:

Order observations from lowest to highest.

- If there is an odd number of observations then the median is the central observation.
- If there is an even number of observations, then the median is the average of the two central observations.

The stem and leaf plot is a useful starting point.

Calculating the Median

For our data, we have 50 observations and so the median will be the average of the 25th and 26th ordered observations

Putting the data in ascending order, we get:

17	24	26	27	37	37	37	38	38	42
42	47	48	48	49	51	51	52	52	52
53	53	53	53	54	55	55	57	57	58
58	58	59	60	61	62	63	63	65	67
68	69	71	73	73	75	76	76	81	91

Thus the median is

$$Me = \frac{54 + 55}{2} = 54.5$$

Mode (Mo)

The mode is the most frequently occurring observation.

From the stem and leaf plot, we observe that the mode is 53.

1	7
2	467
3	7788
4	H 227889
5	M 112223333455778889
6	H 012335789
7	133566
8	1
9	1

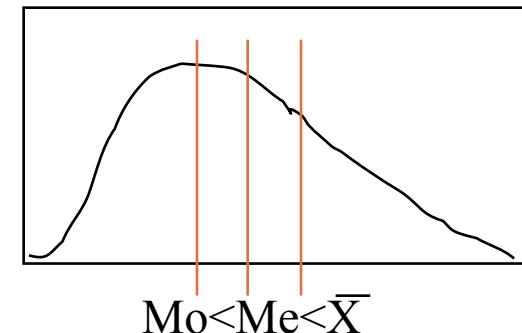
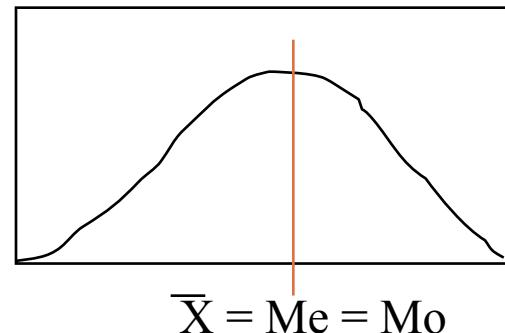
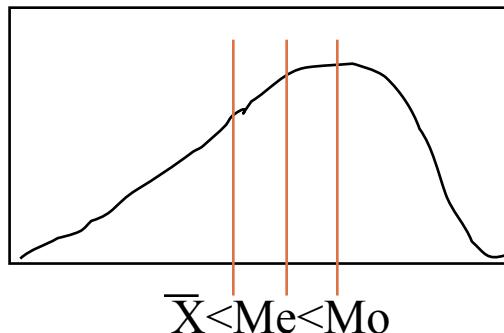
Note that for ungrouped or continuous data, the mode may not exist.

Mean ν Median ν Mode

The mean and median provide the most usual measures of centre for quantitative data.

If the data is symmetrically distributed then either the mean or median are acceptable and the mean is usually preferred.

If the data is skewed or contains exceptionally high or low values then the median is usually preferred.



Example 1

A class of 20 students obtained the following results in a class test:

46	43	58	59	48	67	42	50	55	62
50	44	47	91	54	60	48	57	42	50

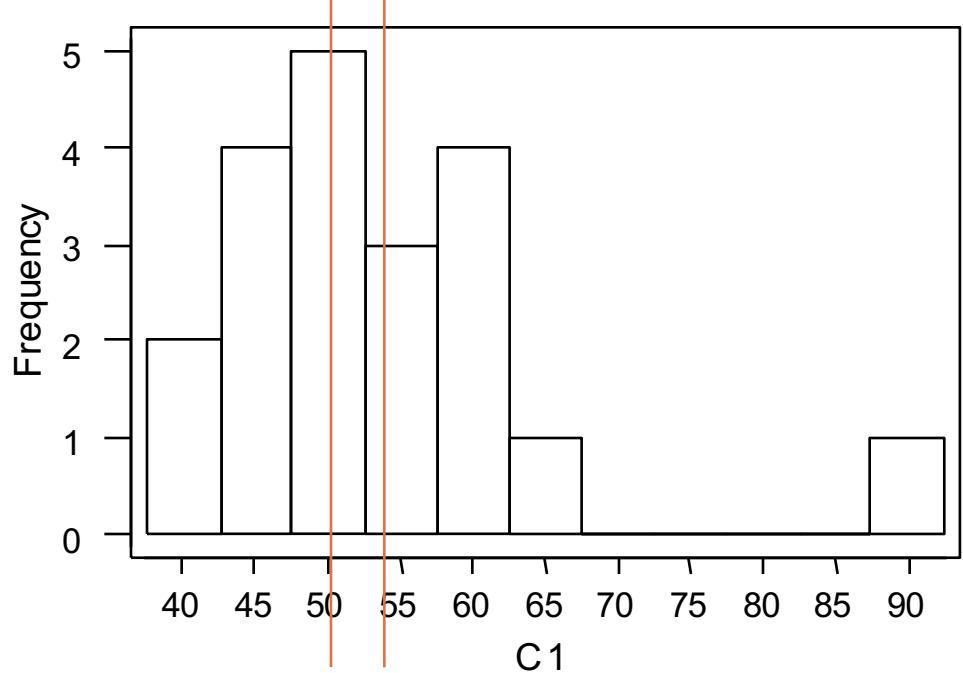
Calculate the mean and median of these data.

Results

Mean = 53.7

Median = 50.0

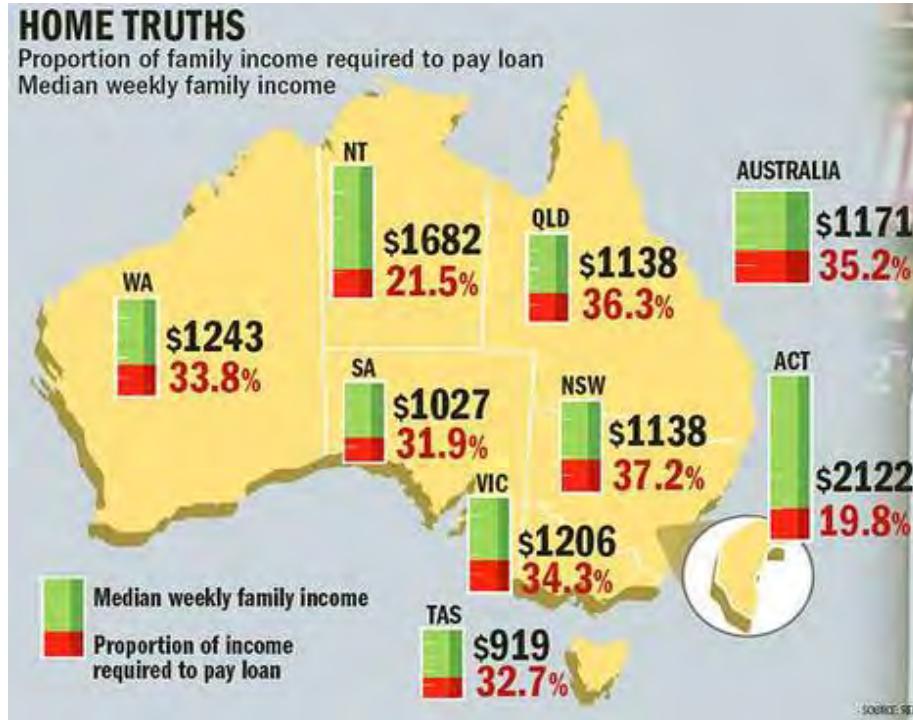
We observe that the exceptionally large observation (91) has distorted the mean.



In this type of situation, the median more accurately represents the centre of the majority of data.

Note: *mean or median?*

An article on housing affordability in *The Age* (on-line) 9th March 2007 contained the illustration below. Why is *median* weekly household income used instead of the *average*?



Trimmed Mean

The $a\%$ trimmed mean provides some compromise between the mean and the median. The highest and lowest $a\%$ of values are trimmed from the data. The mean of the remainder is then calculated.

To calculate the 10% trimmed mean:

17	24	26	27	37	37	37	38	38	42
42	47	48	48	49	51	51	52	52	52
53	53	53	53	54	55	55	57	57	58
58	58	59	60	61	62	63	63	65	67
68	69	71	73	73	75	76	76	81	91

Calculate the mean ignoring the upper and lowest 10%.
Trimmed Mean = 55.0.

Variance & Standard Deviation - s

The standard deviation is perhaps the most commonly used measure of the spread of data.

The variance is calculated as the average of the squared deviations from the mean, adjusted for the fact that we are only considering a sample.

The standard deviation is the square root of the variance.

Two standard formulas are used in practice:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n - 1}}$$

Calculating s - method a

As calculating the standard deviation manually is tedious, we will calculate s for a small data set: 81

$$59 \quad 53 \quad 58 \text{ using } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

$$\bar{x} = \frac{81 + 59 + 53 + 58}{4} = 62.75$$

$$\begin{aligned}s &= \sqrt{\frac{(81 - 62.75)^2 + (59 - 62.75)^2 + (53 - 62.75)^2 + (58 - 62.75)^2}{3}} \\&= \sqrt{154.91} = 12.4\end{aligned}$$

Calculating s - method b

We will now calculate the standard deviation of

$$81 \ 59 \ 53 \ 58 \text{ using } s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n - 1}}$$

$$\begin{aligned} s &= \sqrt{\frac{81^2 + 59^2 + 53^2 + 58^2 - \frac{(81+59+53+58)^2}{4}}{3}} \\ &= \sqrt{154.91} = 12.4 \end{aligned}$$

This method is computationally simpler than first method – and is generally preferred for coding.

Interpreting Standard Deviation 1

For *Normally* distributed data we can say that approximately:

67% of data lies within ± 1 standard deviation of the mean

95% of data lies within ± 2 standard deviations of the mean

99% of data lies within ± 3 standard deviations of the mean

We use this model to establish $X\%$ *Confidence Intervals*.

Confidence Intervals

Using the mean and standard deviation calculations, we can establish confidence intervals for our data.

17	24	26	27	37	37	37	38	38	42
42	47	48	48	49	51	51	52	52	52
53	53	53	53	54	55	55	57	57	58
58	58	59	60	61	62	63	63	65	67
68	69	71	73	73	75	76	76	81	91

mean	54.6
standard deviation	15.1

Mean +/- 1 St Dev	39.5	69.8
Mean +/- 2 St Dev	24.3	84.9
Mean +/- 3 St Dev	9.2	100.1

We see that our data fits the theoretical limits quite well.

Coefficient of Variation

In order to compare the relative variability of several data sets, a proportional measure of variability: the coefficient of variation may be used. This measure is commonly used in financial analysis for comparing the variability of share prices or alternative investments.

For a sample:

$$cv = \frac{s}{\bar{x}}$$

Quartiles

Ranked data can be divided into four quartiles. 25% of the data is less than the first - or lower quartile, 50% lower than the second quartile - or median, 75% of data is less than the third - or upper quartile. In SYSTAT, the upper and lower quartiles are referred to as ‘Hinges’.

For a data set $x_1, x_2 \dots x_n$ arranged in ascending order

We wish to find the Qth quartile, $Q=1, 2 \text{ or } 3$

$$q = (n + 1) \frac{Q}{4} \text{ and } Q = x_q \quad (\text{the required value of } x)$$

When q is non-integer we calculate

$$Q = x_q + r(x_{q+1} - x_q) \text{ where } r \text{ is the fractional part of } q$$

Quartiles - calculating

To find the first and third quartiles of our data set.

17	24	26	27	37	37	37	38	38	42
42	47	48	48	49	51	51	52	52	52
53	53	53	53	54	55	55	57	57	58
58	58	59	60	61	62	62	63	65	67
68	69	71	73	73	75	76	76	81	91

$$q = \frac{1(51)}{4} = 12.75 \text{ thus}$$

$$Q1 = x_{12} + 0.75(x_{13} - x_{12})$$

$$Q1 = 47 + 0.75(48 - 47) = 47.75$$

$$q = \frac{3(51)}{4} = 38.25 \text{ thus}$$

$$Q3 = x_{38} + 0.25(x_{39} - x_{38})$$

$$Q3 = 63 + 0.25(65 - 63) = 63.50$$

We have already calculated $Q2 = \text{Median} = 54.5$

Range, Interquartile Range

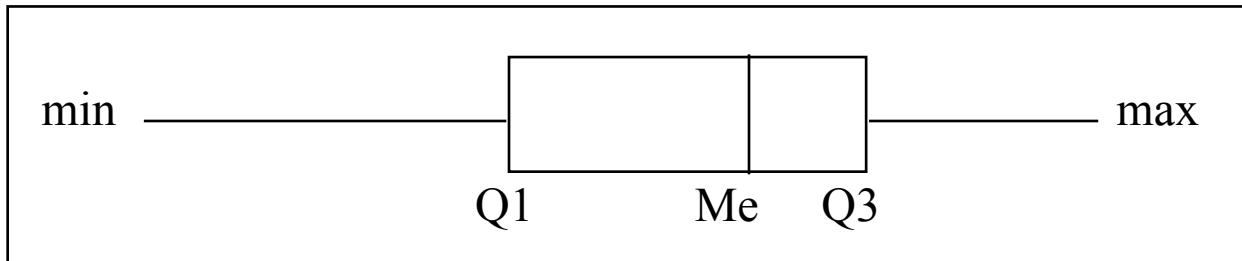
We describe the range of our data as highest value - lowest value
 $= 91 - 17 = 74$

We describe the interquartile range as $Q3 - Q1$
 $= 63.50 - 47.75 = 15.75$

We observe that the middle 50% of data lies within the interquartile range.

Boxplots 1

A boxplot, otherwise known as a box and whisker diagram is, in its simplest form, a five point data summary showing the minimum, maximum, lower quartile, upper quartile and median.



Boxplots are the most useful tools for comparing data sets, and can be drawn horizontally or vertically.

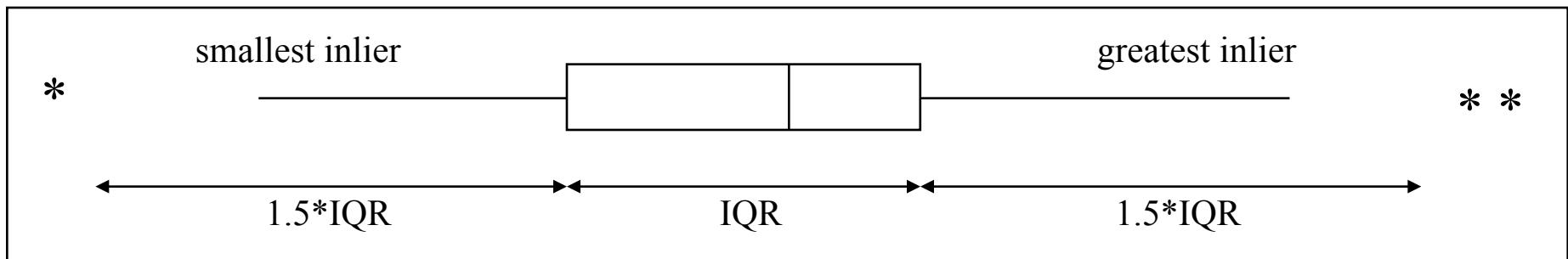
Boxplots 2

An alternative form of the boxplot is to extend the whiskers of the boxplot to include only the inlying values. These can be thought of as data that falls within the main central cluster (roughly speaking ± 2 standard deviations).

Recall that $IQR = Q3 - Q1$.

Inliers are all the values greater than $Q1 - 1.5*IQR$ and less than $Q3 + 1.5*IQR$

Outliers are values outside this range denoted by ‘*’



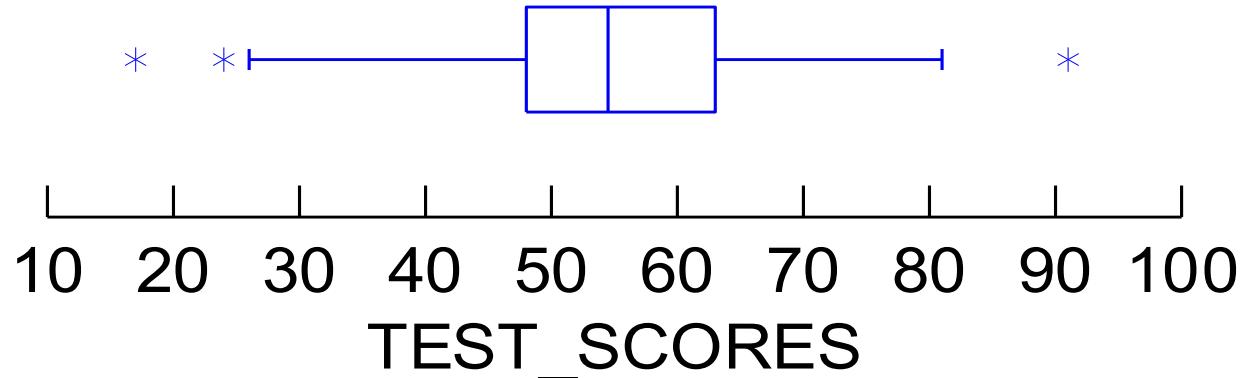
Boxplots 3

From the test score data: $Q3 = 63$, $Q1 = 48$, $IQR = 15$

$Q1 - 1.5 * IQR = 25.5$ and less than $Q3 + 1.5 * IQR = 85.5$

From the data, the smallest inlier is: 26 the greatest inlier is 81.

A boxplot of the test score data drawn in SYSTAT is shown below.



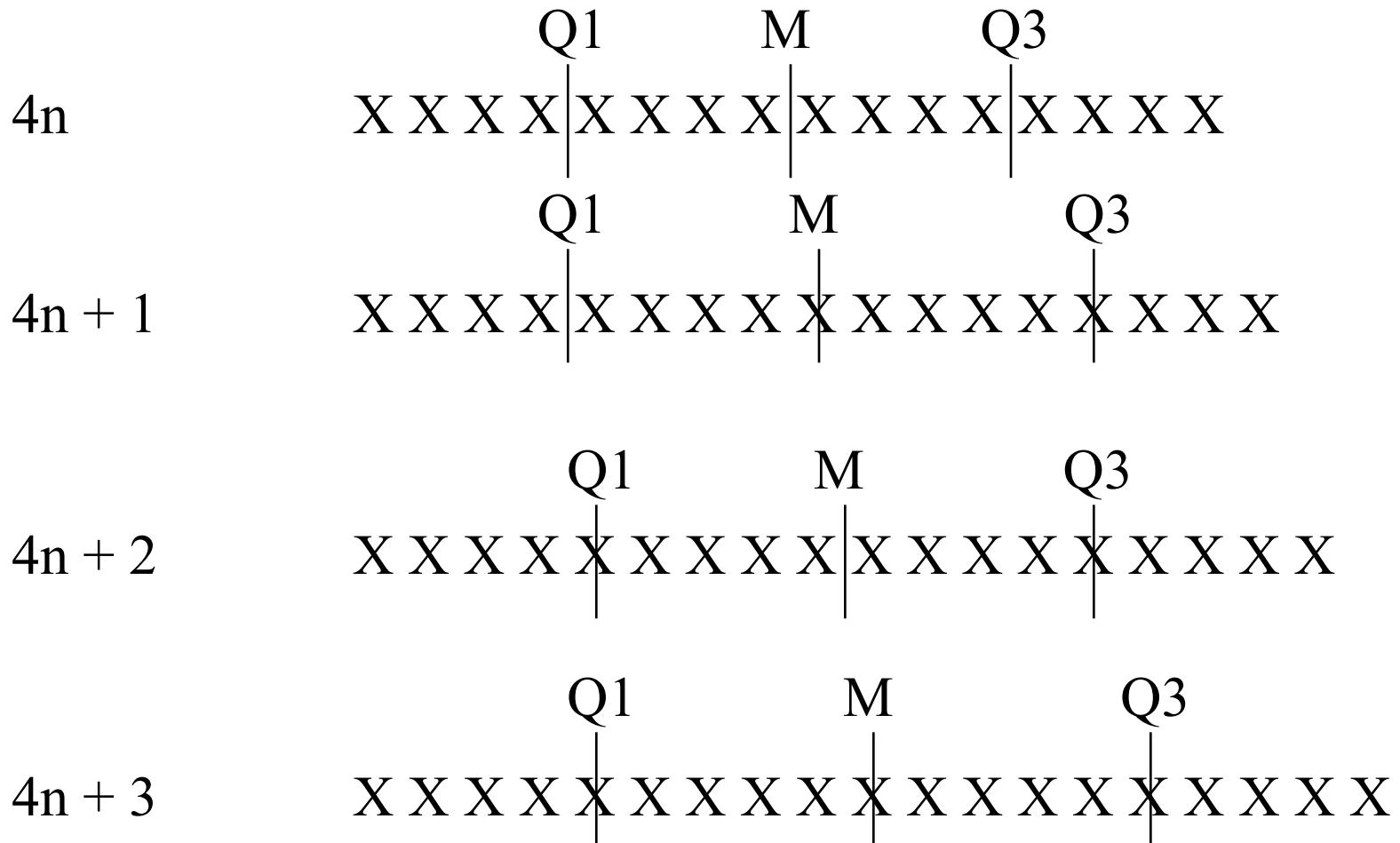
Stem and Leaf Plot

A stem and leaf plot of the data has been produced in SYSTAT with the accompanying data summary. From the stem and leaf plot we can read off the data in ascending order.

Stem and Leaf Plot of variable: TEST_SCORES, N = 50 Minimum: 17.000 Lower hinge: 48.000 Median: 54.500 Upper hinge: 63.000 Maximum: 91.000	
--	--

	<p style="text-align: center;">* * * Outside Values * * *</p> <table><tr><td>1</td><td>7</td></tr><tr><td>2</td><td>4</td></tr><tr><td>2</td><td>67</td></tr><tr><td>3</td><td></td></tr><tr><td>3</td><td>77788</td></tr><tr><td>4</td><td>22</td></tr><tr><td>4</td><td>H 7889</td></tr><tr><td>5</td><td>M 1122233334</td></tr><tr><td>5</td><td>55778889</td></tr><tr><td>6</td><td>H 01233</td></tr><tr><td>6</td><td>5789</td></tr><tr><td>7</td><td>133</td></tr><tr><td>7</td><td>566</td></tr><tr><td>8</td><td>1</td></tr><tr><td>9</td><td>1</td></tr></table> <p style="text-align: center;">* * * Outside Values * * *</p>	1	7	2	4	2	67	3		3	77788	4	22	4	H 7889	5	M 1122233334	5	55778889	6	H 01233	6	5789	7	133	7	566	8	1	9	1
1	7																														
2	4																														
2	67																														
3																															
3	77788																														
4	22																														
4	H 7889																														
5	M 1122233334																														
5	55778889																														
6	H 01233																														
6	5789																														
7	133																														
7	566																														
8	1																														
9	1																														

Quick Quartiles



Percentiles

In the same way that quartiles divide the data into four, we can use percentiles to divide our data into one hundredths.

We can calculate the C^{th} percentile and say that $C\%$ of data lies below this value. The median is the 50^{th} percentile

For a data set $x_1, x_2 \dots x_n$ arranged in ascending order

We wish to find the C^{th} percentile, $C = 0, 1, 2 \dots 100$

$$p = (n+1) \frac{C}{100} \text{ and } P_C = x_p \quad (\text{the required value of } x)$$

When p is non-integer we calculate

$$P_C = x_p + r(x_{p+1} - x_p) \text{ where } r \text{ is the fractional part of } p$$

Percentiles - calculating

We calculate percentiles in the same way as we calculated quartiles.

To find the 5th percentile of the data.

$$p = \frac{5(51)}{100} = 2.55 \text{ thus}$$

$$P_5 = x_2 + 0.55(x_3 - x_2)$$

$$P_5 = 24 + 0.55(26 - 24) = 25.10$$

Data analysts commonly use the 5th 10th, 90th and 95th percentiles by convention when comparing data.

Key Ideas

Hand calculations for small data sets.

Measures of Centre: Mean, Median, Mode,
Trimmed Mean. Median *vs* Mean.

Measures of Spread: Variance and Standard
Deviation.

Quartiles and Percentiles, Boxplots, Quick
Quartiles.

Next week: larger data sets using Excel and
SYSTAT.

FIT1006 Lecture 7 – Pre-reading

Correlation:

The Linear model.

Calculating q and r by hand.

Calculating r using Excel and SYSTAT.

Interpreting q and r .

Visual estimation of q and r .

Textbook:

7th Ed. Sections 4.3, 5.5.

Bivariate Data

Bivariate data is data which consists of pairs of observations made about a single subject, or made at the same time.

For example we might measure a person's height and weight.

We might observe the number of branches that a bank has as well as the number of employees.

More than two observations of a single subject yields *multivariate data*.

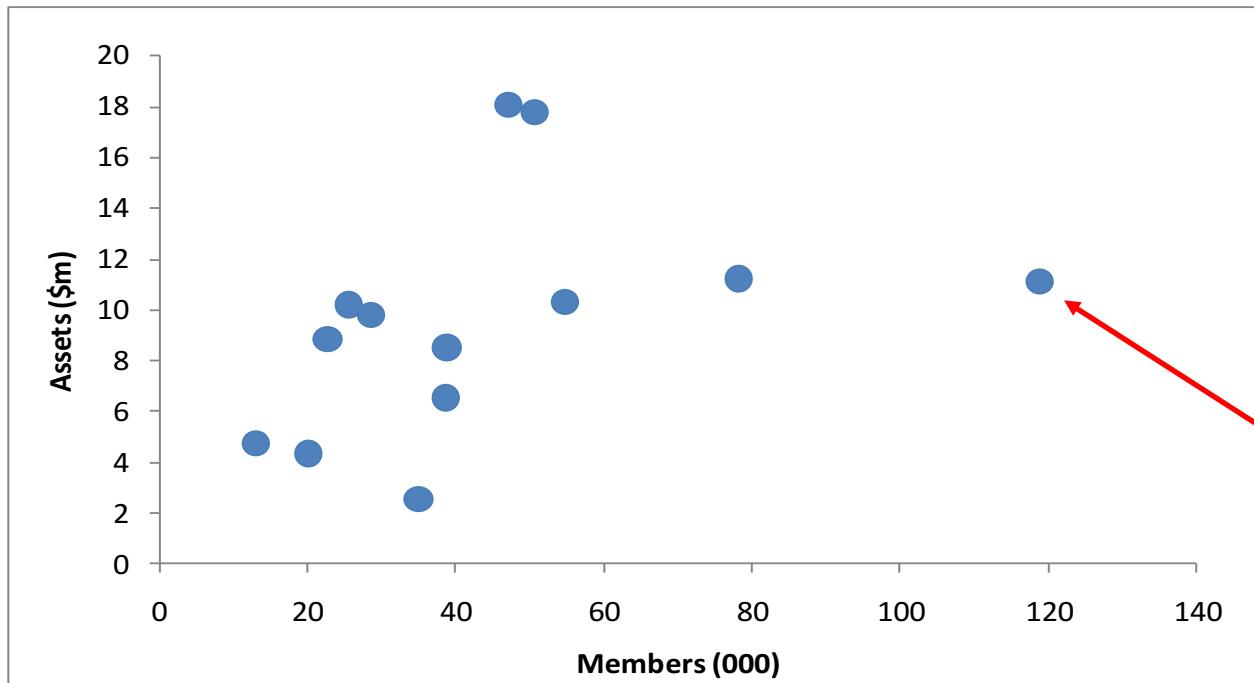
Sample Data

We will first consider the relationship between Total Assets (\$m) and Number of Members (000) the for Major Credit Unions. (Source KPMG financial database)

Members (000)	Assets (\$m)
39	9
78	11
20	4
23	9
47	18
25	10
51	18
29	10
55	10
39	7
119	11
13	5
35	3

Scatterplot

A scatterplot of the data shows both variables together.



Members (000)	Assets (\$m)
39	9
78	11
20	4
23	9
47	18
25	10
51	18
29	10
55	10
39	7
119	11
13	5
35	3

Correlation

Correlation provides a measure of the strength of the linear relationship that exists between the variables in the paired observations.

The most common correlation measure used is Pearson's Product Moment Correlation Coefficient – usually referred to as Pearson's ' r '

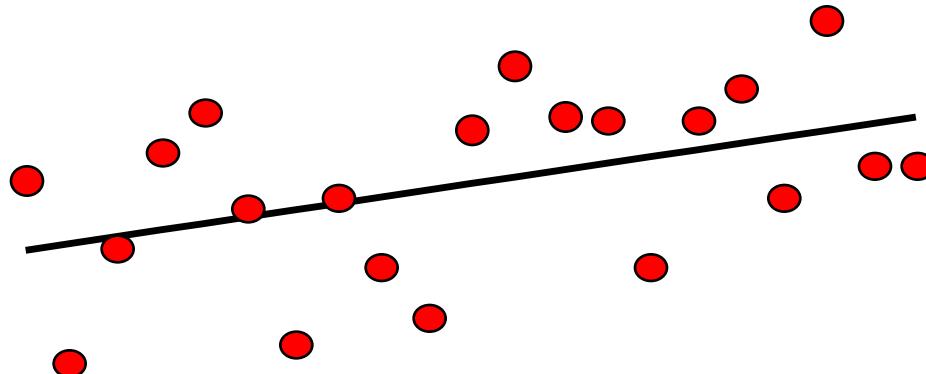
We will also use a robust measure of correlation: q -Correlation.

In both cases the measures of correlation range from -1 to 1, with 1 indicating a strong positive relationship, 0 indicating no relationship and -1 indicating a strong negative relationship.

Linear Relationship

When we determine the degree of correlation between variables we are assuming that the variables have a linear relationship.

For two variables x , and y , we say that $y = ax + b + e$, where e are random, Normally distributed errors.



q-Correlation

To calculate the degree of correlation using this method, we find the horizontal and vertical medians. We use these to divide the data into four groups.

We then count the number of observations in each group, we do not count any observations lying on the median lines.

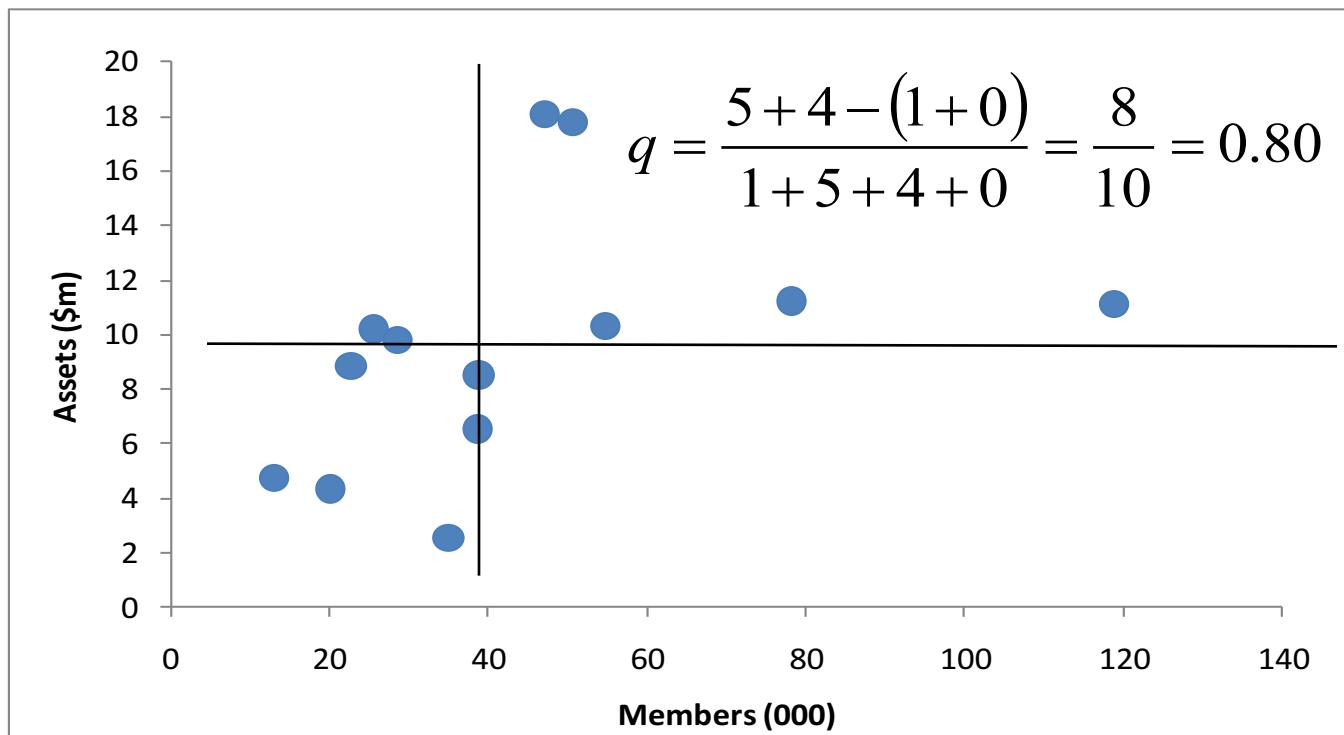
We then calculate the *q*-correlation as follows:

A	B
<hr/>	
C	D

$$q = \frac{N_B + N_C - (N_A + N_D)}{N_A + N_B + N_C + N_D}$$

Example q -Correlation

Using the Major Credit Unions data, and finding the medians by eye:



Pearson's r

Pearson's r is the most commonly used measure of correlation.

S_{xy} is the covariance of x and y .

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum xy - \frac{\sum x \sum y}{n}}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

Example Pearson's r

For our sample data we calculate the following sums and sums of squares: ($n = 13$). Know how to do this on your calculator!

x	y	xx	yy	xy
39	9	1521	81	351
78	11	6084	121	858
20	4	400	16	80
23	9	529	81	207
47	18	2209	324	846
25	10	625	100	250
51	18	2601	324	918
29	10	841	100	290
55	10	3025	100	550
39	7	1521	49	273
119	11	14161	121	1309
13	5	169	25	65
35	3	1225	9	105
573	125	34911	1451	6102

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$
$$= \frac{6102 - \frac{125 \times 573}{13}}{\sqrt{34911 - \frac{(573)^2}{13}} \sqrt{1451 - \frac{(125)^2}{13}}}$$
$$= \frac{592.38}{\sqrt{9654.92} \sqrt{249.07}} = 0.38$$

Excel/SYSTAT

Both programs have Pearson's r as a built in function.

In EXCEL use = CORREL(RANGE1, RANGE2)

In SYSTAT use the menu:

Graph > Plots > Scatterplot

Statistics > Correlations > Simple

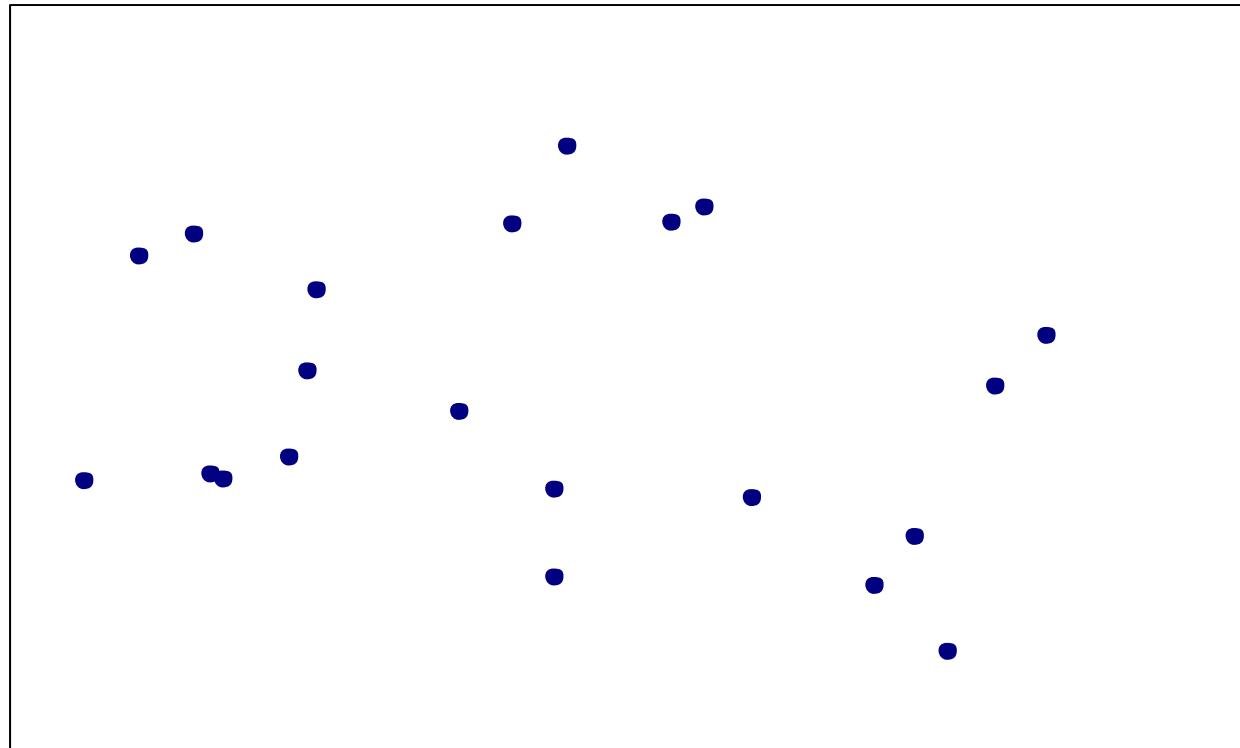
For multivariate data use:

Graph > Multivariate Displays > Scatterplot Matrix
(SPLOM)

Estimating r and q by eye

Practice using the file:

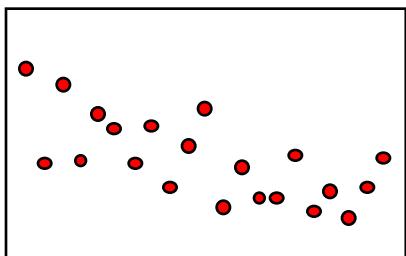
FIT1006 Lecture 7 Correlation.xlsx



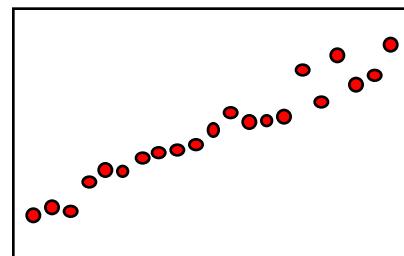
Interpreting Correlation 1

Having calculated the correlation coefficient, we can generally make a statement about the strength of the linear relationship that exists between variables.

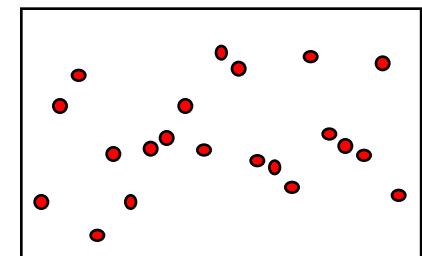
The sign of the coefficient (+ or -) indicates the direction of the trend. The closer to +/-1, the stronger the relationship.



$$r = -0.67$$



$$r = 0.97$$



$$r = 0.05$$

Interpreting Correlation 2

Some Cautions:

As correlation measures the linear relationship between variables, non-linear relationships tend to have low correlations.

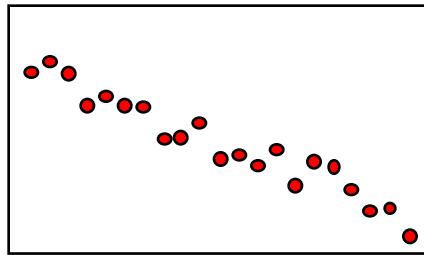
Bivariate data are subject to outliers which tend to decrease the value of correlation coefficient. (Omit outliers and re-calculate)

Correlation does not imply causation. Two variables may have a strong correlation but are not necessarily directly related. (They may be related by a third party)

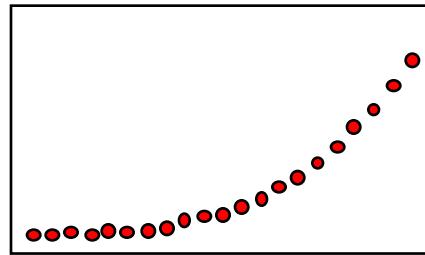
We tend to use correlation comparatively - that is one set of observations have a greater correlation than another set.

Should we or shouldn't we?

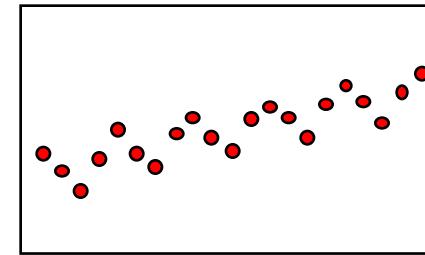
For which of the following data sets is Pearson's r an appropriate measure of correlation?



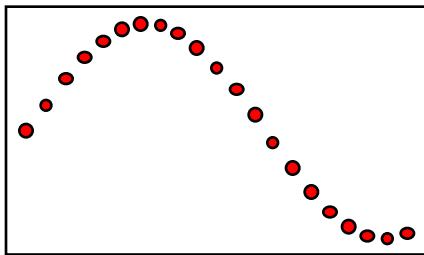
1



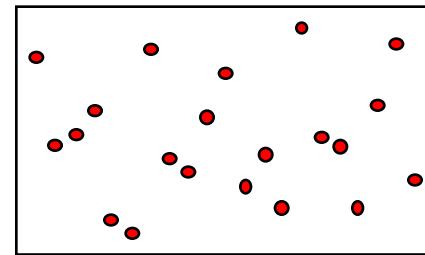
2



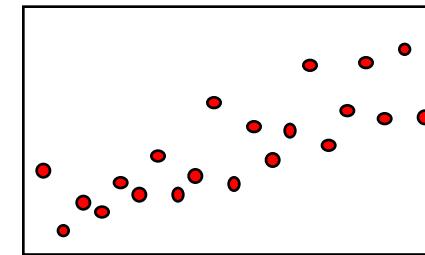
3



4



5



6

Scatterplots over multiple variables

For enrichment: go to <http://www.gapminder.org/>



Scatterplots over multiple variables

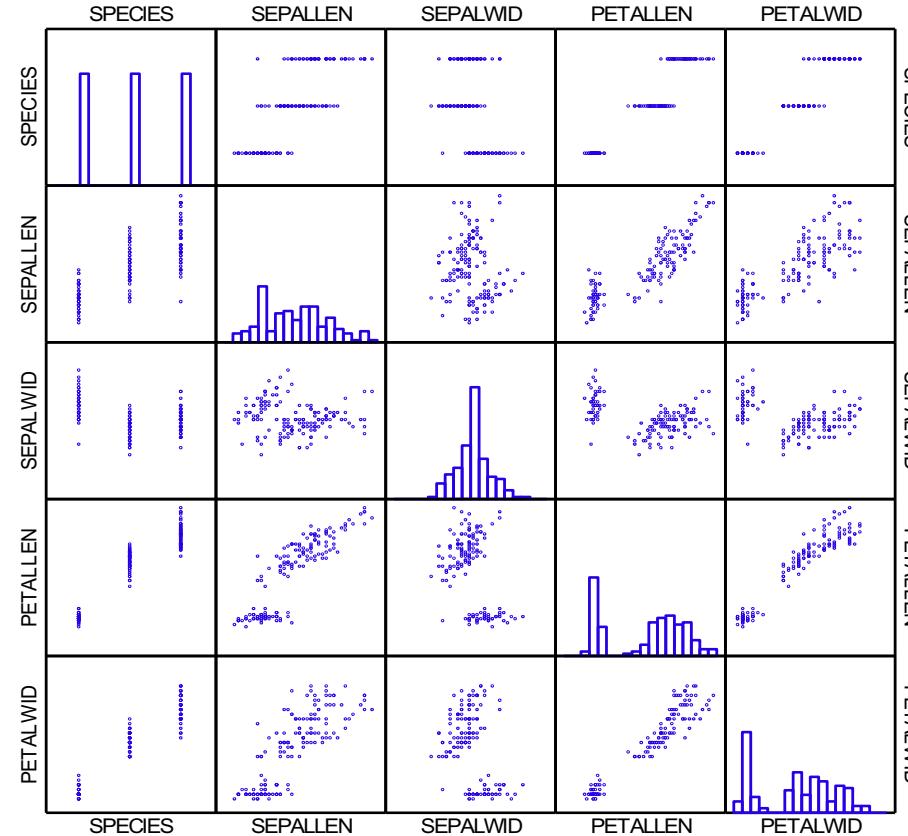
Multivariate display shows: Income, Life expectancy, Geographic region and Population over time.



The Iris Data: Scatterplot Matrix

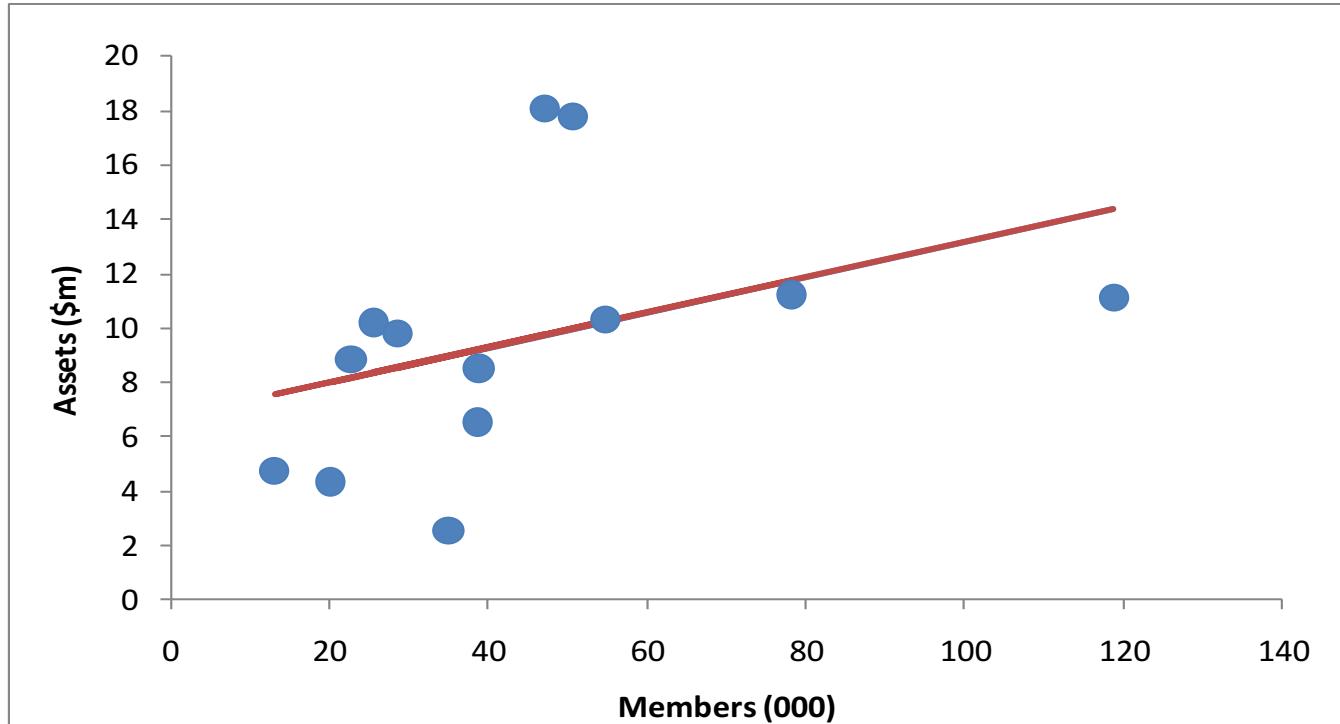
A famous data set. See Wikipedia. (<https://en.wikipedia.org/>)

Gives the sepal and petal width and length measurements for 3 species of iris.



Regression

The equation of the trend is the other important information that we get from bivariate data. This is covered next lecture.



FIT1006 Lecture 8 – Pre-reading

Linear Regression

The linear model,

Estimating coefficients,

Interpreting results, diagnostics.

Textbook:

7th Ed. Sections 15.1 - 15.4, 15.7, 16.1*, 16.2*.

*Additional reading on multiple regression.

Linear Regression

Regression is the practice of describing the relationship between 2 or more quantitative variables. Thus if we know the value of one variable, we can estimate the value of the related variable of interest.

Origin: The 19th century scientist Francis Galton collected data on the heights of fathers and their sons. He found that tall fathers had slightly shorter sons and that short fathers had slightly taller sons. Thus in each case there was a regression (reversion) to the mean. Over time the details of the investigation have been forgotten but the name has stuck to this method of modelling.

The purpose of regression

To find the underlying relationship between variables, to develop a model (or equation) of our data.

We may use the model for prediction or extrapolation.

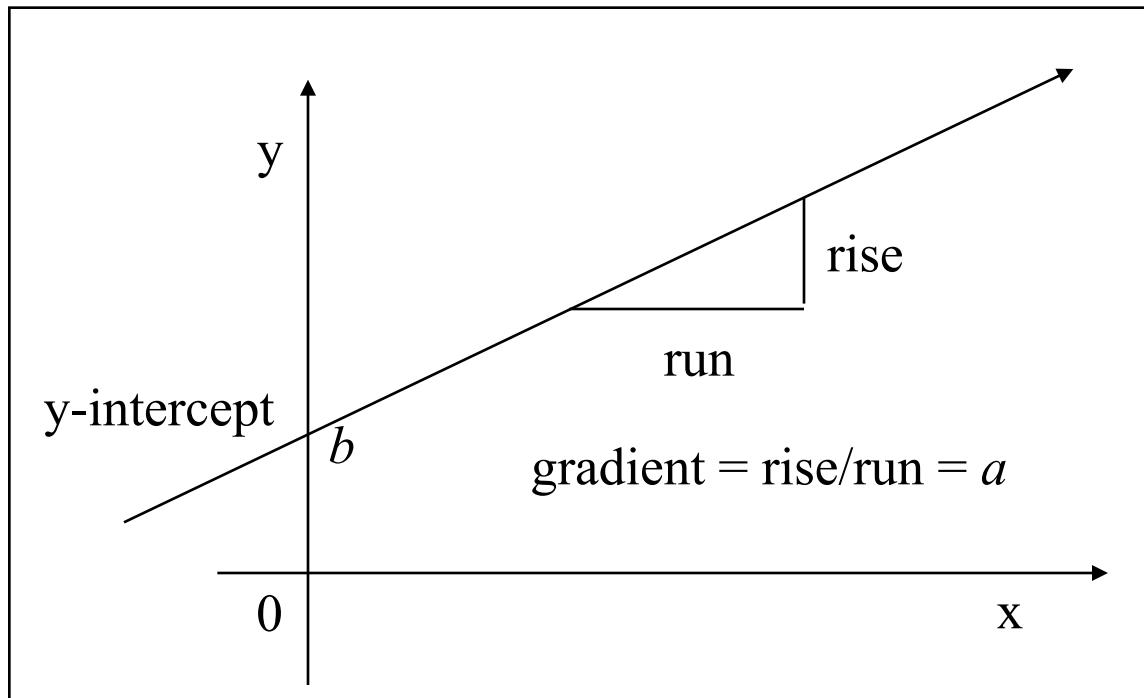
For example:

The industry standard for the number of staff employed given the annual turnover of the company

The time it takes to serve a customer given the number of items they have purchased, or the total value of items.

The equation of a straight line

We can use the basic equation of a straight line as the model for our regression equation. A line with gradient a and y -intercept b has equation $y = ax + b$.

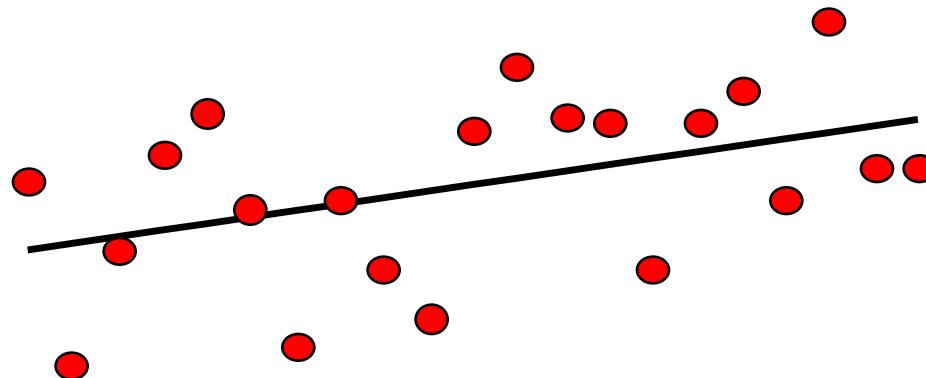


The underlying assumption

When we calculate the regression of y on x , we are assuming that the relationship between x and y is linear and thus we can say that $y = ax + b + e$, where e are random, normally distributed errors.

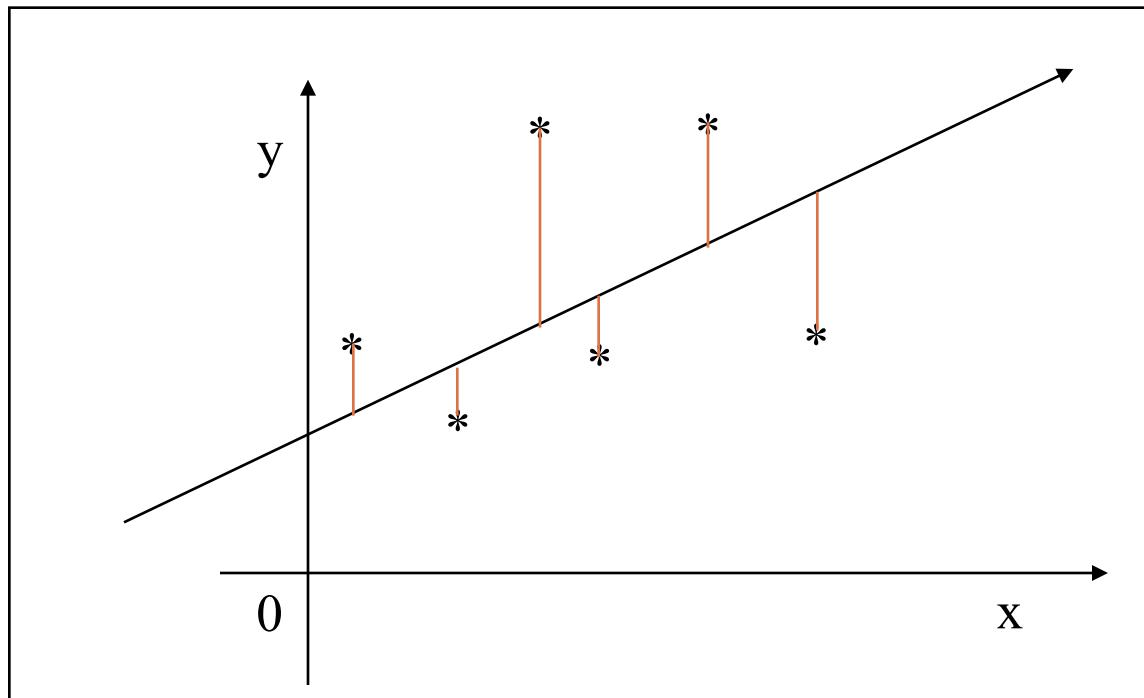
We want to find the value of a and b .

(Note that the textbook uses slightly different notation)



The basic idea

We want to minimise the sum of the squared errors, or differences between the fitted model (line) and the data.



Least Squares Regression

Ordinary Least Squares Regression seeks to minimise the sum of squared errors in the data.

To express the regression of y on x as $y = ax + b$, we calculate:

$$a = \frac{S_{xy}}{S_{x^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad b = \bar{y} - a\bar{x}$$

Sample Data

We will first model the relationship between Total Assets (\$m) and Number of Members (000) the for Major Credit Unions. (Source KPMG financial database)

We first use a small selection of data (13 pairs) and then consider the larger set.

See FIT1006 Lecture 7 & 8.xlsx

Members (000)	Assets (\$m)
39	9
78	11
20	4
23	9
47	18
25	10
51	18
29	10
55	10
39	7
119	11
13	5
35	3

Example

Using the same table as the previous lecture.

X	Y	XX	YY	XY
39	9	1521	81	351
78	11	6084	121	858
20	4	400	16	80
23	9	529	81	207
47	18	2209	324	846
25	10	625	100	250
51	18	2601	324	918
29	10	841	100	290
55	10	3025	100	550
39	7	1521	49	273
119	11	14161	121	1309
13	5	169	25	65
35	3	1225	9	105
573	125	34911	1451	6102

$$a = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$
$$= \frac{6,102 - \frac{573 \times 125}{13}}{34,911 - \frac{(573)^2}{13}}$$
$$= 0.06$$
$$b = \bar{y} - a\bar{x} = \frac{125}{13} - 0.06 \frac{573}{13} = 6.71$$

Interpreting the results

Thus we can describe our line of best fit as: $y = 0.06x + 6.71$.

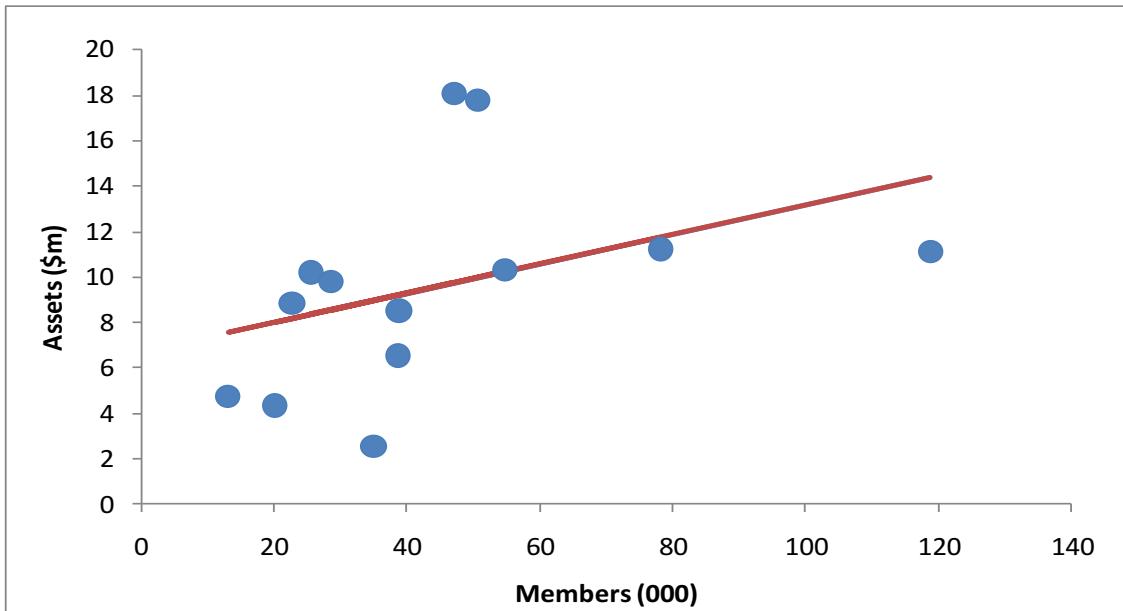
In terms of the data, we can say that “total assets (*in m*) of each of the major credit unions can be calculated as $0.06 \times \text{number of members} / 1,000 + 6.71$ ”. This is too precise for common-language discussion...

Rough rule of thumb: -----.

Benchmarking, if a company had ----- members, we would expect them to have ----- in assets.

Line of Best Fit

Using our equation, $y = 0.06x + 6.71$, we can draw a line of best fit for the data.



Members (000)	Assets (\$m)	Line of Best Fit
39	9	9.2
78	11	11.8
20	4	8.0
23	9	8.2
47	18	9.7
25	10	8.4
51	18	10.0
29	10	8.6
55	10	10.2
39	7	9.2
119	11	14.4
13	5	7.6
35	3	9.0

Slope	0.06
Y Intercept	6.71

Discussion of the model

Note that the model fits the main group quite well.

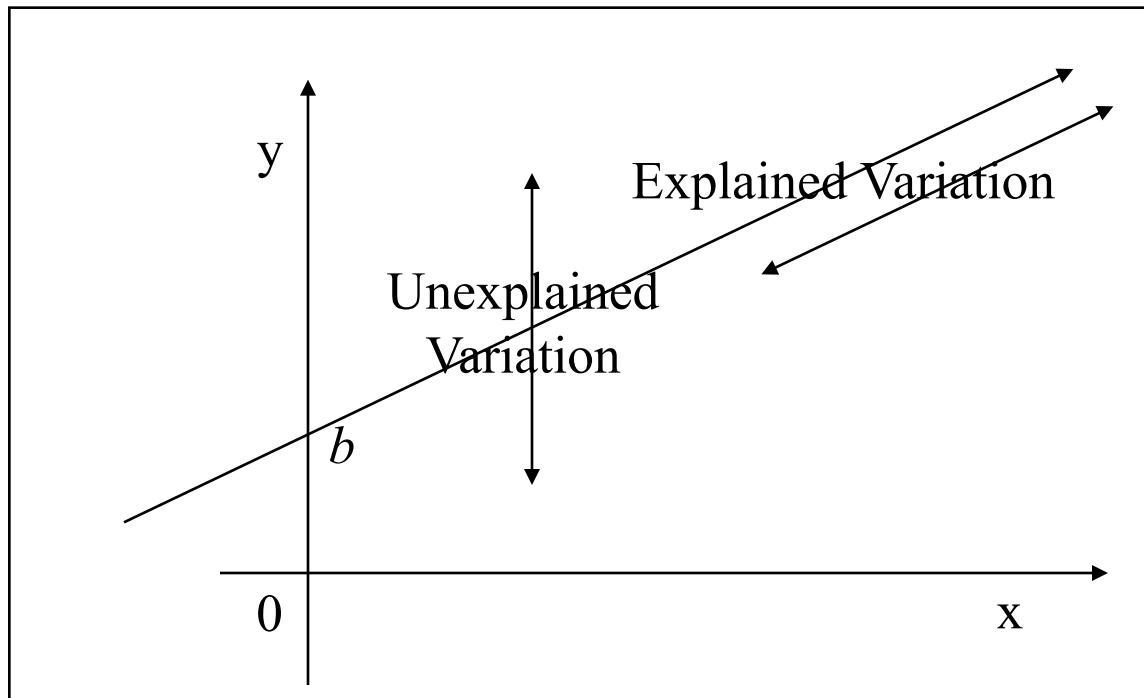
Observe that there are several outliers, in practice one would identify which companies these referred to.

The model predicts that even with no members, a credit union would have about \$6m in assets.

Note that the largest x value, at the end of the data cluster tends to exert undue influence (*leverage*) on the LOBF.

How good is the fit?

One measure of how well the regression model is the proportion of variation in y that is explained by the regression equation.



Coefficient of Determination

- The coefficient of determination is the proportion of variation in y that is explained by variation in x through the regression equation.
- The coefficient of determination is r^2 - the square of Pearson's correlation coefficient r .
- We generally calculate $100 r^2$ and express the coefficient of determination as a percentage.
- In the last lecture we calculated the correlation of the data as $r = 0.38$. By calculating r^2 we can say that the model explains 15% of the variation in y .

Regression in EXCEL

Regression is one of the built in analysis functions in EXCEL

You can also calculate the formulas manually with the following formulas:

$$a = \text{SLOPE}(y \text{ values}, x \text{ values})$$

$$b = \text{INTERCEPT}(y \text{ values}, x \text{ values})$$

also

$$r^2 = \text{CORREL}(y \text{ values}, x \text{ values})^2$$

Regression in SYSTAT

SYSTAT calculates regression and gives a diagnostic output of the fitted model. (*We will not be considering all diagnostic output at this stage*).

Select: Statistics > Regression > Linear > Least Squares

ASSETS_M is the dependent variable (the one we are trying to predict).

MEMBERS_000 is the independent variable – the one that is free to change...

The model and residuals can be saved to a data file.

SYSTAT Output (a) report

Dependent Variable	ASSETS_M
N	48
Multiple R	0.425
Squared Multiple R	0.181
Adjusted Squared Multiple R	0.163
Standard Error of Estimate	4.657

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std.		Tolerance	t	p-Value
			Coefficient	Tolerance			
CONSTANT	5.318	1.394	0.000	.	3.814	0.000	
MEMBERS_000	0.108	0.034	0.425	1.000	3.185	0.003	

SYSTAT Output (a) report

Analysis of Variance

Source	SS	df	Mean Squares	F-Ratio	p-Value
-----+-----					
Regression	220.099	1	220.099	10.147	0.003
Residual	997.818	46	21.692		

*** WARNING *** :

Case 11 has large Leverage (Leverage : 0.386)

Durbin-Watson D-Statistic | 1.881
First Order Autocorrelation | 0.058

Information Criteria

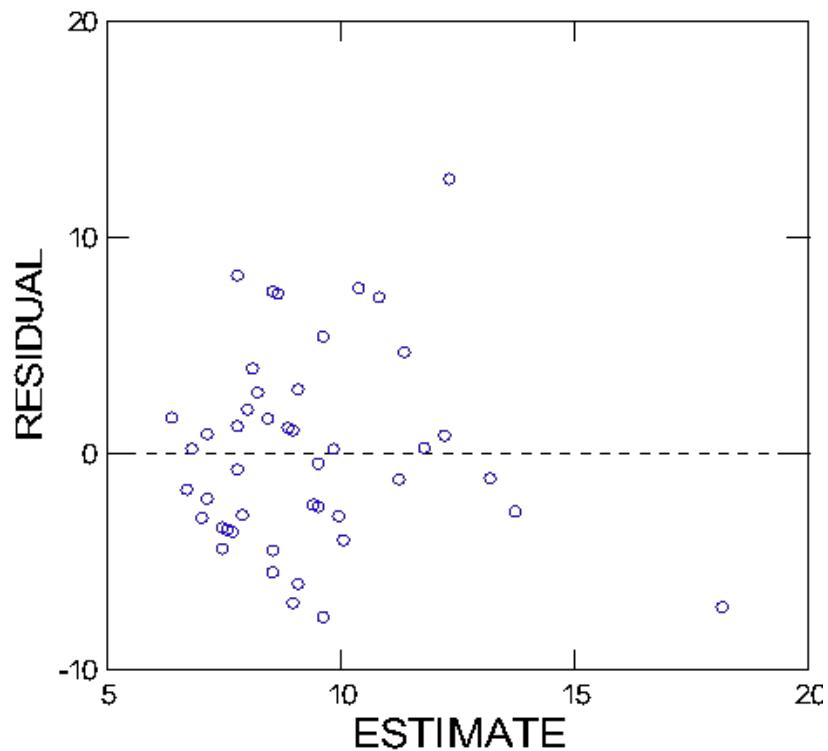
AIC | 287.868
AIC (Corrected) | 288.413
Schwarz's BIC | 293.481

Residuals have been saved.

SYSTAT Output (b) residuals

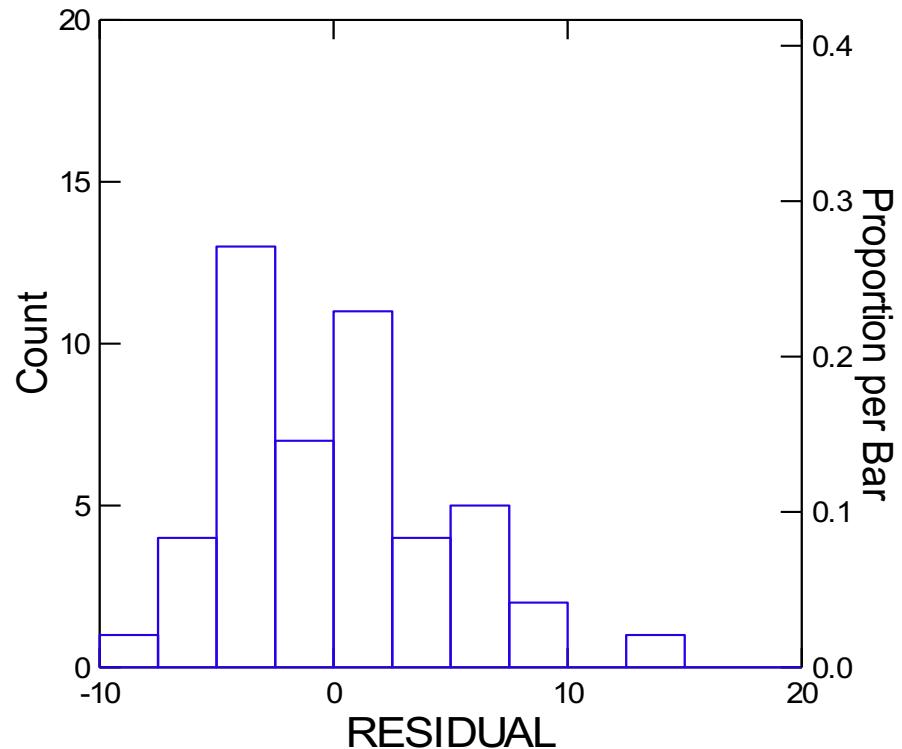
SYSTAT also has the option to calculate and save residuals, which are the difference between fitted values and the actual data. Residuals should be Normally distributed under the assumptions of the linear model. On the right is a scatterplot of the residuals.

Plot of Residuals vs. Predicted Values



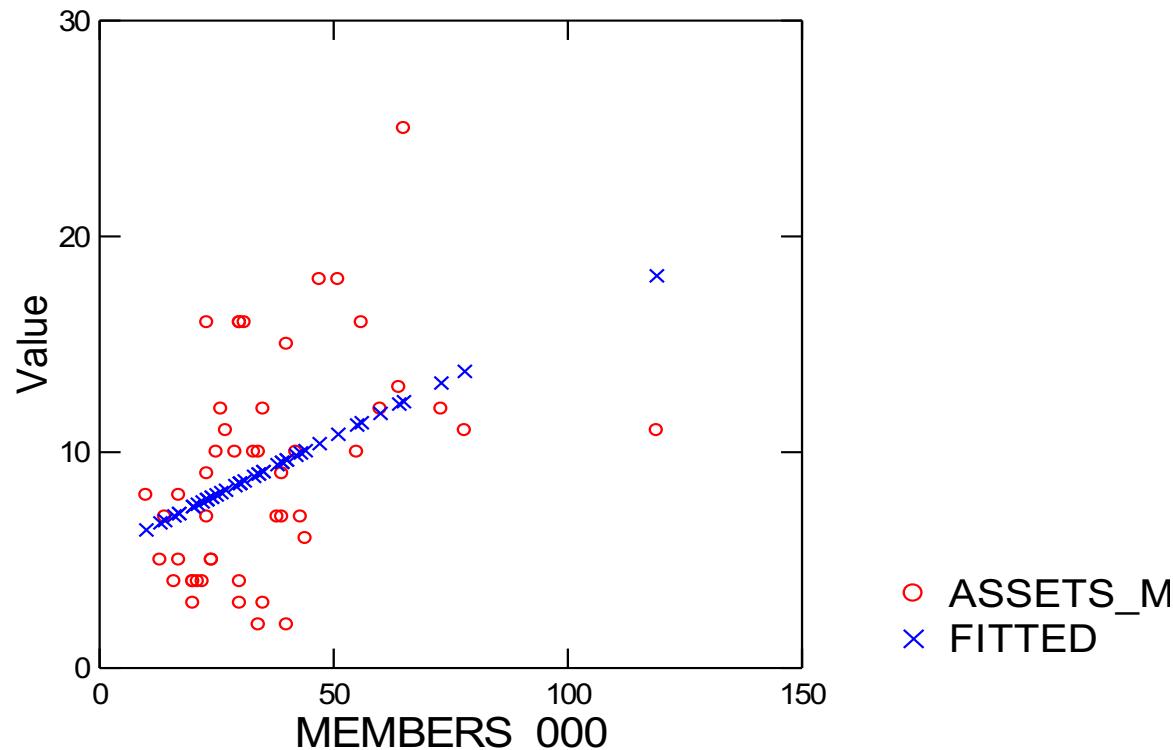
SYSTAT Output (c) residuals

A histogram of the residuals shows that they are (approximately) normally distributed. Note the outlier.



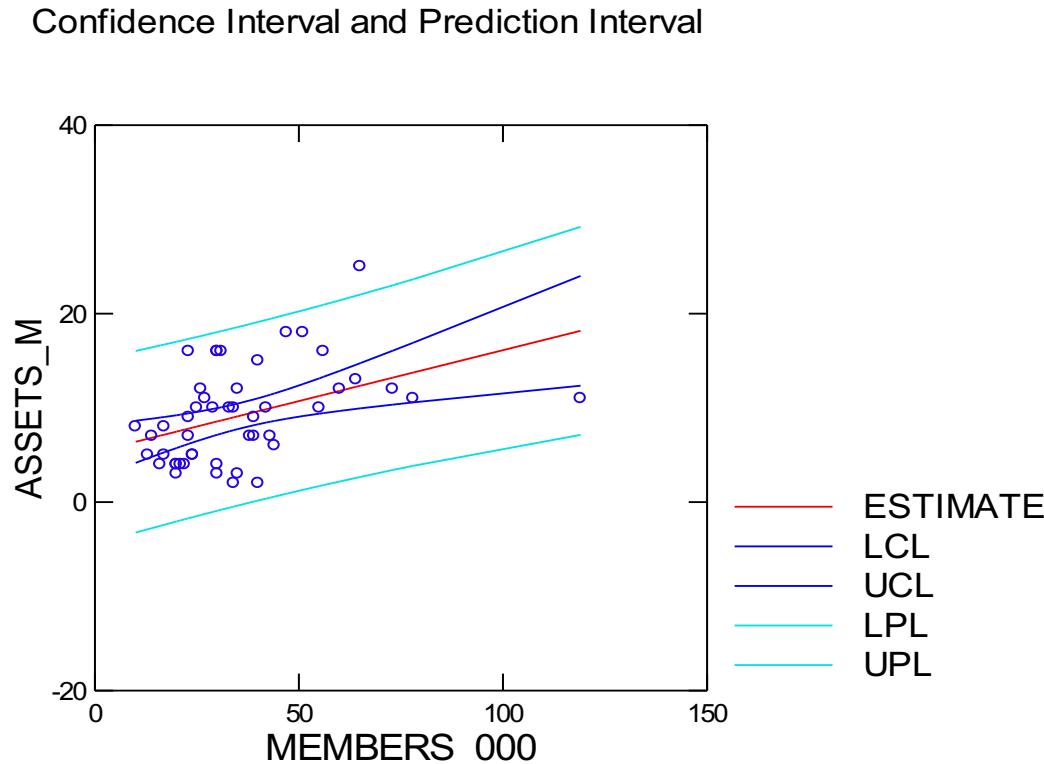
SYSTAT Output (d) model

A scatterplot of the original data with the a plot of the fitted model
(also saved as a separate file when the regression was calculated).



SYSTAT Output (e) Prediction CI

The prediction confidence interval shows graphically how the predicted model becomes less accurate at the extremes of the prediction interval.



One More Thing

If we interchange x and y of our model we get a different regression equation, not just the inverse equation.

Why?

The following material is optional.

Multiple Regression

Sometimes we may want to predict the value of a variable as a function of a number of input variables in order (usually) to increase the accuracy of the model.

In the following example we will first determine the number of employees as a function of assets only using Ordinary Least Squares Regression.

We will then determine the number of employees of a credit union as a function of both the credit union's assets and the number of members. This procedure is called multiple regression, whereby the value of y is predicted using two variables: x_1 and x_2 .

OLS Regression in SYSTAT

Dep Var: EMPLOYEES N: 48 Multiple R: 0.332 Squared multiple R: 0.110

Adjusted squared multiple R: 0.091 Standard error of estimate: 50.690

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	49.244	15.172	0.000	.	3.246	0.002
ASSETS_M	3.464	1.453	0.332	1.000	2.384	0.021

*** WARNING ***

Case 11 is an outlier (Studentized Residual = 7.314)
Case 21 has large leverage (Leverage = 0.227)

Durbin-Watson D Statistic 1.935

First Order Autocorrelation 0.020

Multiple Regression in SYSTAT

Dep Var: EMPLOYEES N: 48 Multiple R: 0.906 Squared multiple R: 0.821

Adjusted squared multiple R: 0.813 Standard error of estimate: 23.012

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-1.597	7.871	0.000	.	-0.203	0.840
MEMBERS_000	2.470	0.185	0.932	0.818	13.349	0.000
ASSETS_M	-0.694	0.729	-0.066	0.818	-0.952	0.346

*** WARNING ***

Case 11 has large leverage (Leverage = 0.433)

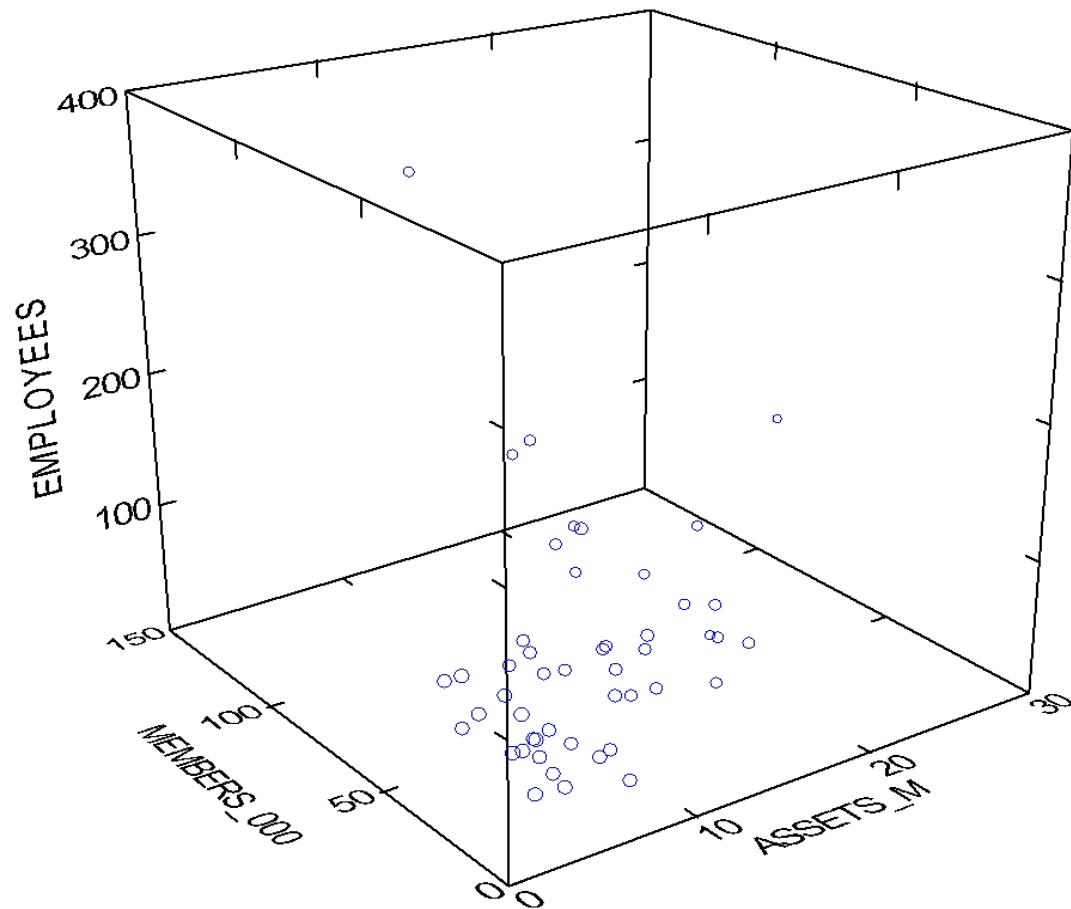
Case 11 is an outlier (Studentized Residual = 3.465)

Durbin-Watson D Statistic 2.163

First Order Autocorrelation -0.094

Multiple Regression in SYSTAT

A 3D plot of the data showing the main cluster and several outliers.



Differences between models

y = employees, x_1 = assets(m), x_2 = members(000)

OLS Regression model: $y = 3.4 x_1 + 49.2$

$r = 0.33, r^2 = 11\%$ 2 exceptional observations.

Multiple Regression Model $y = -0.7 x_1 + 2.5 x_2 - 1.6$

$r = 0.91, r^2 = 82\%$ 1 exceptional observation.

Comparing models we see that including membership gives a better prediction of the number of employees than assets alone.

FIT1006 Lecture 9 – Pre-reading

Introduction to Probability

- The probability of an event.
- Set notation and set operations for probability.
- Probability distributions.
- The mean and variance of a probability distribution.

Textbook:

7th Ed. Sections 6.1, 7.1 – 7.3.

Why Study Probability?

So far we have used descriptive statistics to summarise collections of data.

To make a definitive comparison between groups of data we need a way of measuring the reliability that our conclusion is correct: probability.

Probability Theory is the formal method for determining the likelihood of randomly occurring events.

We will first look at probability as a study in its own right and then use it as the basis for the technique of hypothesis testing .

Some terms...

Probability theory uses the term ‘experiment’ to describe an activity that leads to a single well defined result or outcome.

Rolling a die is an experiment.

The number showing uppermost is the outcome.

An event is a combination of one or more outcomes:

Eg. Two dice are thrown. An event might be that the total shown is 7, which can be obtained from a variety of outcomes such as (1,6) (2,5) (3,4) etc.

Random Variable

When we conduct an experiment, we are interested in observing some outcome or event.

We refer to the value of the event as a random variable.

Some examples of random variables are:

The number of customers entering a shop during one hour.

The number shown on one die.

The number of red cars in the student car park today.

The amount of money in a person's bank account.

Sample Space

The set of all possible outcomes of an experiment is described as the sample space.

The sample space for tossing a coin:

$$\text{SS} = \{\text{Head}, \text{Tail}\}$$

The sample space for throwing a die is:

$$\text{SS} = \{1, 2, 3, 4, 5, 6\}$$

The Probability of an Event

The probability of an event is the chance that it will occur.

To calculate the probability of a desired event we determine the number of outcomes resulting in the event we are interested in as a proportion of the of all possible events.

$$\text{Probability of an event} = \frac{\text{Number of desired outcomes}}{\text{All outcomes}}$$

A probability of 0 means an event will never occur.

A probability of 1 means an event will always occur.

Example

We throw two dice and define a random variable x as the sum shown on the uppermost faces. What is the probability of *throwing a 7*?

Sample Space:

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

$$P(x = 7) = \frac{6 \text{ outcomes resulting in } 7}{36 \text{ outcomes in total}} = \frac{1}{6}$$

Determining Probability

Objective Probability

- Limit of relative frequency - based on an analysis of repeated outcomes.
- Logical deduction - (previous slide) - based on an analysis of all outcomes. (Sometimes called the Classical Method)
- Empirically - based on historical observations.

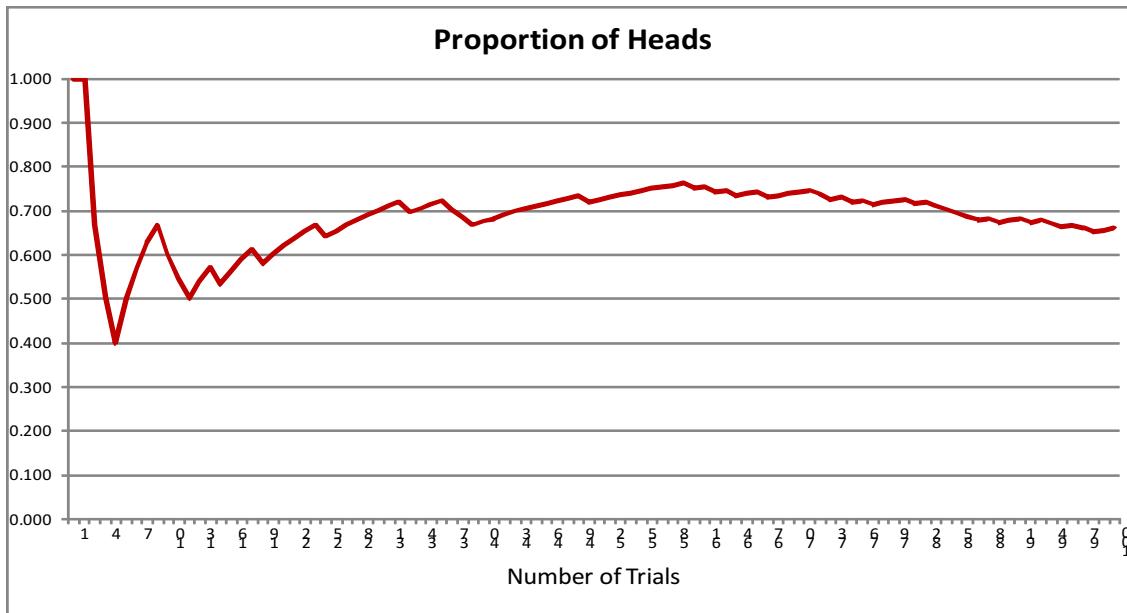
Subjective Probability

- Based on personal assessment using intuition or judgement. This is a method often employed for business decision making.

Limit of Relative Frequency

We can calculate the probability of throwing heads in one toss of a *biased* coin by repeated tossing of a coin. We calculate the average number of heads (heads/total tosses) thrown at each toss. (1 = head, 0 = tail)

The graph below shows the running average over 100 tosses.



Sets

We define a set as a collection of objects, often related by some common property, and often use them to describe events or sample spaces.

Notation. Use $\{ \}$ to mean '*the set of*'

Let A represent the outcomes of throwing a die, then $A = \{1, 2, 3, 4, 5, 6\}$

Let Y represent the set of even numbers, then

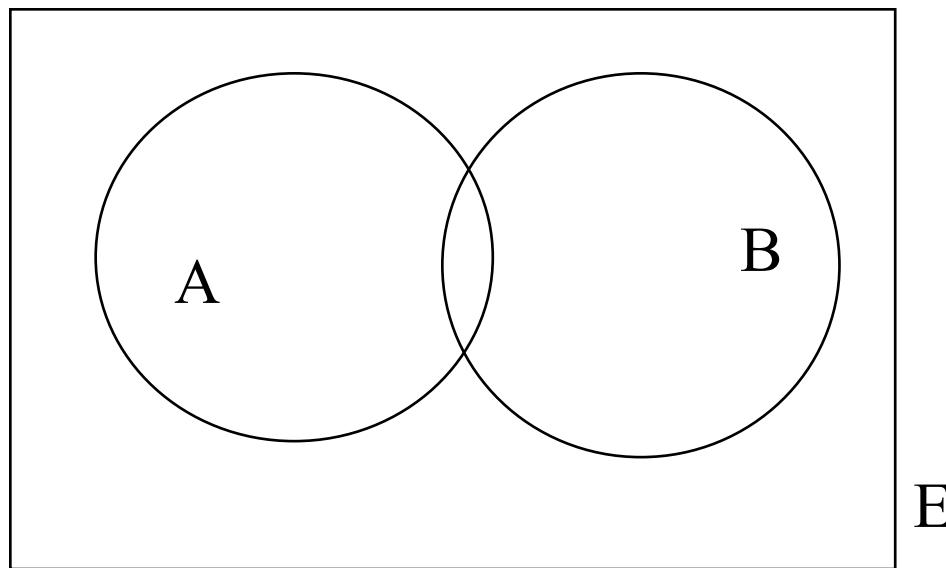
$$Y = \{\dots, -4, -2, 0, 2, 4, 6, \dots\}$$

By convention we refer to the set of all outcomes, or the sample space as the *Universal Set (E)*.

Venn Diagrams

Named after John Venn, these diagrams are a standard way of representing sets.

We can show intersection, union, complementary sets etc.



Relationships Between Sets

Having defined sets, we can define several standard relationships between sets.

These are:

Disjoint Sets

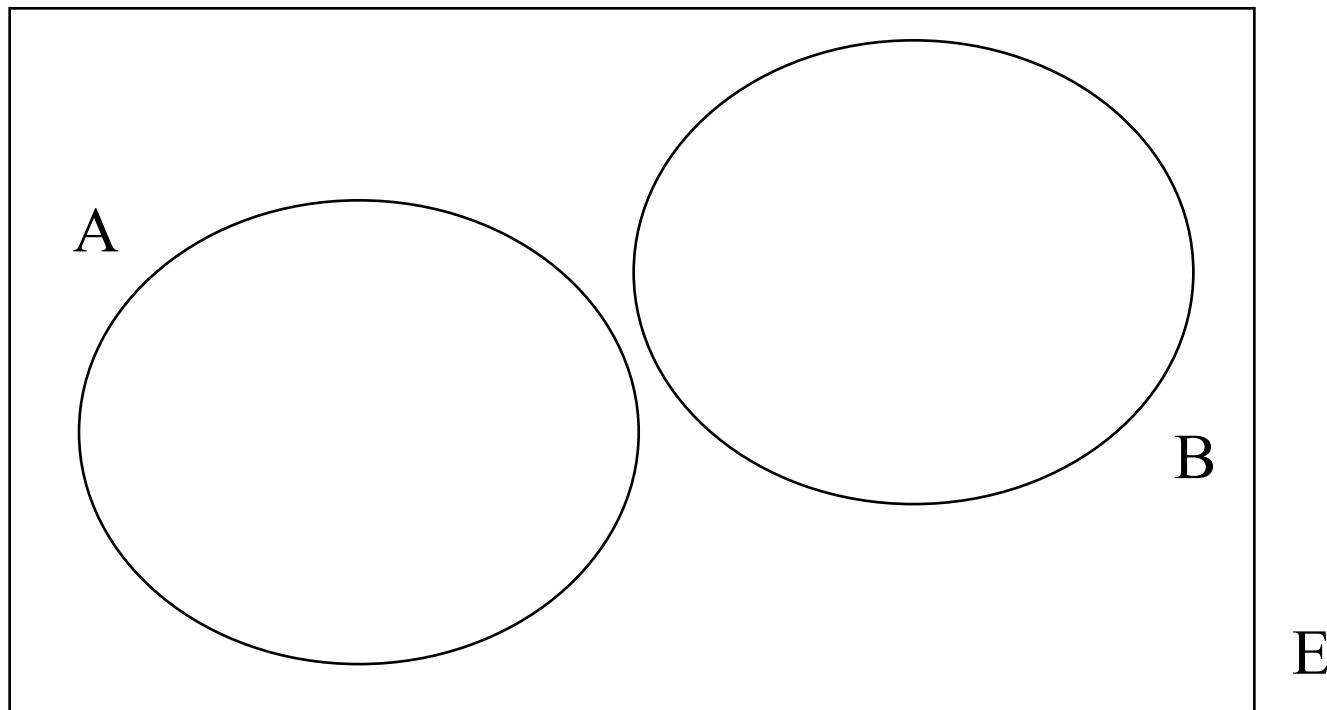
Complement

Union

Intersection

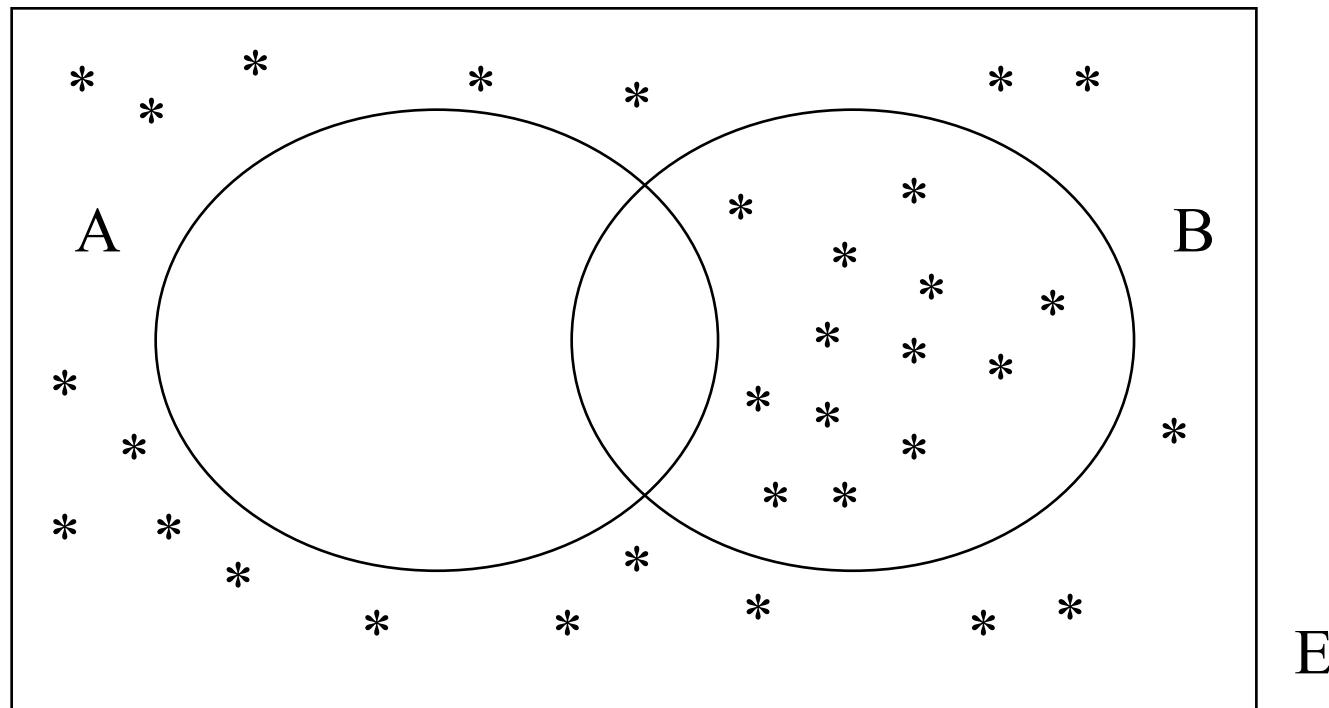
Disjoint Sets

“There is no intersection of A and B”



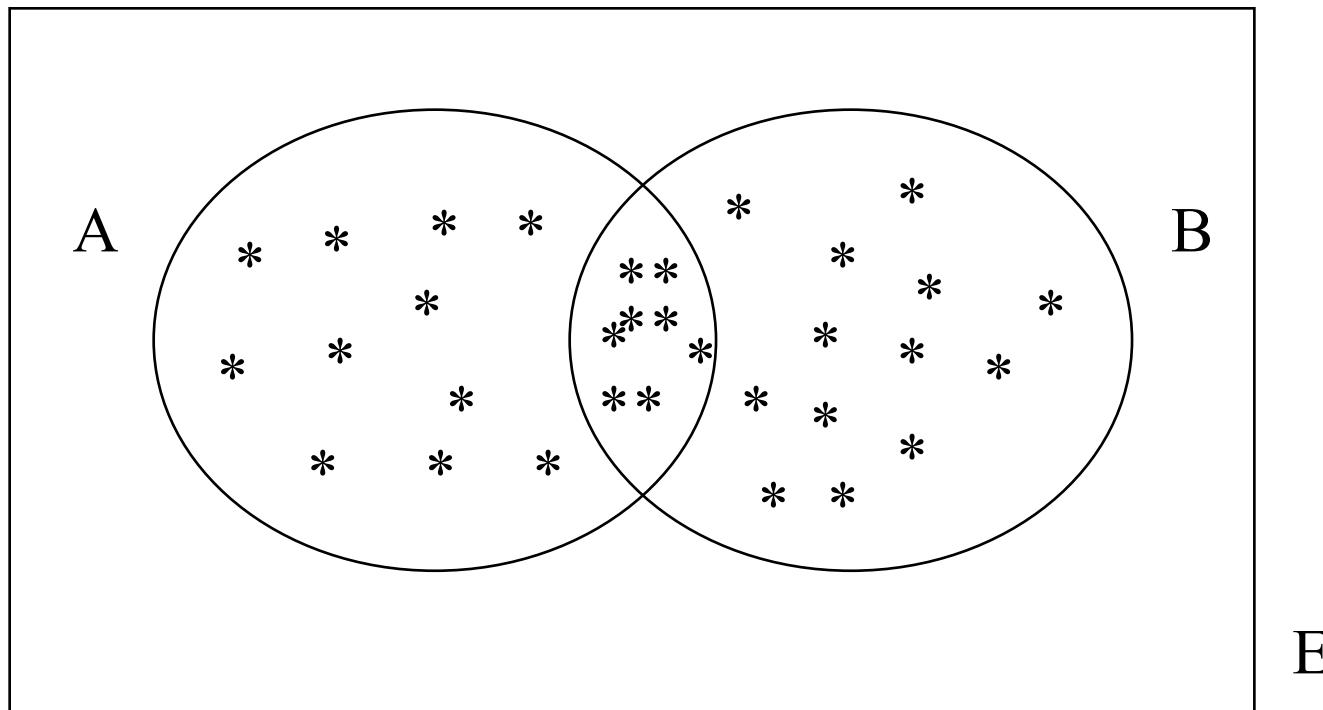
Complement \bar{A}

“*Everything not in A*”



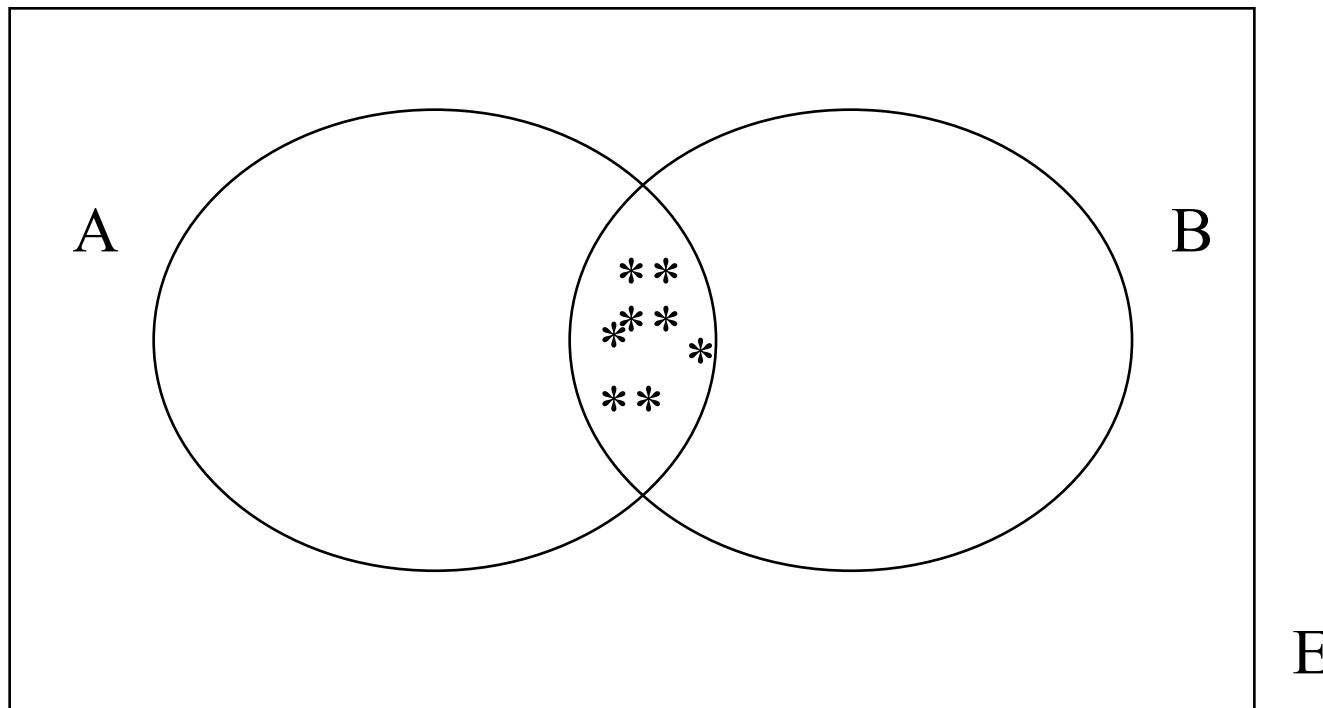
Union $A \cup B$

“Everything in A or B”



Intersection $A \cap B$

“Everything in A and B”



Rules of Probability

Based on the relationship between sets, we can define the standard rules of probability.

Probability of complementary events:

If the probability that an event occurs is $P(A)$, Then the probability that the event does not occur is $P(\bar{A})$ where $P(\bar{A}) = 1 - P(A)$.

The addition rule for the union of events:

If events A and B have probabilities of occurring respectively $P(A)$ and $P(B)$,
Then the probability of A or B occurring is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Example

We throw two dice and define a random variable x as the sum shown on the uppermost faces.

What is the probability of *throwing a 7* ?

What is the probability of not *throwing a 7* ?

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

$$P(X = 7) = \frac{1}{6}$$

$$P(\overline{X = 7}) = 1 - \frac{1}{6} = \frac{5}{6}$$

Example

We throw two dice. What is the probability of
(A) *throwing a 7 or* (B) *one die showing 6* ?

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

$$P(A) = \frac{6}{36}$$

$$P(B) = \frac{10}{36}$$

$$P(A \cap B) = \frac{2}{36}$$

$$P(A \cup B) = \frac{6}{36} + \frac{10}{36} - \frac{2}{36} = \frac{14}{36}$$

Expected Value of a random variable

Having defined a random variable, we can determine the expected, or average value that the variable would have. This is similar to the mean for statistical data.

Generally speaking, we have a random variable X , which can take on a range of values, x_i , each with a certain probability of occurring, $P(x_i)$.

The expected value, $E(X)$, (or the mean) is then the sum of outcomes multiplied by their respective probabilities of occurrence, thus:

$$E(X) = \mu = \sum_{i=1}^n x_i P(x_i)$$

E. V. of a Function of a R.V.

In the same way as we determine the expected value of a random variable as the sum of all outcomes multiplied by their probability, the expected value of a function of a random variable is similarly determined. thus:

Let $y = f(X)$, then

$$E(Y) = \sum_{i=1}^n f(x_i)P(x_i)$$

Variance of a random variable

In the same way that the variance of statistical data is the ‘average’ squared deviations, we can calculate the variance and standard deviation of a random variable.

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

Giving the computational formula of:

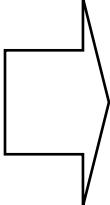
$$= \sum_{i=1}^n x_i^2 P(x_i) - \mu^2$$

Example

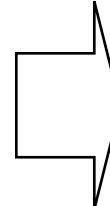
Three coins are tossed. Let X denote the number of heads showing.
Calculate the mean and variance of X .

Our set of outcomes is $\{\text{hhh}, \text{hht}, \text{hth}, \text{htt}, \text{thh}, \text{tht}, \text{tth}, \text{ttt}\}$ (see next slide). To calculate the mean and variance of X we first need to write out the probability distribution of X .

X	P(x)
0	0.125
1	0.375
2	0.375
3	0.125

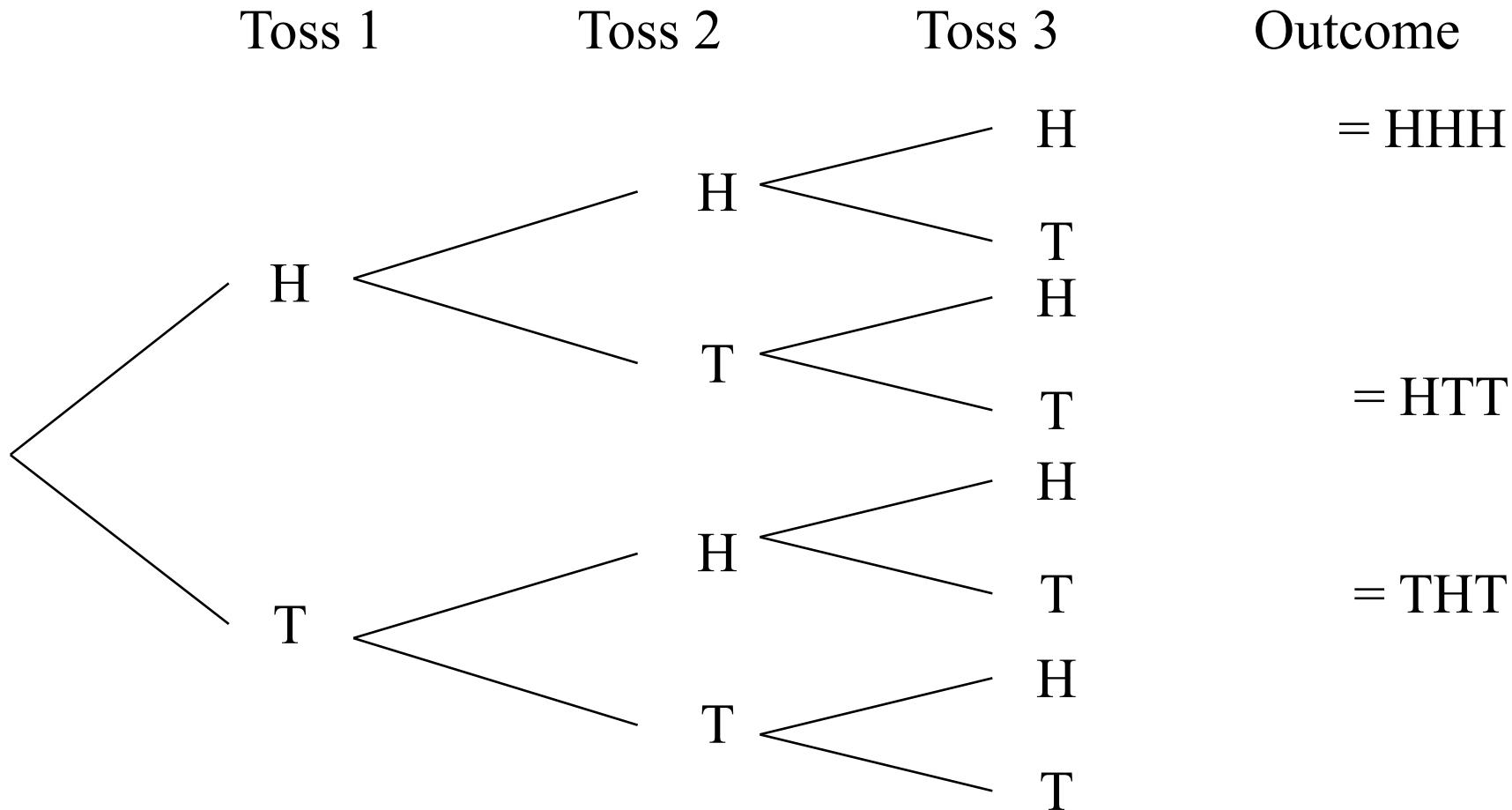


X	P(x)	$X \cdot P(X)$	$X^2 \cdot P(X)$
0	0.125	0.000	0.000
1	0.375	0.375	0.375
2	0.375	0.750	1.500
3	0.125	0.375	1.125
		1.500	3.000



$$\begin{aligned} \text{Thus } E(X) &= 1.5 \\ E(X^2) &= 3 \\ \text{Var}(X) &= 3 - 1.5^2 \\ &= 3 - 2.25 \\ &= 0.75 \end{aligned}$$

Three Coin Tosses



Independent Events

If two events are independent then the probability of either event occurring has no effect on the probability of the other event occurring.

For two independent events A and B,

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

For tosses of a coin, let A be the outcome of a head with the first toss and B the outcome of a head with the second toss. Then $P(A \cap B) = P(A) * P(B) = 0.5 * 0.5 = 0.25$

Question (for you to finish)

Two dice are thrown. Let Y be the sum of faces uppermost.

- 1 Write out the table of throw results
- 2 Write out a table of sums (Y)
- 3 Write down the probability distribution of Y
- 4 Calculate the mean and variance of Y
- 5 Let Z be the product of faces uppermost.
- 6 Do the same for Z .

Distribution of X (sum of faces)

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

x	f	p(x)	x*p(x)	x**p(x)
2	1	0.03	0.06	0.11
3	2	0.06	0.17	0.50
4	3	0.08	0.33	1.33
5	4	0.11	0.56	2.78
6	5	0.14	0.83	5.00
7	6	0.17	1.17	8.17
8	5	0.14	1.11	8.89
9	4	0.11	1.00	9.00
10	3	0.08	0.83	8.33
11	2	0.06	0.61	6.72
12	1	0.03	0.33	4.00
	36		7.00	54.83
		mean =	7.00	
		var =	5.83	
		stdev =	2.42	

FIT1006 Lecture 10 – Pre-reading

Probability continued:

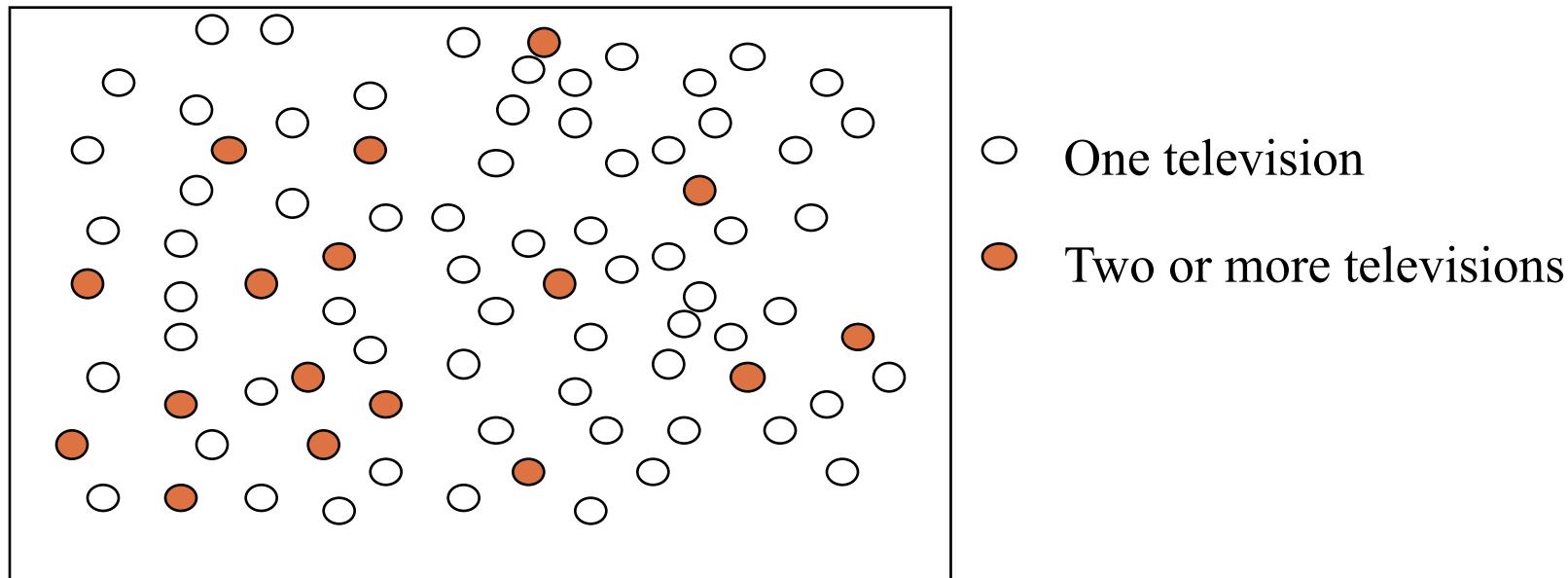
- Independent and conditional events.
- Probability trees.
- Bayes' Theorem.
- Background mathematics for probability distributions.

Textbook:

7th Ed. Sections 6.2 – 6.6.

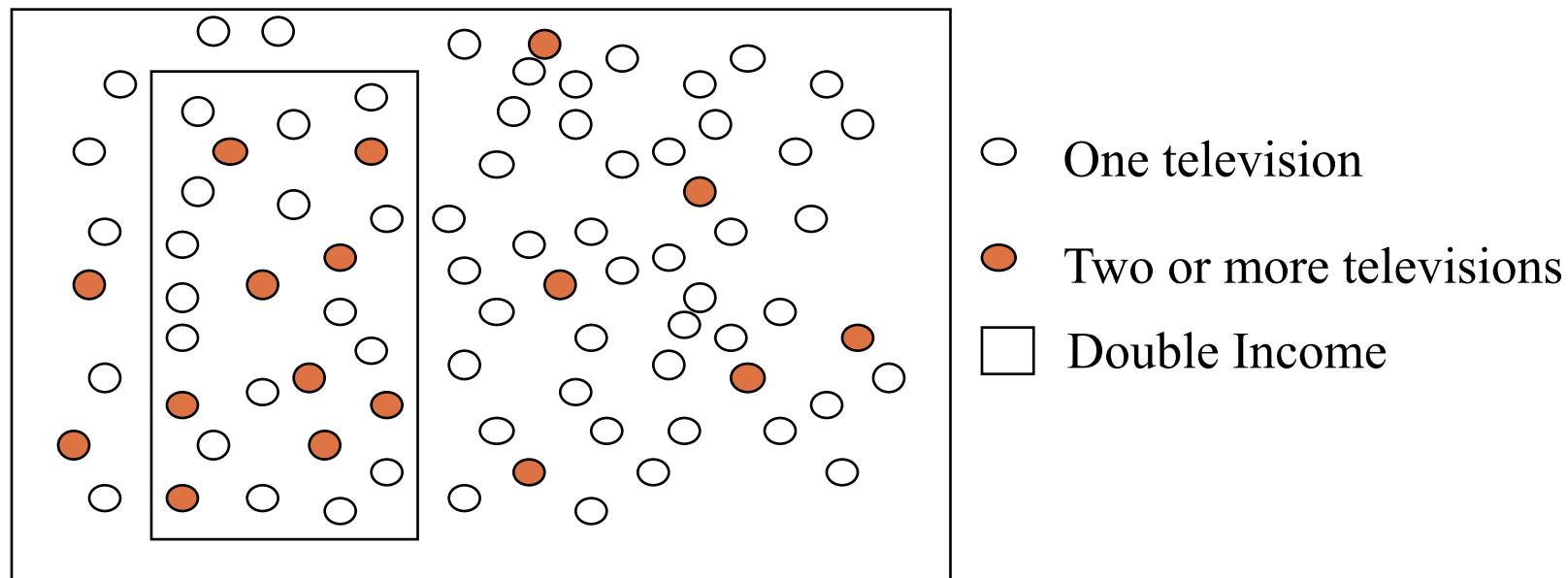
Sample (Survey) Data

We are interested in predicting the number of televisions in a household. From a survey of households we find that 17 out of 82 households have two or more televisions. Thus the probability that a household chosen at random will have two or more televisions is 0.21.



Additional Information

When we identify the families with a double income, we see that the probability of having two or more televisions is now 9 out of 25 or 0.36. *We can refine our estimate of the distribution of televisions with some extra information. Bayes' Theorem is a systematic way of doing this.*



Conditional Probability

Using D to represent a family having a double income and T to represent a family having two or more televisions we can write the following.

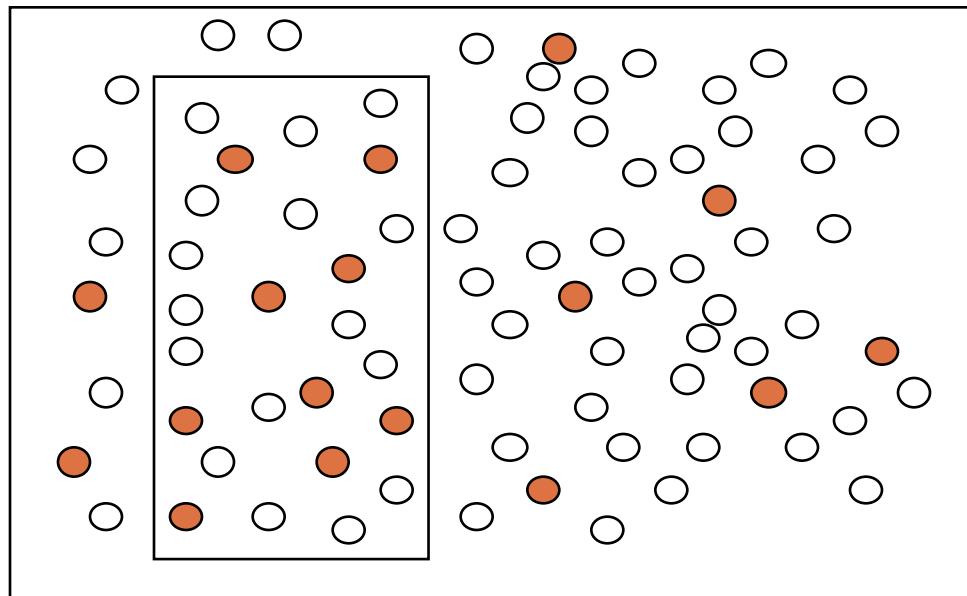
$$P(D) = 25/82 = 0.30 \text{ and } P(T) = 17/82 = 0.21$$

We can also write: $P(T | D) = 9/25 = 0.36$

We express this last probability as “The probability of having two or more televisions given that the family has a double income.” This is a conditional probability because T is conditioned on (affected by) D.

Conditional Probability

We can define a conditional probability more formally as the probability that one event occurs (is true) given that another event has occurred (is true).



We define the conditional probability of B given A as:

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

Conditional Probability

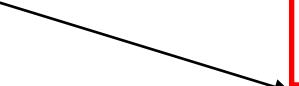
Using D to represent a family having a double income and T to represent a family having two or more televisions.

We can write: $P(T | D) = P(T \cap D)/P(D) = (9/82)/(25/82) = 0.36$

The content of the original diagram can be abstracted as a

Contingency Table. The highlighted cells refer to the case above.

D' means D complement



	T	T'	Σ
D	9	16	25
D'	8	49	57
Σ	17	65	82

Independent Events 1

We describe events as being independent when the outcome of one event does not affect the outcome of the others.

For example, if I toss two dice the number showing on one is independent of the number showing on the other.

Put in another way, the probability of getting a 6 on the second die is unaffected by throwing a 6 on the first.

The case of television ownership and family income was an example of events which were not independent.

When events are not independent, there exists an opportunity to make a refined estimate of a probability based on whether or not another event has occurred.

Independent Events 2

Put more formally, if two events are independent then the probability of either event occurring has no effect on the probability of the other event occurring.

For two independent events A and B,

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

and $P(A \cap B) = P(A) * P(B)$

For tosses of a coin, let A be the outcome of a head with the first toss and B the outcome of a head with the second toss.

Then $P(A \cap B) = P(A) * P(B) = 0.5 * 0.5 = 0.25$

Mutually Exclusive Events

Mutually exclusive events are events which cannot all occur in the same trial.

For example, a person tosses a coin, the outcome of head and tails is mutually exclusive. The event that a coin lands heads up means that the event of tails up cannot occur.

A person may choose product A or B

A family may have a single or double income.

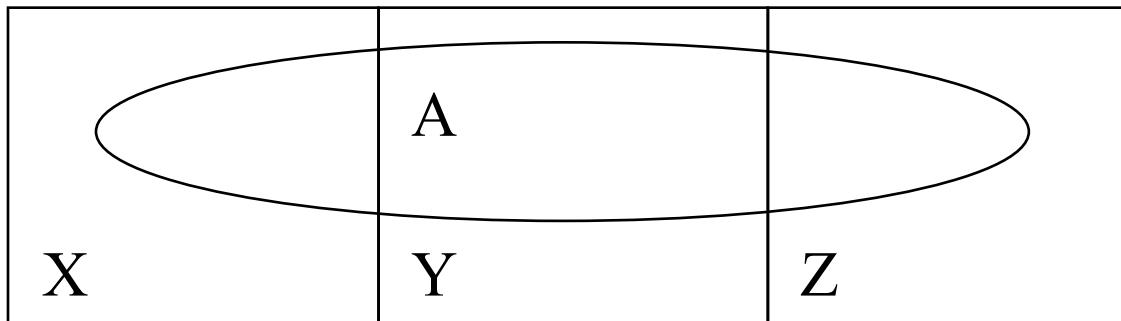
For two mutually exclusive events (or outcomes)

A and B, $P(A \cap B) = 0$.

The Law of Total Probability

The sets below represent the 3 outcomes X, Y and Z as well as the outcome A. Then X, Y and Z are *mutually exclusive* and *collectively exhaustive* because they do not intersect and together they cover the total sample space (or universe).

Let $A = (A \cap X) \cup (A \cap Y) \cup (A \cap Z)$. Then by the law of total probability: $P(A) = P(A \cap X) + P(A \cap Y) + P(A \cap Z)$.



• • •

One last piece of theory before we look at Bayes' Theorem.

We can re-express the conditional probability formula to put the intersection of sets as the subject of the equation.

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

thus

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

Bayes' Theorem

Bayes' Theorem provides a formal method for updating the probability of an event when the occurrence of that event is affected (conditional) on some other event.

The stages of a Bayesian problem:

- 1 Start with the *Prior* probability – this is the probability of an event in the absence of any other information. Sometimes called the state of nature.
- 2 Receive additional information as conditional probabilities.
- 3 Update the Prior probability using the additional information to determine the *Posterior* probability. The following slides show how!

• • •

The point of Bayes' Theorem is this:

Without any other information, we know something about the distribution of televisions in the community: the (*prior probability*), $P(T)$.

Additional information is given to us that associates rates of television ownership based on whether families have a double or single income.

We can use the information on family incomes to update our estimates of television ownership (*posterior probability*), $P(T/D)$.

Example

Expressing the original question in Bayesian terms:

$$P(D | T) = 9/17 \text{ and } P(D | T') = 16/65 \text{ and } P(T) = 17/82$$

First find $P(D)$

Then find $(T | D)$

	T	T'	Σ
D	9	16	25
D'	8	49	57
Σ	17	65	82

Evaluating

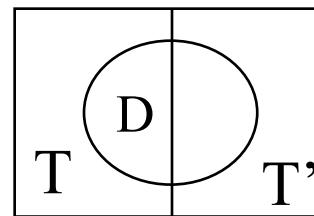
$$P(D | T) = 9/17 \text{ and } P(D | T') = 16/65 \text{ and } P(T) = 17/82$$

$$P(D \cap T) = P(D | T) * P(T) = 9/17 * 17/82 = 9/82 \text{ and}$$

$$P(D \cap T') = P(D | T') * P(T') = 16/65 * 65/82 = 16/82$$

By the law of total probability,

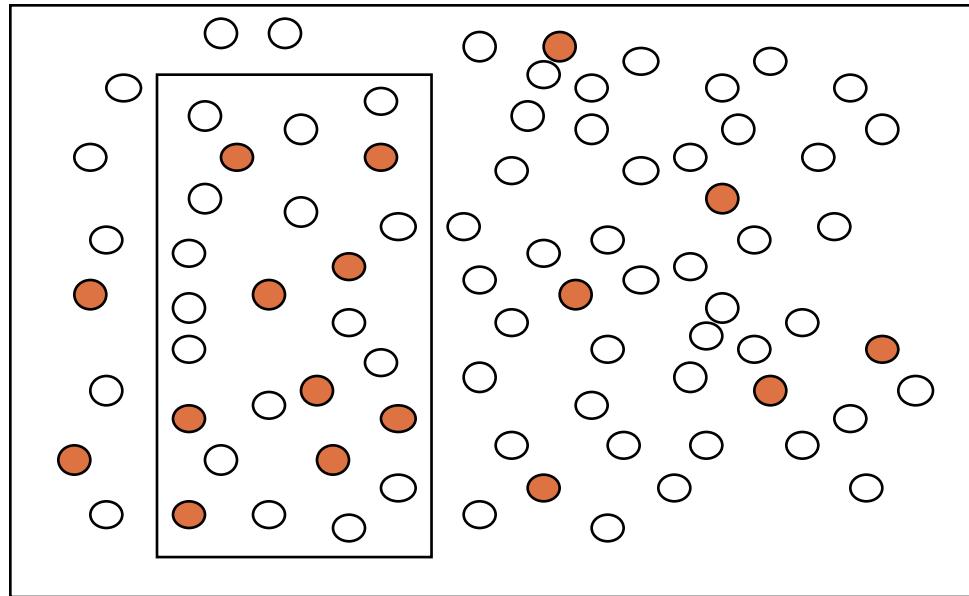
$$P(D) = 9/82 + 16/82 = 25/82$$



$$\text{Consequently } P(T | D) = (9/82) / (25/82) = 9/25 = 0.36$$

Which we can verify from slide 6.

The Initial Problem



- One television
- Two or more televisions
- Double Income

$$P(T | D) = (9/82) / (25/82) = 0.36$$

	T	T'	Σ
D	9	16	25
D'	8	49	57
Σ	17	65	82

Example: as a tree diagram

Expressing the original question in Bayesian terms:

$$P(D | T) = 9/17 \text{ and } P(D | T') = 16/65 \text{ and } P(T) = 17/82$$

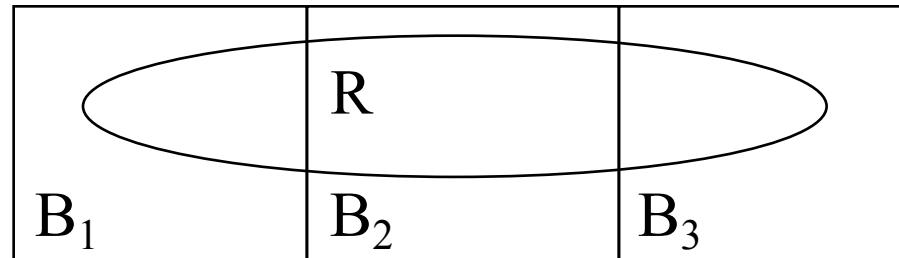
First find $P(D)$, Then find $(T | D)$.

Bayes' Theorem

Formal statement:

For an event with outcomes $B_1, B_2 \dots B_n$, and event R . what is the probability that outcome B_x occurred given that event R has occurred?

$$P(B_x | R) = \frac{P(B_x)P(R|B_x)}{\sum_{j=1}^n P(B_j)P(R|B_j)} = \frac{P(B_x \cap R)}{P(R)}$$



Example 2

A plant has two machines. Machine A produces 60% of the output with the fraction defective being 0.02. Machine B produces 40% of the output with the fraction defective being 0.04. The quality control inspector finds a defective part awaiting pack and ship. What is the probability that it was produced by Machine A?

Solution A - Notation

Let E_i be the event of the part being produced by machines A and B. Let D be the event that the part is defective. The probability that any given part is produced by machine A is therefore $P(E_1) = 0.6$ and by machine B is $P(E_2) = 0.4$. The conditional probability that a part is defective given that it was produced by machine A is $P(D|E_1) = 0.02$. The joint probability that a part is produced by machine A and is defective is (by the multiplication law for statistically dependent events) $P(E_1)P(D|E_1)$.

Solution B – table formulation

By Bayes' rule, the probability that a defective part comes was produced by machine A is

$$P(E_1|D) = \frac{P(E_1)P(D|E_1)}{P(E_1)P(D|E_1) + P(E_2)P(D|E_2)}$$

Probability of
manufacturing a defective

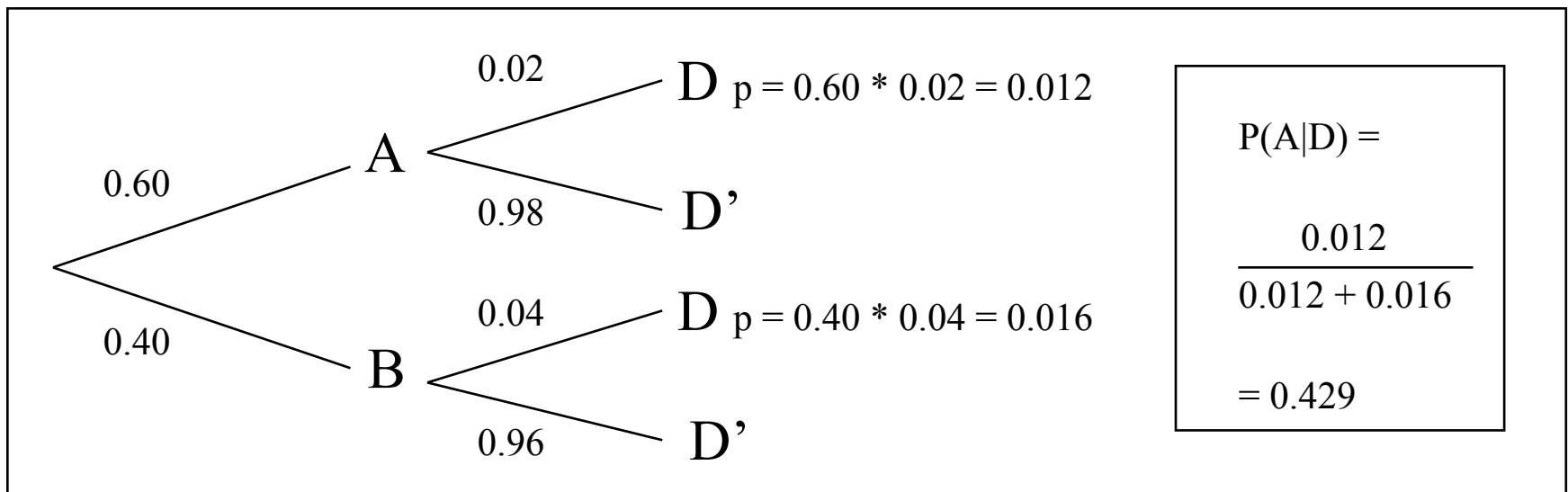
The computation is made easier by means of the following worksheet:

Event E_i	Prior $P(E_i)$	Conditional $P(D E_i)$	Joint $P(E_i)P(D E_i)$	Posterior $P(E_i D)$
Machine A	0.60	0.02	0.012	0.012/0.028 = 0.429
Machine B	0.40	0.04	0.016	0.016/0.028 = 0.571
			0.028	

This shows that even though Machine A produces the greater output, the revised or posterior probability indicates that the quality control effort should first be directed towards Machine B.

Solution C – alternative notation

We can visualise the problem using a tree diagram to determine the probabilities of various outcomes. We can then calculate the posterior probability of the component being manufactured by machine A given that it is defective (D).



Example 2

A kitchenware manufacturer decides that she will not undertake a production plan for a new style of copper-based saucepan unless the probability of the venture being successful is at least 0.8. Her prior estimate of this probability is 0.6. However to ‘test the water’ she produces a small number of saucepan sets and places them in a leading kitchenware store. From past experience she has found that of 20 products which were successful, the customer reaction to a sample was favourable in 15 cases. Further, of 13 products which failed, a favourable customer reaction was recorded in only two cases. If the reaction to the saucepans is favourable, should she go ahead with manufacturing plans?

Solution

Let S be the event that the *Venture* is a success

Let T be the event that the *Sample Trial* is a success

Prior $P(S) = 0.6$

Know that $P(T | S) = 15/20$ and $P(T | S') = 2/13$

Want to find $P(S | T)$

$$P(S | T) = P(S \cap T) / P(T)$$

$$T = (T \cap S) \cup (T \cap S')$$

$$\begin{aligned} P(T) &= P(T \cap S) + P(T \cap S') \\ &= P(T | S) * P(S) + P(T | S') * P(S') \\ &= 15/20 * 0.6 + 2/13 * 0.4 = 0.450 + 0.062 \end{aligned}$$

$$P(S | T) = P(S \cap T) / P(T) = 0.450 / 0.512 \approx 0.88$$

Summary

Characteristics of Bayesian type problems:

Prior probability is known. Additional information is available as a conditional probability.

Use the prior and conditional probabilities to determine the joint probabilities.

Law of total probability is required to construct posterior probabilities.

Two approaches for calculation

Tree diagrams (Your lecturer's favourite method)

Tabular form with conditional, joint, posterior probabilities.

Background for next week...

In the next lecture we will study several probability distributions. You may need to be able to make sense of the following notations and symbols if you are going to do hand calculations:

Exponential

Factorial

Combinatorial

Probability Distributions

For many problems, we can define a probability distribution for the set of outcomes that a variable can assume.

There are two main classes of probability distributions:

Discrete - for random variables which can only take on prescribed values *and*

Continuous - for random variables that can take on any value over a certain range.

Exponents

We describe the notation a^b as a raised to the power of b .

This is defined formally as $a^b = \underbrace{(a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdots a)}_{b \text{ times}}$

a and b can take on non - integer values and we often use the number ' e ' as a base. $e \approx 2.7182\dots$

Using a calculator you should be able to calculate expressions such as: 6^2 , e^2 , $e^{-0.5}$, 0.3^{10} , 0.994^{10} , $e^{-0.44}$, 8^8 , -0.001^2 , e^{-20} .

You should be able to use the $[y^x]$ and $[e^x]$ or $[\exp]$ keys on your calculator.

Factorial

Factorial notation, '!' is easiest to understand with an example.

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, $1! = 1$, and $0! = 1$ by convention.

Formally, $n! = n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3) \cdots 3 \cdot 2 \cdot 1$

Combinations

We use the notation nC_x or $\binom{n}{x}$ to describe the number of different ways we can select x objects at a time from a group of n objects.

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

The number of ways that we can select 4 students from a class of 10

$$\text{students is given by } {}^{10}C_4 = \frac{10!}{4!(10-4)!} = \frac{10!}{4! \cdot 6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

The number of ways that we can select 3 cards from a deck of 52 cards

$$\text{is given by } {}^{52}C_3 = \frac{52!}{3!(52-3)!} = \frac{52!}{3! \cdot 49!} = \frac{52 \cdot 51 \cdot 50}{3 \cdot 2 \cdot 1} = 22100$$

FIT1006 Lecture 11 – Pre-reading

Discrete probability distributions:

The Binomial Distribution

The Poisson Distribution

Textbook:

7th Ed. Sections 7.1, 7.2, 7.6, 7.7.

Revision for hand calculations

Depending on your calculator, you may not need to know how to manually calculate some of the probability distributions we will encounter this week.

If you do need to manually calculate the probability distributions, the following formulas may be useful.

Exponents

We describe the notation a^b as a raised to the power of b .

This is defined formally as $a^b = \underbrace{(a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdots a)}_{b \text{ times}}$

a and b can take on non-integer values and we often use the number ' e ' as a base. $e \approx 2.7182\ldots$

Using a calculator you should be able to calculate expressions such as: 6^2 , e^2 , $e^{-0.5}$, 0.3^{10} , 0.994^{10} , $e^{-0.44}$, 8^8 , -0.001^2 , e^{-20} .

You should be able to use the $[y^x]$ and $[e^x]$ or $[\exp]$ keys on your calculator.

Factorial

Factorial notation, '!' is easiest to understand with an example.

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, $1! = 1$, and $0! = 1$ by convention.

Formally, $n! = n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3) \cdots 3 \cdot 2 \cdot 1$

Combinations

We use the notation nC_x or $\binom{n}{x}$ to describe the number of different ways we can select x objects at a time from a group of n objects.

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

The number of ways that we can select 4 students from a class of 10

$$\text{students is given by } {}^{10}C_4 = \frac{10!}{4!(10-4)!} = \frac{10!}{4! 6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

The number of ways that we can select 3 cards from a deck of 52 cards

$$\text{is given by } {}^{52}C_3 = \frac{52!}{3!(52-3)!} = \frac{52!}{3! 49!} = \frac{52 \cdot 51 \cdot 50}{3 \cdot 2 \cdot 1} = 22100$$

The Binomial Distribution

We use the Binomial Distribution to determine the ‘number of successes’, each with a probability p of occurring in n independent trials.

Typical questions:

A coin is tossed 50 times, what is the probability that 22 heads will be tossed? What is the probability of more than 40 heads being tossed?

A machine produces bolts and from past experience it is known that there is a 0.01 probability that a bolt produced will be defective. If a box contains 100 bolts, what is the probability that there are less than 5 defectives in the box.

Formal Statement

The use of the Binomial Distribution is valid under the following conditions:

- 1 Trials are independent
- 2 There are only two outcomes for each trial
- 3 The probability of success in each trial is constant

For n independent trials, each with a probability p , of success, we define the the number of successes X , as a Binomial Distribution with the following formula :

$$P(X = x) = {}^nC_x p^x (1 - p)^{(n-x)} \text{ for } x = 0,1,2,3\dots n$$

Example (single value)

A fair coin is tossed 10 times and the number of heads appearing uppermost is counted. What is the probability that 8 heads are thrown?

We have $n = 10$, $p = 0.5$ and $x = 8$

thus $P(x) = {}^nC_x p^x (1-p)^{(n-x)}$

$$P(8) = {}^{10}C_8 0.5^8 (1-0.5)^{(10-8)}$$

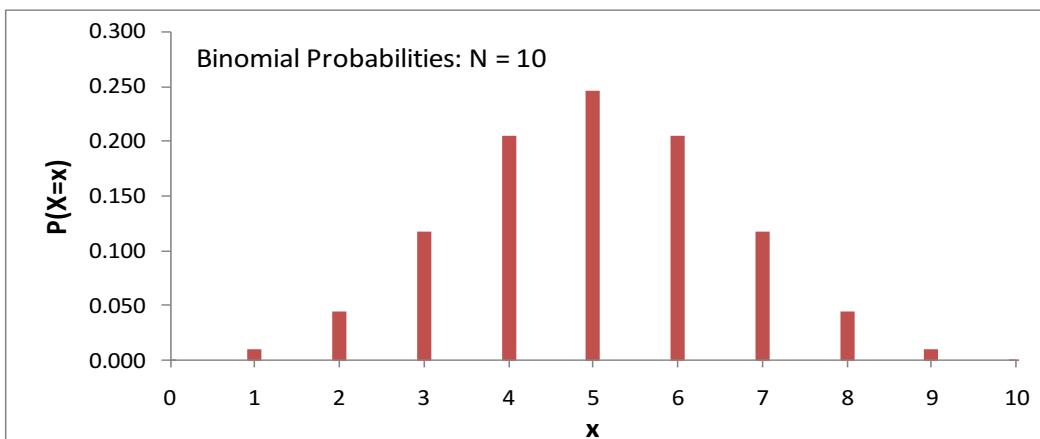
$$= \frac{10 \times 9}{2 \times 1} \times 0.5^8 \times 0.5^2$$

$$= 0.0439$$

Example (distribution)

A fair coin is tossed 10 times and the number of heads appearing uppermost is counted. What is the probability distribution for the number of heads thrown?

Using tables constructed in Excel:
Probabilities are points.



n =	10
p =	0.50
0	0.0010
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.0010
sum =	1.0000

Tables

It is usual to use tables of Binomial Probabilities:

Table gives $P(X=x)$ for $X = Bi(n,p)$								
n =	10							
p =	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70
0	0.5987	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000
1	0.3151	0.3874	0.2684	0.1211	0.0403	0.0098	0.0016	0.0001
2	0.0746	0.1937	0.3020	0.2335	0.1209	0.0439	0.0106	0.0014
3	0.0105	0.0574	0.2013	0.2668	0.2150	0.1172	0.0425	0.0090
4	0.0010	0.0112	0.0881	0.2001	0.2508	0.2051	0.1115	0.0368
5	0.0001	0.0015	0.0264	0.1029	0.2007	0.2461	0.2007	0.1029
6	0.0000	0.0001	0.0055	0.0368	0.1115	0.2051	0.2508	0.2001
7	0.0000	0.0000	0.0008	0.0090	0.0425	0.1172	0.2150	0.2668
8	0.0000	0.0000	0.0001	0.0014	0.0106	0.0439	0.1209	0.2335
9	0.0000	0.0000	0.0000	0.0001	0.0016	0.0098	0.0403	0.1211
10	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0060	0.0282

Cumulative Probabilities

Frequently we want to calculate the probability that the number of successes lies within a certain range. For example the probability that the number of heads thrown in the previous trial is less than 4.

In this case we would calculate the probability that 0, 1, 2 and 3 heads are thrown.

$$P(x = 0) = 0.0010$$

$$P(x = 1) = 0.0098$$

$$P(x = 2) = 0.0439$$

$P(x = 3) = 0.1172$ thus, summing these gives

$$\mathbf{P(x < 4) = 0.1719}$$

Binomial Probabilities in EXCEL

We can use the built-in functions in EXCEL to calculate binomial probabilities.

For $X = Bi(n,p)$

$$P(X = x) = BINOM.DIST(x, n, p, \text{false})$$

The cumulative probability is given by

$$P(X \leq x) = BINOM.DIST(x, n, p, \text{true})$$

... or learn to use your calculator for these.

Problem

A farmer produces free range eggs which are collected, graded and packaged into cartons of 12. The probability that a defective egg gets through the packaging process is 0.05. What is the probability that a carton will contain more than one defective egg.

Now, $n = 12$, $p = 0.05$,

We want $P(x > 1)$ ie $P(x = 2) + P(x = 3) + \dots$

A quicker way to use our knowledge of complementary events.

$$\begin{aligned}P(x > 1) &= 1 - P(x = 0) - P(x = 1) && * \\&= 1 - 0.5404 - 0.3413 = 0.1183\end{aligned}$$

Sample Calculations

For the previous example, n=12, p=0.05.

$$P(x) = {}^nC_x p^x (1-p)^{(n-x)}$$

$$P(0) = {}^{12}C_0 0.05^0 (1-0.05)^{(12-0)}$$

$$= 1 \times 1 \times 0.95^{12}$$

$$= 0.5404$$

$$P(1) = {}^{12}C_1 0.05^1 (1-0.05)^{(12-1)}$$

$$= 12 \times 0.05^1 \times 0.95^{11}$$

$$= 0.3413$$

$$P(2) = {}^{12}C_2 0.05^2 (1-0.05)^{(12-2)}$$

$$= \frac{12 \times 11}{2 \times 1} \times 0.05^2 \times 0.95^{10}$$

$$= 0.0987$$

The Poisson Distribution

We use the Poisson distribution to determine the number of occurrences of a random event, distributed over time or space.

Typical Poisson Distribution questions:

On average 100 customers per day visit a particular shop. What is the probability that 10 customers enter the shop over one hour?

A fabric is known to have, on average, one defect per 10 meters. What is the probability that 5 meters of fabric will have two defects?

We may also use the Poisson distribution as an approximation to the Binomial distribution.

Formal Statement

The use of the Poisson Distribution is valid under the following conditions:

- 1 Trials record the number of occurrences of a random event distributed over time or space.
- 2 The number of occurrences is theoretically unlimited

For a random event occurring with an average rate λ , the mean number of occurrences over a period or area t is given by $\mu = \lambda t$. The number of occurrences over the time or period x , has a Poisson distribution with the formula :

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!} \text{ for } x = 0, 1, 2, 3, \dots, \infty$$

Example

A certain fabric has a probability of 0.05 that a given metre will have a defect. What is the probability that a 10m length will have 3 defects?

We have $t = 10$, $\lambda = 0.05$ thus $\mu = 0.5$

$$\begin{aligned}P(x) &= \frac{e^{-\mu} \mu^x}{x!} \\P(3) &= \frac{e^{-0.5} \times 0.5^3}{3!} \\&= \frac{0.6065 \times 0.125}{6} \\&= 0.0126\end{aligned}$$

Tables

We see that the same result is obtained from tables:

Table gives $P(X=x)$ for $X = \text{Poi}(\mu)$								
μ	0.5	1.0	2.0	3.0	4.0	5.0	6.0	7.0
0	0.6065	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009
1	0.3033	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149	0.0064
2	0.0758	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446	0.0223
3	0.0126	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892	0.0521
4	0.0016	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339	0.0912
5	0.0002	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606	0.1277
6	0.0000	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606	0.1490
7	0.0000	0.0001	0.0034	0.0216	0.0595	0.1044	0.1377	0.1490
8	0.0000	0.0000	0.0009	0.0081	0.0298	0.0653	0.1033	0.1304
9	0.0000	0.0000	0.0002	0.0027	0.0132	0.0363	0.0688	0.1014
10	0.0000	0.0000	0.0000	0.0008	0.0053	0.0181	0.0413	0.0710

Poisson Probabilities in EXCEL

You can use the built-in functions in EXCEL to calculate Poisson probabilities.

For $X = Poi(\mu)$

$$P(X = x) = POISSON.DIST(\mu, x, \text{false})$$

The cumulative probability is given by

$$P(X \leq x) = POISSON.DIST(\mu, x, \text{true})$$

... or learn to use your calculator for these.

Table of Cumulative Probabilities

Table gives $P(X \leq x)$ for $X = \text{Poi}(\mu)$											
mu	0.5	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
0	0.6065	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001	0.0000
1	0.9098	0.7358	0.4060	0.1991	0.0916	0.0404	0.0174	0.0073	0.0030	0.0012	0.0005
2	0.9856	0.9197	0.6767	0.4232	0.2381	0.1247	0.0620	0.0296	0.0138	0.0062	0.0028
3	0.9982	0.9810	0.8571	0.6472	0.4335	0.2650	0.1512	0.0818	0.0424	0.0212	0.0103
4	0.9998	0.9963	0.9473	0.8153	0.6288	0.4405	0.2851	0.1730	0.0996	0.0550	0.0293
5	1.0000	0.9994	0.9834	0.9161	0.7851	0.6160	0.4457	0.3007	0.1912	0.1157	0.0671
6	1.0000	0.9999	0.9955	0.9665	0.8893	0.7622	0.6063	0.4497	0.3134	0.2068	0.1301
7	1.0000	1.0000	0.9989	0.9881	0.9489	0.8666	0.7440	0.5987	0.4530	0.3239	0.2202
8	1.0000	1.0000	0.9998	0.9962	0.9786	0.9319	0.8472	0.7291	0.5925	0.4557	0.3328
9	1.0000	1.0000	1.0000	0.9989	0.9919	0.9682	0.9161	0.8305	0.7166	0.5874	0.4579
10	1.0000	1.0000	1.0000	0.9997	0.9972	0.9863	0.9574	0.9015	0.8159	0.7060	0.5830

Problem

Customers enter a shop at an average rate of one per minute. What is the probability that in a 10 minute period more than 7 customers will enter?

Now, $\lambda = 1$, $t = 10$, and $mu = 10$

We want $P(x > 7)$ ie $P(x = 8) + P(x = 9) + \dots$

$$P(x > 7) = 1 - P(x = 7) - P(x = 6) - \dots - P(x = 0)$$

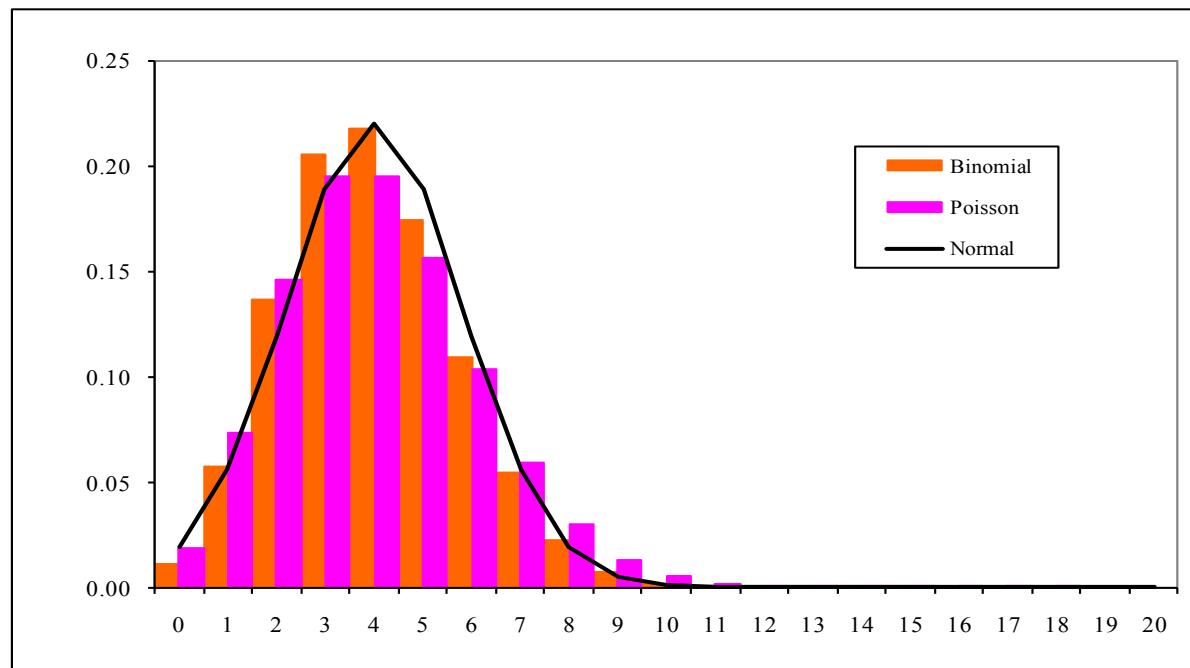
(we can use the table of cumulative probabilities here)

$$P(x > 7) = 1 - 0.2202 = 0.7798$$

Poisson Approximation to the Binomial

When n is large and p is small we let $mu = np$. The graph shows visually why this works:

$n = 20, p = 0.2$. Try this for other values of p .



Poisson Approximation to the Binomial

A farmer produces free range eggs which are collected, graded and packaged into cartons of 12. The probability that a defective egg gets through the packaging process is 0.05. What is the probability that a carton will contain more than one defective egg.

Now, $n = 12$, $p = 0.05$,

When n is large and p is small, we let $\mu = np$

$$P(X = x) = Poi(0.6)$$

$$\begin{aligned}P(x > 1) &= 1 - P(x = 0) - P(x = 1) \\&= 1 - 0.5488 - 0.3293 = 0.1219\end{aligned}\quad *$$

Using a Poisson approximation to the Binomial Distribution

Sample Calculations

For the previous question, $\mu = 0.05 * 12 = 0.6$

$$P(x) = \frac{e^{-\mu} \mu^x}{x!}$$

$$\begin{aligned} P(0) &= \frac{e^{-0.6} \times 0.6^0}{0!} \\ &= \frac{0.5488 \times 1}{1} \\ &= 0.5488 \end{aligned}$$

$$\begin{aligned} P(1) &= \frac{e^{-0.6} \times 0.6^1}{1!} \\ &= \frac{0.5488 \times 0.6}{1} \\ &= 0.3293 \end{aligned}$$

$$\begin{aligned} P(2) &= \frac{e^{-0.6} \times 0.6^2}{2!} \\ &= \frac{0.5488 \times 0.36}{2} \\ &= 0.0988 \end{aligned}$$

Mean and Variance

The table below contains a summary of the formula for each distribution, as well as the mean and variance of each.

Distribution	Formula	Mean	Variance
Binomial	$P(X = x) = {}^nC_x p^x (1 - p)^{(n-x)}$	np	$np(1 - p)$
Poisson	$P(X = x) = \frac{e^{-\mu} \mu^x}{x!}$	μ	μ

Binomial vs Poisson

Binomial: 2 parameters, n & p

Series of n trials

Maximum known

Poisson: 1 parameter, μ

Maximum theoretically infinite

Know how to calculate Binomial or Poisson probabilities by hand or with your calculator.

FIT1006 Lecture 12 Pre-reading

The Normal Distribution

Normal approximation to the Binomial & Poisson Distributions

Textbook:

7th Ed. Section 8.3.

The Normal Distribution

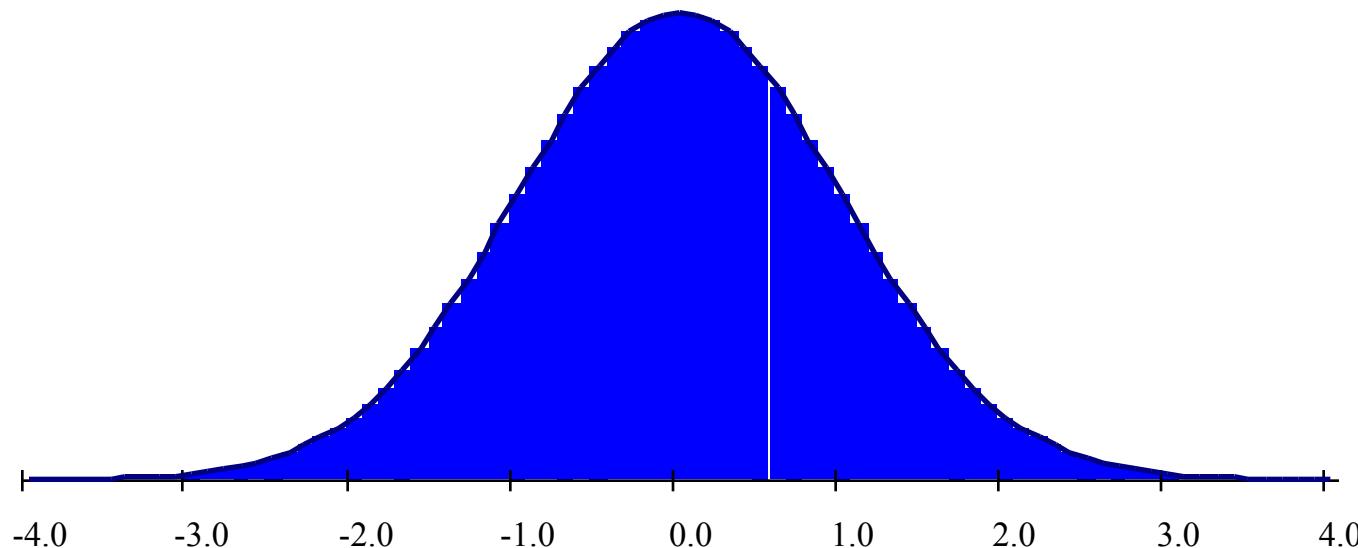
The most important distribution in statistics

Arises when we measure a large number of nearly identical objects subject to random fluctuations - height, weight, response time. (Used a lot in biometrics).

The Normal Distribution arises when we take the sum or means of a large number of observations from any distribution and thus provides the basis for sampling theory.

The Bell Curve

The shape of the Normal distribution can be seen below. The shape is often referred to as the bell curve. The distribution below has a mean of 0 and a standard deviation of 1 and is called the *Standard Normal*.



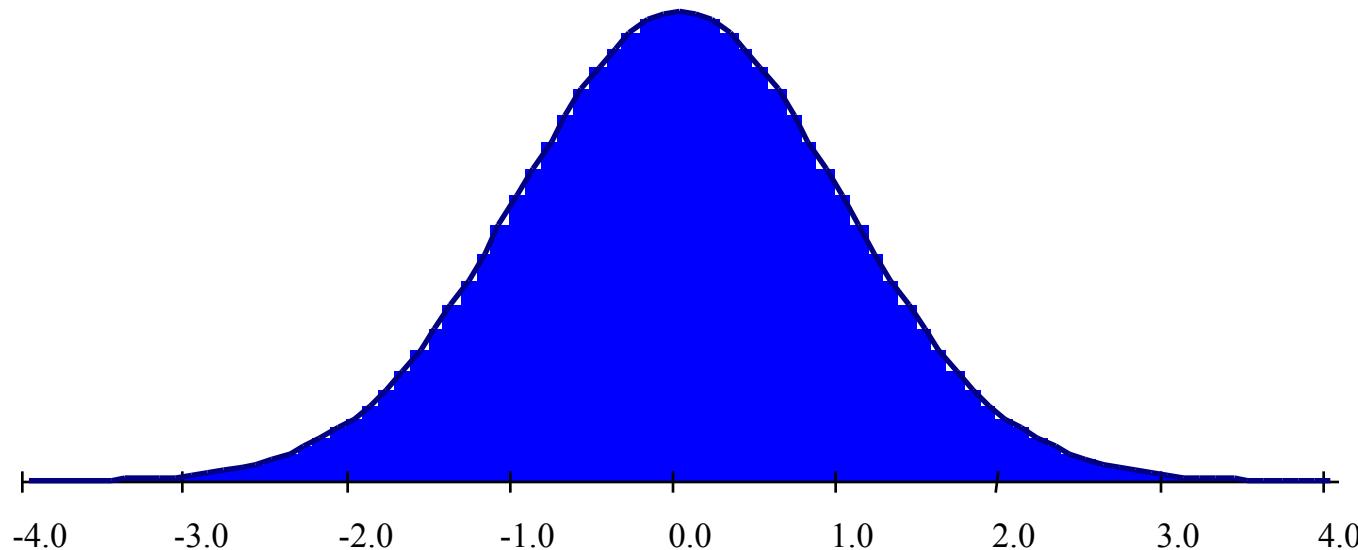
General Properties 1

The total area under the curve = 1. Approximately:

68% of the area is within 1 standard deviation of the mean.

95% of the area is within 2 standard deviations of the mean.

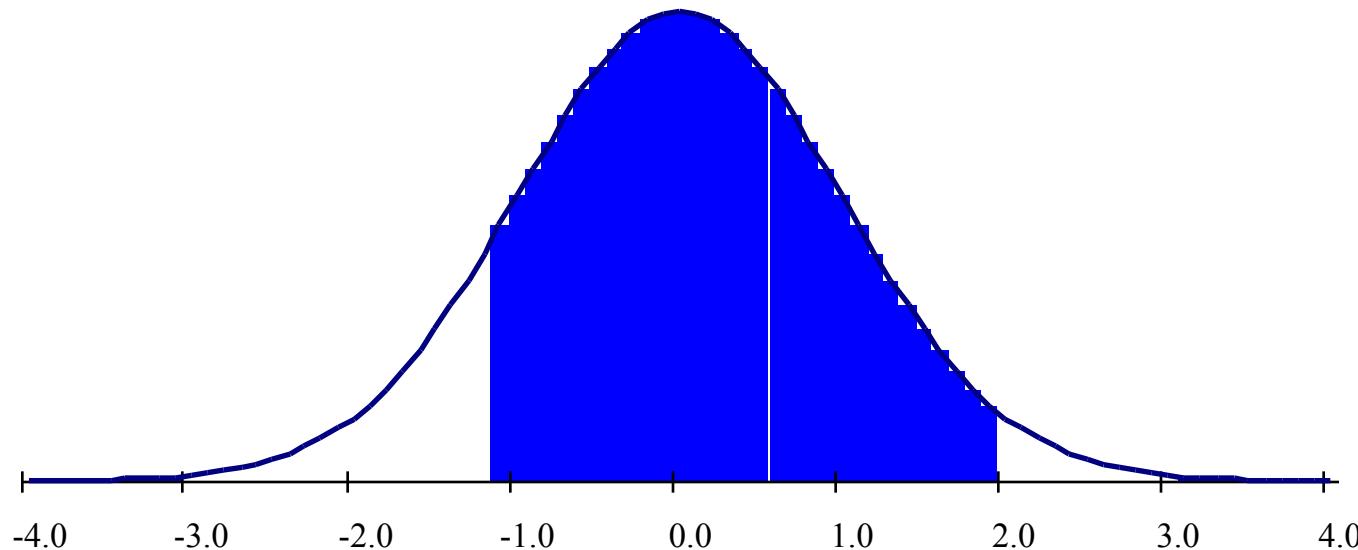
99.7% of the area is within 3 standard deviations of the mean.



General Properties 2

As the Normal Distribution is a continuous distribution, the probability that Z equals a particular value is zero. Instead we find the probability that Z lies within a particular range by calculating the area under the curve.

Eg. $P(Z < 2)$, $P(Z > 1.2)$, $P(-1.1 < Z < 2.0)$ etc.



Calculating Normal Probabilities

Because the integral for the Normal function doesn't have a closed form expression, the usual approach to calculating probabilities is by using a table of Cumulative probabilities (the CDF), or through the built in Excel functions (which are good approximations of the exact values), *or your calculator!*

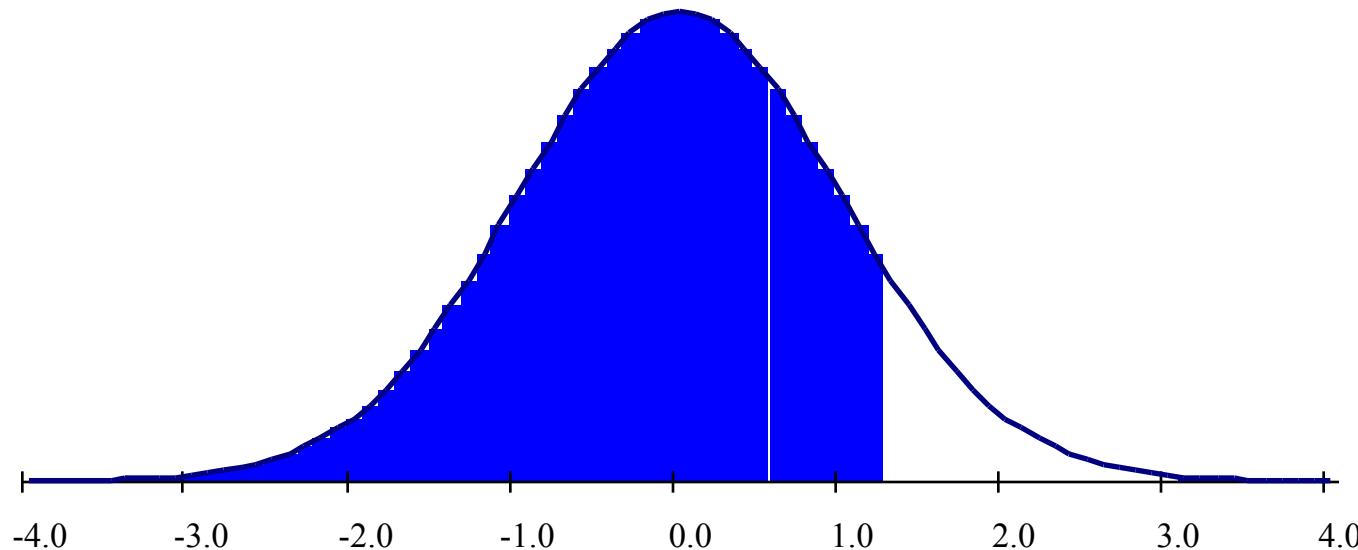
The Excel Function: =NORMDIST(Z,Mean,Stdev,TRUE).

Standard Normal CDF Table

Cumulative Probabilities for the Standard Normal Distribution										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

$P(Z < a)$ when a is positive

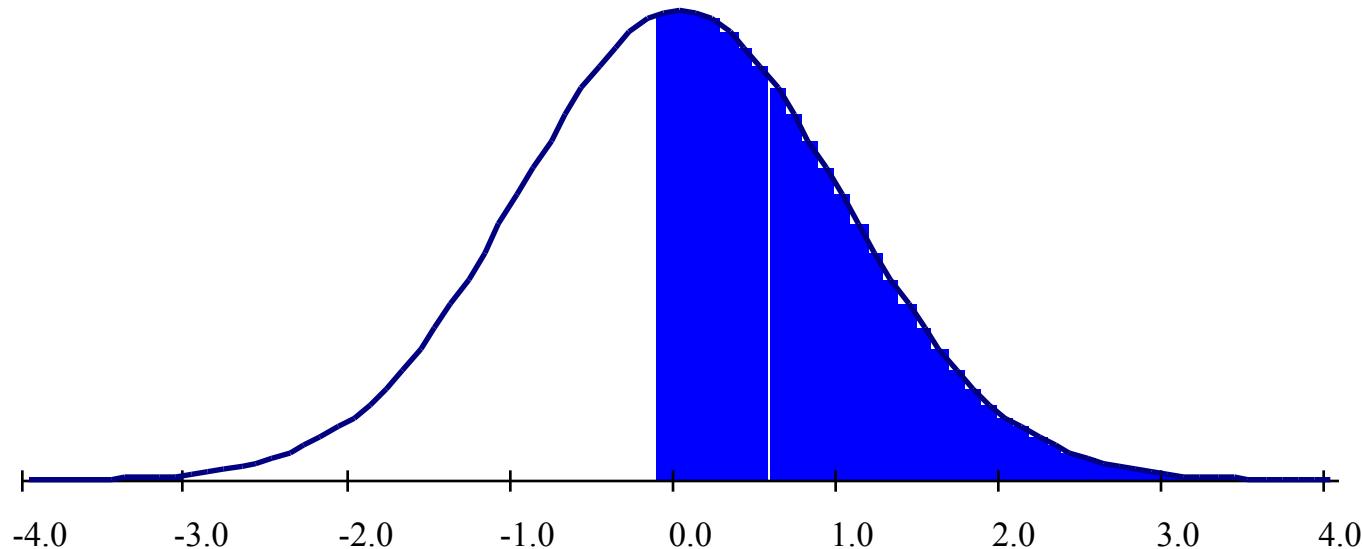
$P(Z < 1.33) = 0.9082$ from tables or use
 $NORMDIST(1.33, 0, 1, true)$ in Excel.



$P(Z > a)$ when a is negative

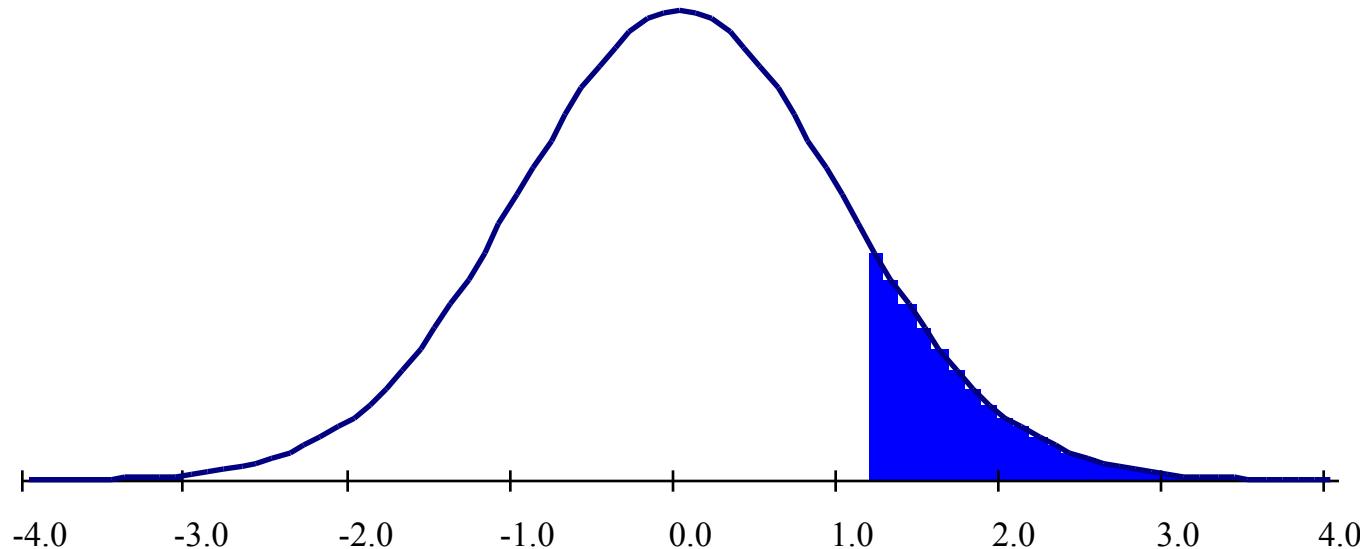
$$P(Z > -0.07) = P(Z < 0.07) = 0.5279$$

when a is negative.



$P(Z > a)$ when a is positive

$$P(Z > 1.21) = 1 - P(Z < 1.21) = 1 - 0.8869 = 0.1131$$

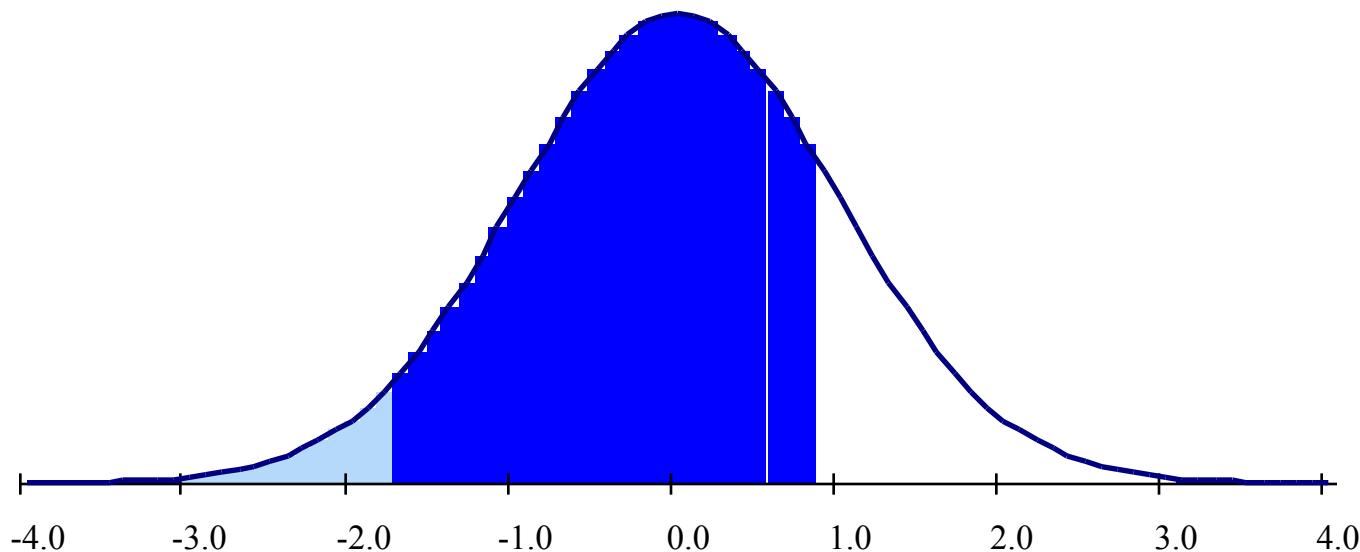


$P(Z < a)$ when a is negative

Try yourself

$$P(a < Z < b)$$

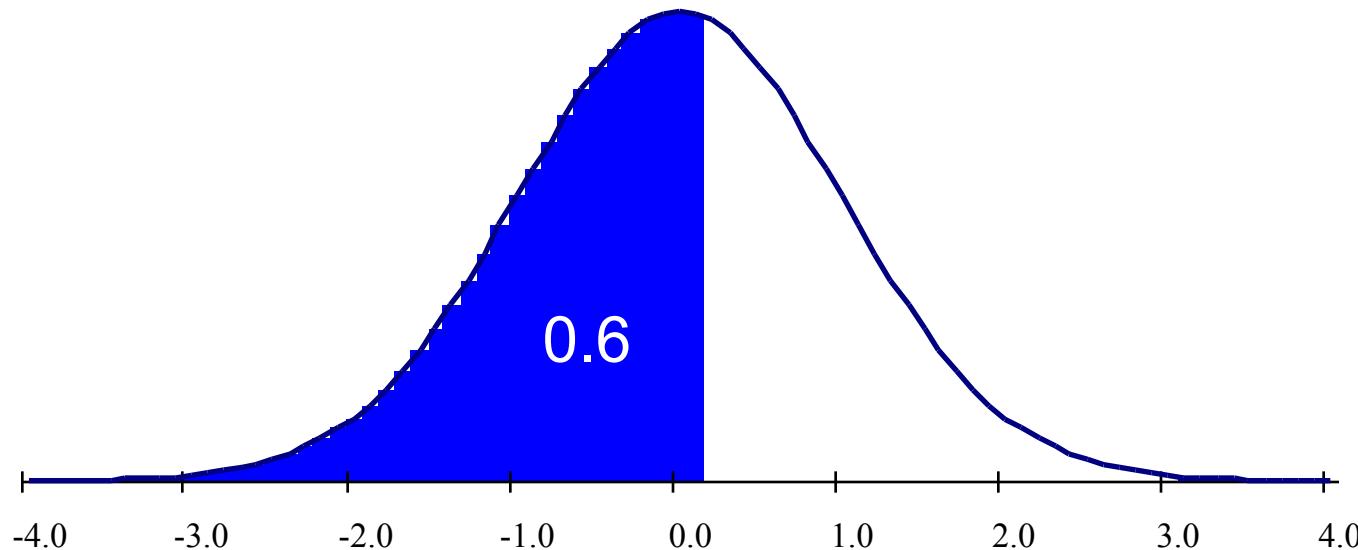
$$\begin{aligned} P(-1.77 < Z < 0.95) &= P(Z < 0.95) - P(Z < -1.77) \\ &= P(Z < 0.95) - P(Z > 1.77) \\ &= 0.8289 - (1 - 0.9616) \\ &= 0.7905 \end{aligned}$$



Inverse CDF

We frequently want to find a value of Z for a given probability.

For example find a such that $P(Z < a) = 0.6$. We can read the tables in reverse and find that $a = 0.255$. We can use NORMINV(0.6,0,1) in Excel.



The Normal Distribution in Practice

The Standard Normal distribution has mean = 0 and standard deviation = 1. This tends to limit the applications of Standard Normal to everyday problems.

However, as the Normal distribution is completely defined by the mean and variance (or standard deviation) we can apply the Standard normal to any problem by standardising the variable of interest.

Standardising Variables

Let X be a normal variate with mean μ and variance σ^2 .

We can write $X \approx N(\mu, \sigma^2)$. We can standardise X by

use of the formula: $z = \frac{x - \mu}{\sigma}$ and $Z \approx N(0,1)$. In this

way we can apply the Standard Normal probabilities to any problem.

Example

A machine manufactures bolts which have a length of 40mm with a variance of 4mm² what is the probability that a bolt manufactured by the machine has a length greater than 43 mm?

From the problem, $X \approx N(\mu = 40, \sigma^2 = 4)$. We standardise X

using the formula: $z = \frac{x - \mu}{\sigma}$, thus $P(X > 43)$ becomes

$$P\left(Z > \frac{43 - 40}{2}\right) \text{ and } Z \approx N(0,1).$$

Thus we calculate $P(Z > 1.5) = 1 - 0.9332 = 0.0668$

Example

The number of customers entering a shop is known from historical information to be normally distributed with a mean of 365 and a standard deviation of 33. What is the probability that on any given day the number of customers will be less than 320?

From the problem, $X \approx N(\mu = 365, \sigma^2 = 33^2)$. thus $P(X < 320)$ becomes $P\left(Z < \frac{320 - 365}{33}\right)$ and $Z \approx N(0,1)$.

Thus we calculate $P(Z < -1.36) = 1 - 0.9131 = 0.0869$

Example

The proprietor of the shop described in the previous question wants to set her staffing levels. She wants to be 80% sure of being able to meet customer demand. What number of customers per day should she plan for?

From the problem, $X \approx N(\mu = 365, \sigma^2 = 33^2)$. We want to find α such that $P\left(Z < \frac{\alpha - 365}{33}\right) = 0.8$. From the tables, $P(Z < 0.845) = 0.8$ and so $\alpha = 0.845 \times 33 + 365 = 392.85$, approximately 393 customers per day.

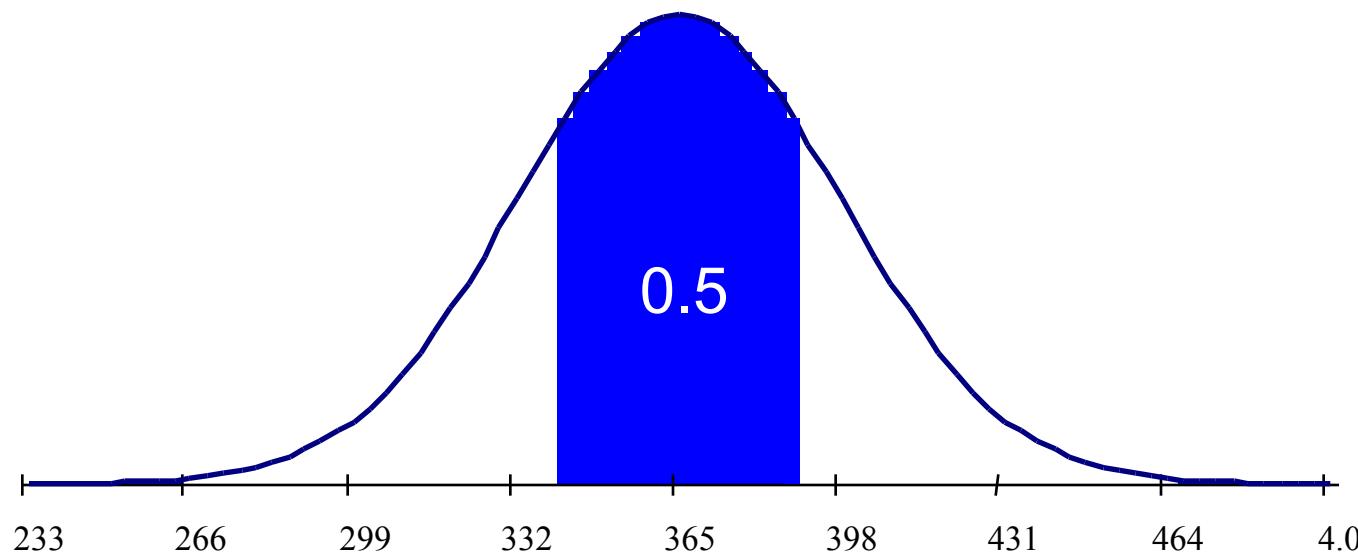
Example

For the preceding problem data, over what range would the proprietor expect the middle 50% of customer demand to range?

From the problem, $X \approx N(\mu = 365, \sigma^2 = 33^2)$. We want to find α and β such that $P\left(\frac{\alpha - 365}{33} < Z < \frac{\beta - 365}{33}\right) = 0.5$. By the symmetry of the distribution and from tables, $P(Z < 0.675) = 0.75$ and so $365 \pm 0.675 \times 33$ gives our upper and lower values as $\alpha = 342.7$ and $\beta = 387.2$

• • •

The 50% confidence interval from the previous question.



Normal Approximation of Binomial

When n is large and p is small, we can use the Normal distribution to approximate the Binomial distribution.

We use the mean and variance of the Binomial Distribution to give us the parameters of the Normal Distribution and proceed as before.

Because a point probability = 0 in a continuous distribution, (e.g. $P(X=7) = 0$) we make a continuity correction that assumes the probability is determined over an interval of 1 unit when we approximate a discrete distribution with a continuous one.

If we wanted to determine $P(X = 25)$ for a binomial problem, we would use $P(24.5 < X < 25.5)$ as the required interval using a Normal approximation.

Example

Students attempt a multiple choice test consisting of 100 questions, each with 5 possible responses. What is the probability that a student will score of 30 or more just by guessing?

Let X be the student's score. $X \sim Bi(100, 0.2)$

We can approximate this with $X \sim N(20, 16)$

$$P(X > 29.5), X \sim N(20, 16) = P(Z > 2.375), Z \sim N(0, 1)$$

$$= 1 - 0.9912 = 0.0088$$

(not a good strategy for success!)

Normal Approximation of Poisson

When $\mu \geq 5$, we can use the Normal distribution to approximate the Poisson distribution.

We use the mean and variance of the Poisson Distribution to give us the parameters of the Normal Distribution and proceed as before, using a continuity correction.

Example

I have a 100 metre roll of carpet, with a 0.05 probability of a defect in any metre. What is the probability that the roll will contain two or fewer defects?

Let X be the number of defects. $X = \text{Poi}(5)$

We can approximate this with $X = N(5, 5)$

$$\begin{aligned} P(X \leq 2), X = \text{Poi}(5) &\approx P(Z < 2.5), X = N(5, 5) \\ &= P(Z < (2.5 - 5)/\sqrt{5}), Z = N(0, 1) \\ &= 0.1318 \text{ (exact is 0.1247)} \end{aligned}$$

FIT1006 Lecture 17

Estimating population parameters using a sample:

Theoretical Sampling Distributions

Introduction to sampling.

The Central Limit Theorem.

The sampling distribution of the mean and proportion.

Estimation

Estimating population parameters using a sample.

Estimating a population parameter

- We now look at making estimates of a population parameter based on samples.
- We are frequently interested in the mean of a population, or the proportion of a population exhibiting a certain characteristic.
- We look at how we determine the accuracy of our estimate of these parameters, based on the value of the parameter in question and the sample size.

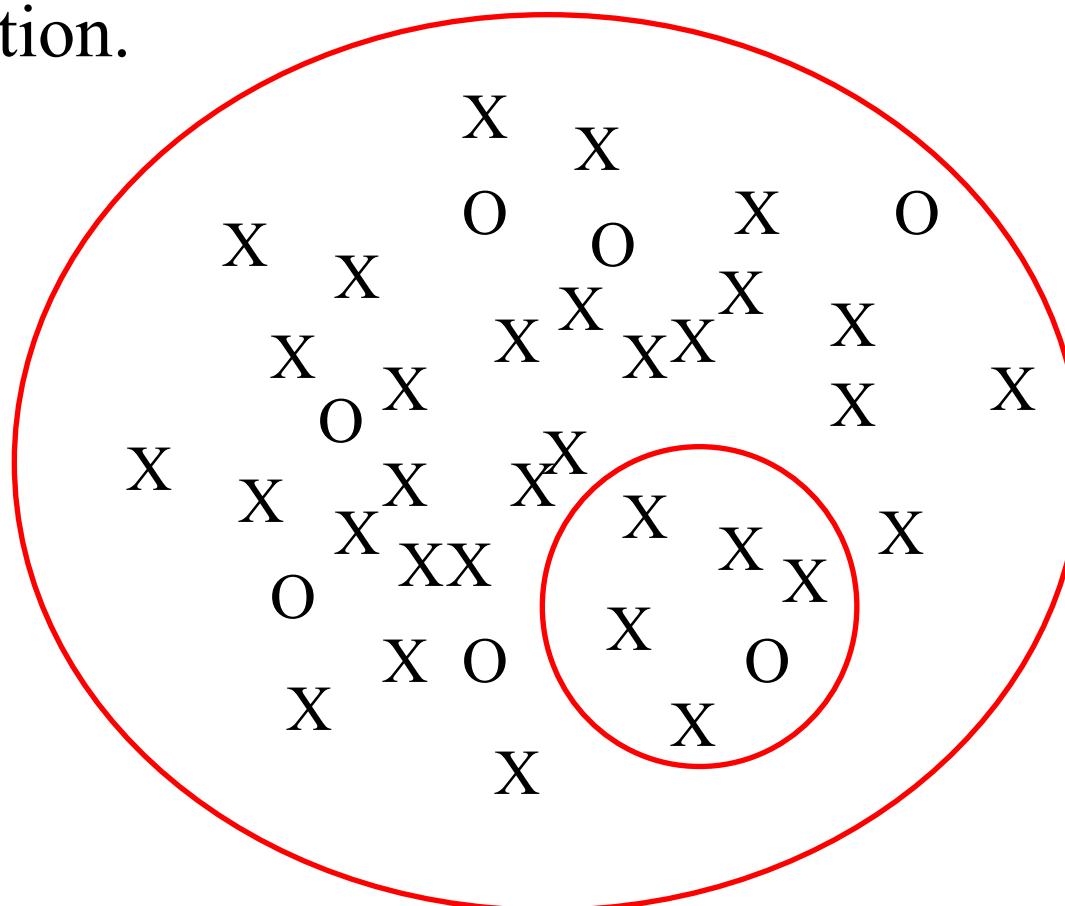
Covered in Lecture 13

Estimation

Part 1. The behaviour of samples

Populations and Samples

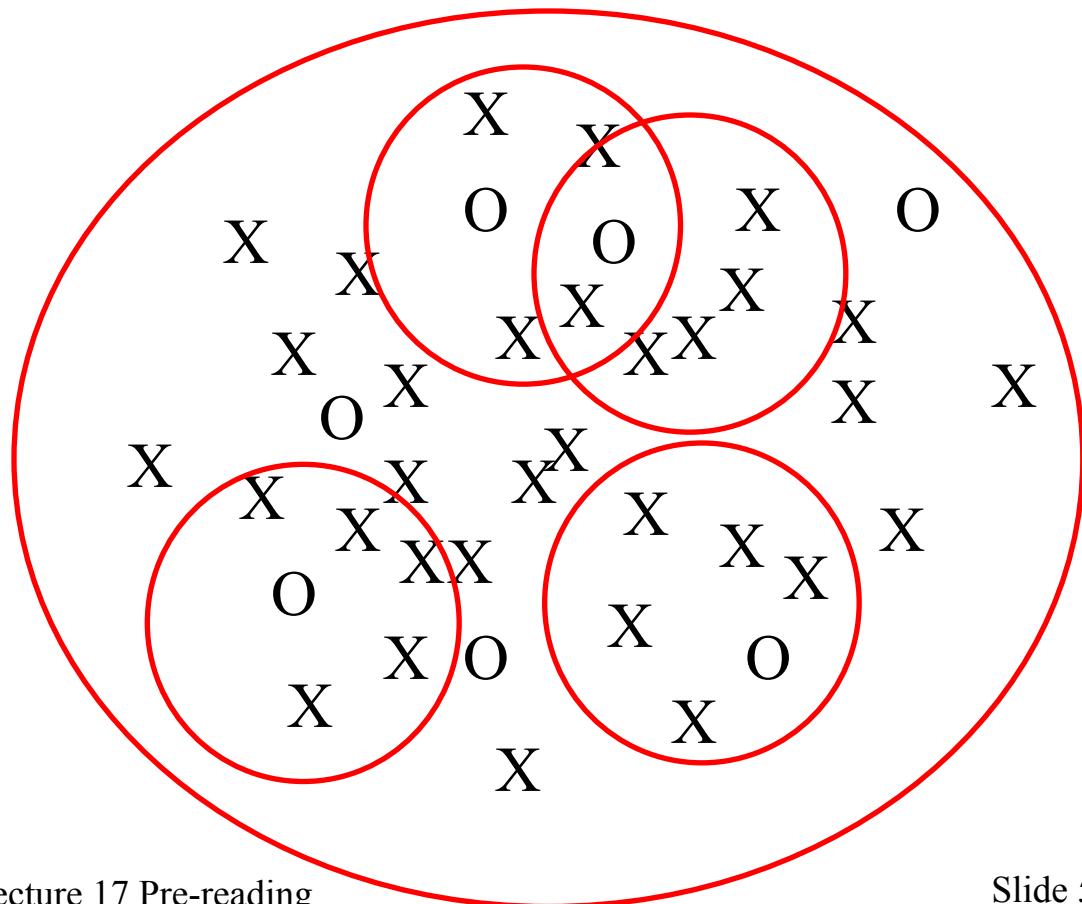
We want to use a sample to make an inference about a population.



Populations and Samples

Taking different random samples of the same size from a population may yield different means.

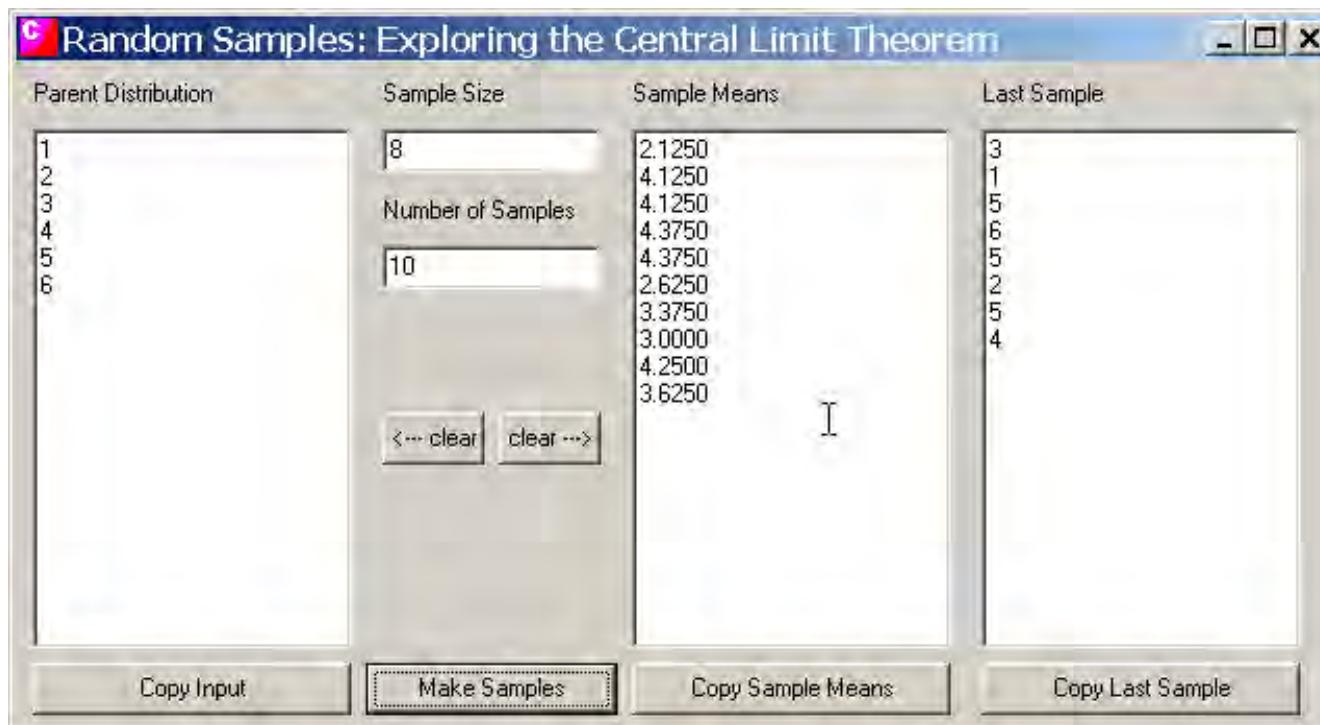
Thus, the sample mean is itself a random variable having its own distribution.



Covered in Lecture 13

CLTProject.exe

This application (See Tutorial 8 for details) lets you take multiple samples from a population and observe the variability in the samples as a function of sample size.



A Binomial distribution problem

The following slide shows samples taken from a population where, for example:

0 = right handed ($p = 0.9$)

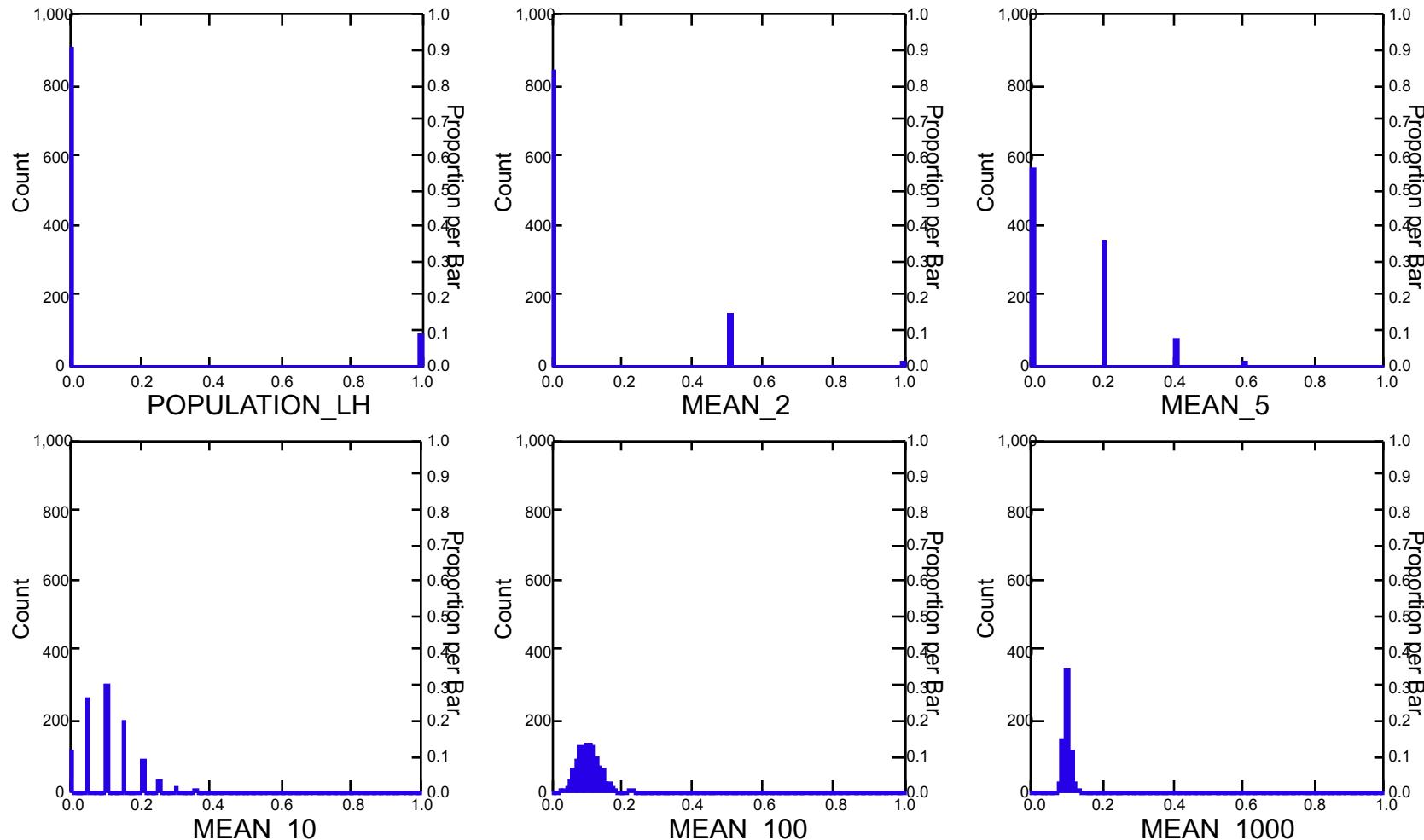
1 = left handed ($p = 0.1$)

Samples of size 1, 2, 5, 10, 100, 1000 are taken and the means calculated.

1000 samples were taken with replacement. (*That means each sample was chosen observed and put back into the population*)

Effect of sample size

Covered in Lecture 13



Observations

As sample size gets larger, 3 things happen:

- 1 Histogram goes from having a Binomial distribution to approaching a Normal distribution.
- 2 Sample mean converges to the population mean.
- 3 Variance of the sample mean decreases – inversely proportional to sample size.

Covered in Lecture 13

Estimation

Part 2. The Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem is fundamental to inferential statistics.

The main idea is that if we take large enough sample from a population, we find that *regardless of the distribution of the parent population*, the sample mean is:

1. Normally distributed around the population mean.
2. The variance of the sample mean is the population variance divided by the size of the sample.

Conditions for the CLT to hold

Covered in Lecture 13

- 1 Samples must be sufficiently large ($n \geq 30$).
- 2 Samples must be of equal size.
- 3 Sampling must be carried out with replacement.

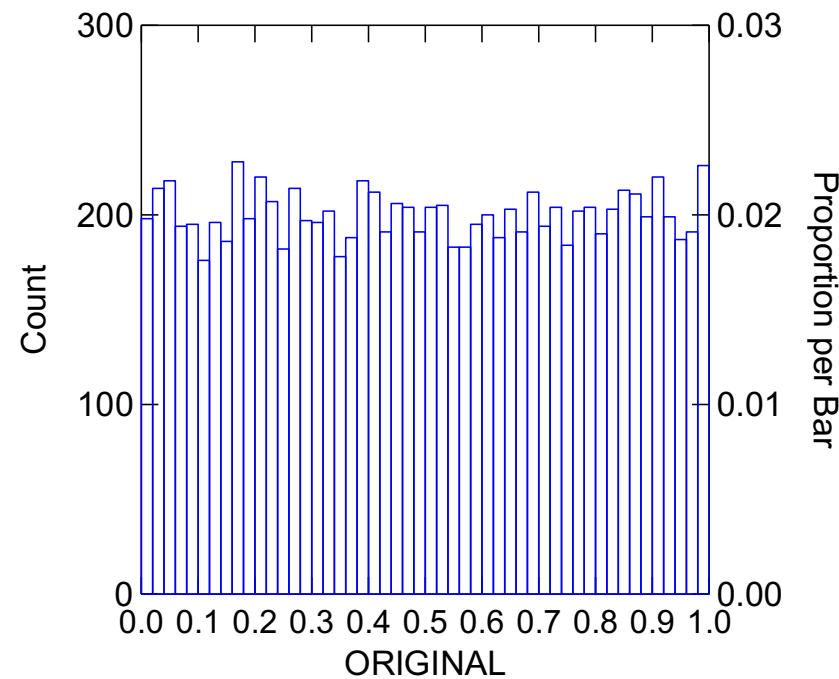
In practice we usually only take and analyse one sample from a population. The conditions above are used to establish the validity of the CLT.

CLT demonstration

10000 uniformly $[0,1]$ distributed random numbers were generated using SYSTAT. A histogram of them appears below.

Data generated using:

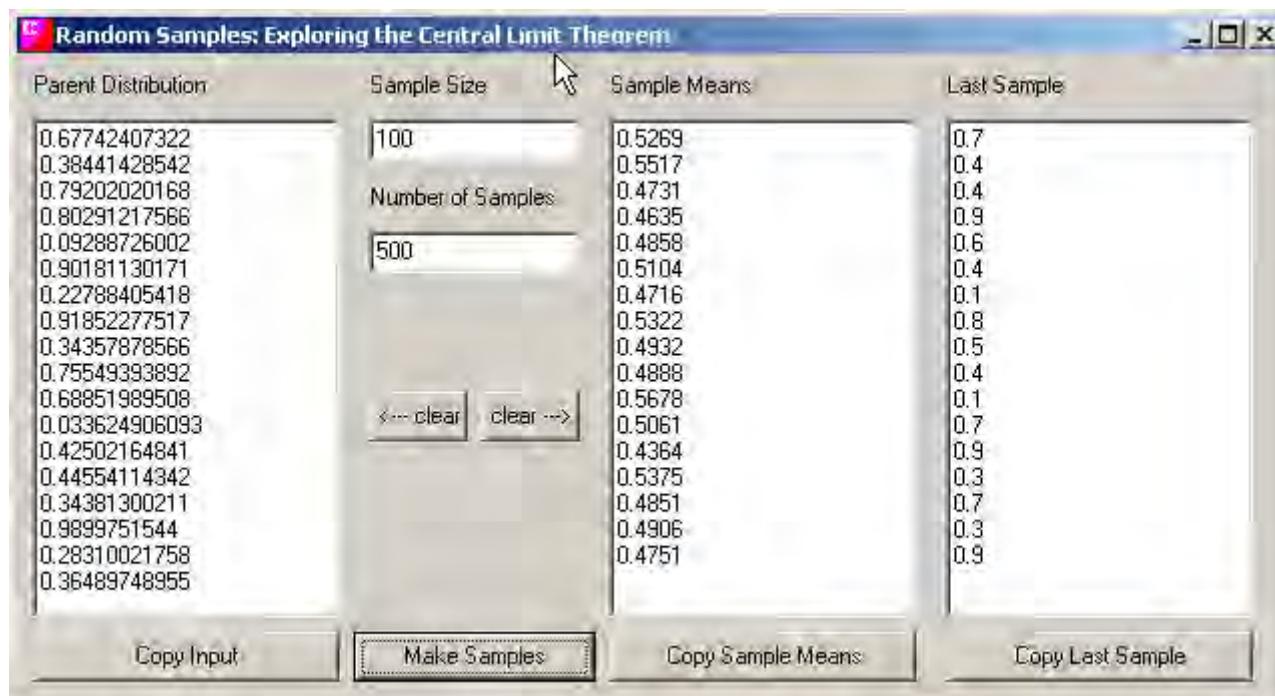
```
Utilities > Basic >  
    BASIC  
    NEW  
    REPEAT=10000  
    LET a=URN  
    SAVE d:\Random_10000_Uniform  
    RUN
```



Covered in Lecture 13

• • •

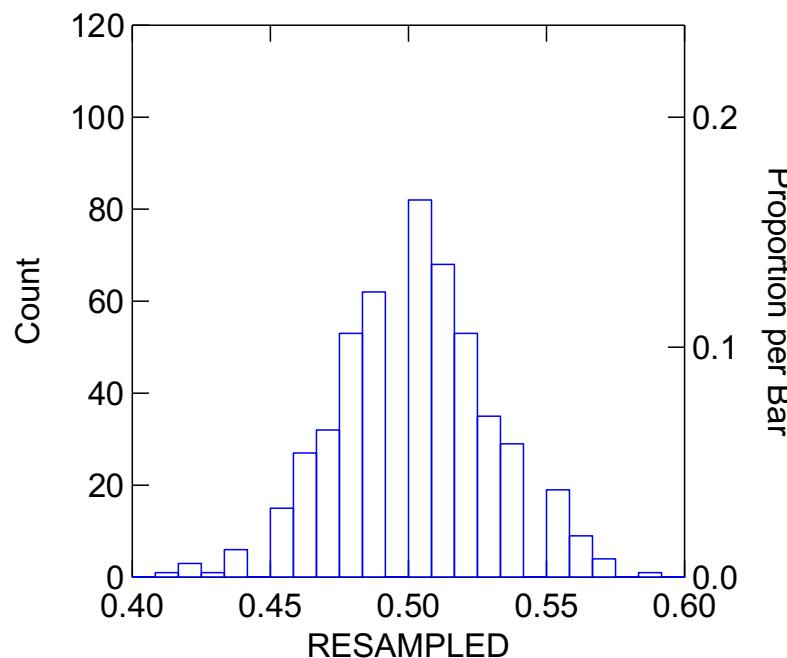
CLTProject.exe calculates the mean of 500 samples, each of size 100.



File: FIT1006 Lecture 16 CLT.syz

• • •

The randomly generated data was saved as text, copied and pasted into CLTProject.exe. 500 samples of size 100 were taken and the mean calculated. A histogram of the means is below.



• • •

Comparing the descriptive statistics for both the original data
and the 500 samples of size 100.

	ORIGINAL	RESAMPLED
N of cases	10000	500
Minimum	0.000	0.410
Maximum	1.000	0.590
Median	0.499	0.500
Mean	0.501	0.501*
Standard Dev	0.290	0.028*
N 1 of 4	0.247	0.480
N 2 of 4	0.499	0.500
N 3 of 4	0.754	0.520

Estimation

Part 3. The sampling distribution of means and proportions

Notation, main characters:

Parameter	Population	Sample
Mean	μ	\bar{x}
Standard Deviation	σ	s
Proportion	π	p

$\sigma_{\bar{x}}$ = standard error of the sample mean

σ_p = standard error of the sample proportion

When the population parameters are unknown (which is usually the case) use the sample values as estimates...

The Sampling Distribution of the Mean

From the CLT, if we take a sample of size n ,

From a population with mean μ and variance σ^2

Then, as n increases:

The sample mean, $\bar{x} \rightarrow \mu$, and $\text{variance}(\bar{x}) \rightarrow \frac{\sigma^2}{n}$

thus $\bar{x} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ for n large.

Example 1

A sample of 100 accounts were taken from a population of accounts with mean = \$2000 and standard deviation \$500. What is the probability that the sample mean will be less than 2050?

From the population, $\mu = 2000$, $\sigma = 500$

For the sample, $\bar{x} = 2000$, $n = 100$

$$\text{thus } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{100}} = 50$$

and $\bar{x} \approx N(2000, 50^2)$

$$P(\bar{x} < 2050) = P\left(z < \frac{2050 - 2000}{50}\right)$$

$$= P(z < 1), z \approx N(0, 1^2) = 0.8413$$

The Sampling Distribution of a Proportion

If we take a sample of size n ,

From a population with proportion π of interest

Then, from the CLT, as n increases:

Sample proportion, $p \rightarrow \pi$, variance(p) $\rightarrow \frac{\pi(1-\pi)}{n}$

Thus $p = \pi$, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ for n large,

$np, n(1-p) \geq 5$

Example 2

It is thought that the proportion of left handed people in the population is 10%. What is probability that a sample of 100 people taken at random would have a proportion of left handers less than 0.12?

$$\pi = 0.1, n = 100$$

$$E(p) = 0.1, \text{Var}(p) = \frac{0.1 \times 0.9}{100} = 0.03^2$$

thus $p \approx N(0.1, 0.03^2)$

$$\begin{aligned} P(p < 0.12) &= P\left(z < \frac{0.12 - 0.1}{0.03}\right) \\ &= P(z < 0.67), z \approx N(0, 1^2) = 0.7486 \end{aligned}$$

Estimation

Part 4. Creating a confidence interval for a population parameter

Estimates

Two types of estimates:

- Point Estimates, where we estimate the actual (exact) value of a population parameter.
- Because point estimates are rarely correct it is more usual to define an Interval Estimate. This the range over which we expect the value of the population parameter vary with a given level of confidence.

Deriving a confidence interval

In the following slides a confidence interval is derived based on our understanding of the Normal distribution.

To simplify learning the basic technique, this lecture assumes that we know the population variance (*not true in practice*) or that the sample size is large enough that the Central Limit Theorem is true.

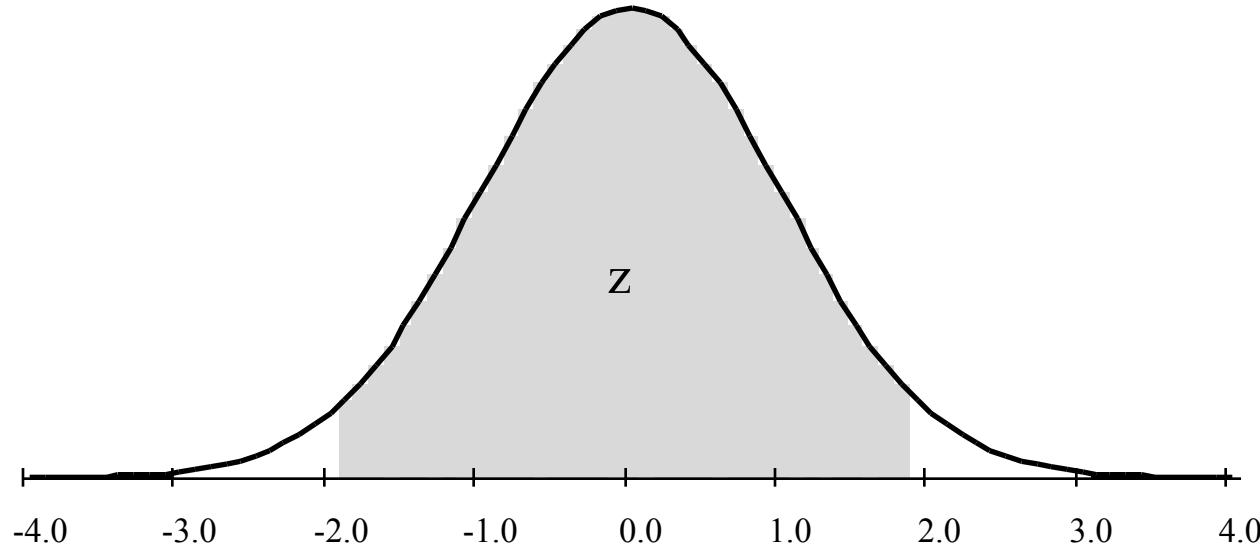
In the following lecture the model is adjusted for the case when sample sizes are small and the population variance is unknown.

95% Confidence Interval

From the Standard Normal distribution:

$$P(-1.96 < z < 1.96) = 0.95.$$

This is a 95% confidence interval for z .

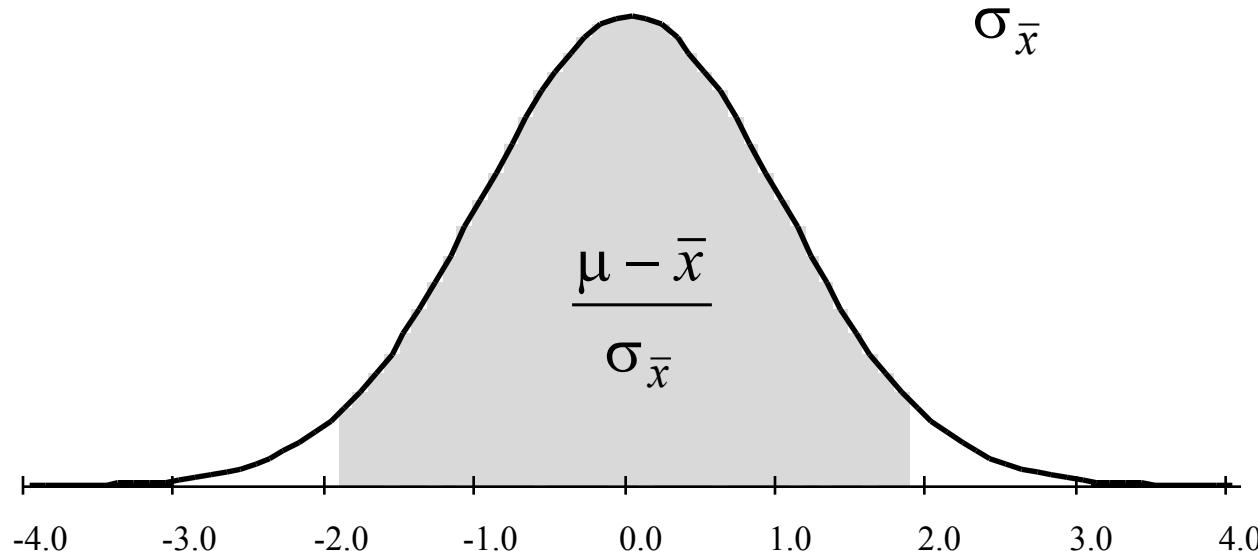


95 % C.I. for μ

Problem: using a sample taken from a population, estimate population mean, μ , using \bar{x} and σ known.

Construct a 95% C.I. for the population mean.

Standardised error of the estimate is $\frac{\mu - \bar{x}}{\sigma_{\bar{x}}}$.



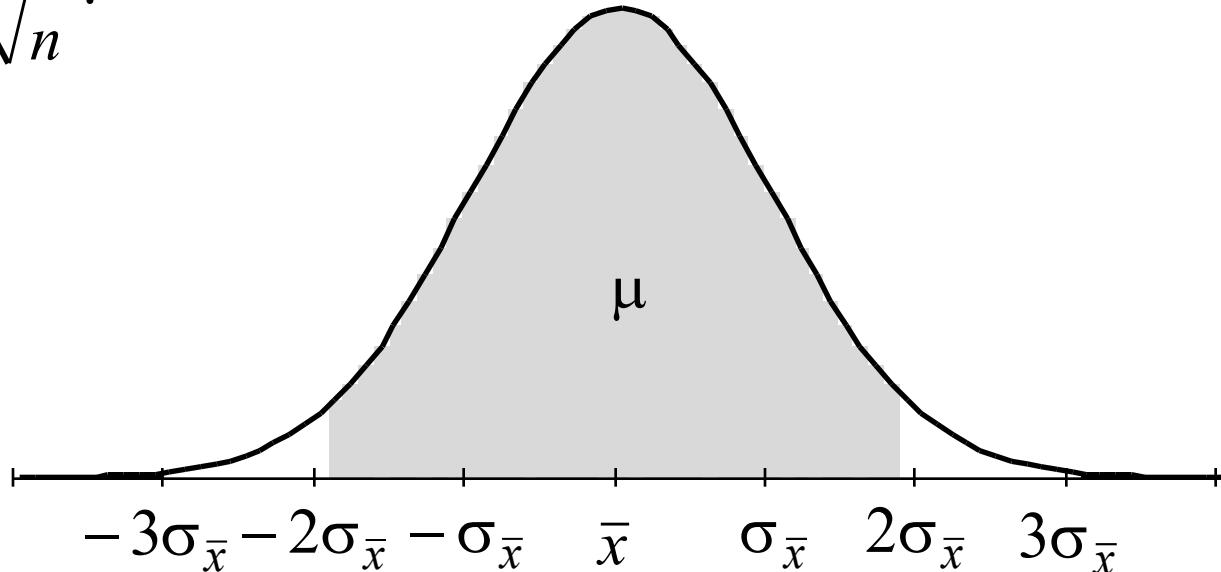
95 % C.I. for μ

Rescale the distribution, by un-standardising so now:

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

So a 95% C.I. for μ is : $\mu = \bar{x} \pm 1.96\sigma_{\bar{x}}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$



95 % C.I. for μ

Algebraic derivation (*you can skip if you want*)

The true value of the population mean is μ which is unknown.

Take a sample, calculate \bar{x} and s (sample mean and st dev).

The standard error (deviation) of \bar{x} is $\sigma_{\bar{x}}$, which is unknown.

The standardised error of the estimate of μ is $\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$,

which has a normal $N(0,1)$ distribution.

95 % C.I. for μ continued

Algebraic derivation (*you can skip if you want*)

$$P(-1.96 < z < 1.96) = 0.95$$

So $P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < 1.96\right) = 0.95$, manipulating

$$P(-1.96\sigma_{\bar{x}} < \bar{x} - \mu < 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

Thus a 95% C.I. for μ based on the sample mean \bar{x} is :

$$\mu = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

Finally, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to calculate C.I.

Example

A sample of 100 students were sampled and their age recorded. Summary statistics: $\bar{x} = 20.1$, $\sigma = 1.2$

Calculate a 95% C.I. for μ , the average age of students at the university.

$$\mu = \bar{x} \pm 1.96\sigma_{\bar{x}} \text{ where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

$$\mu = 20.1 \pm 1.96 \frac{1.2}{\sqrt{100}} = 20.1 \pm 1.96 \times 0.12$$

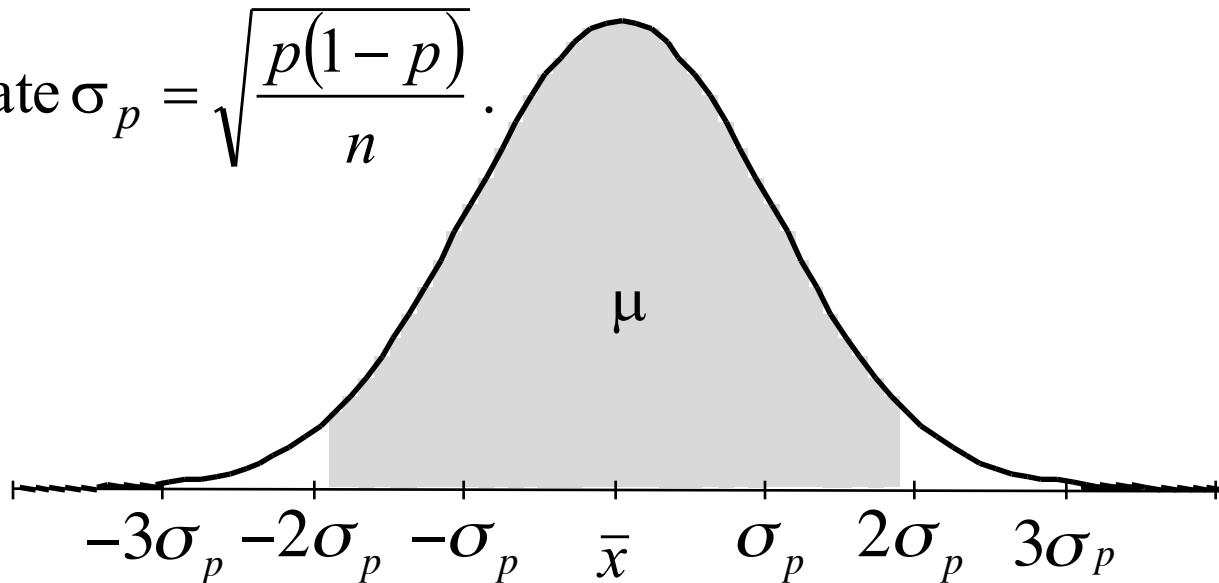
95 % C.I. for π

In the same way that μ was estimated:

$$P(p - 1.96\sigma_p < \pi < p + 1.96\sigma_p) = 0.95$$

So a 95% C.I. for π is : $\pi = p \pm 1.96\sigma_p$

Estimate $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.



Example

A sample of 100 students were sampled and 12 left-handed students counted.

Calculate a 95% C.I. for π , proportion of left-handed students at the university.

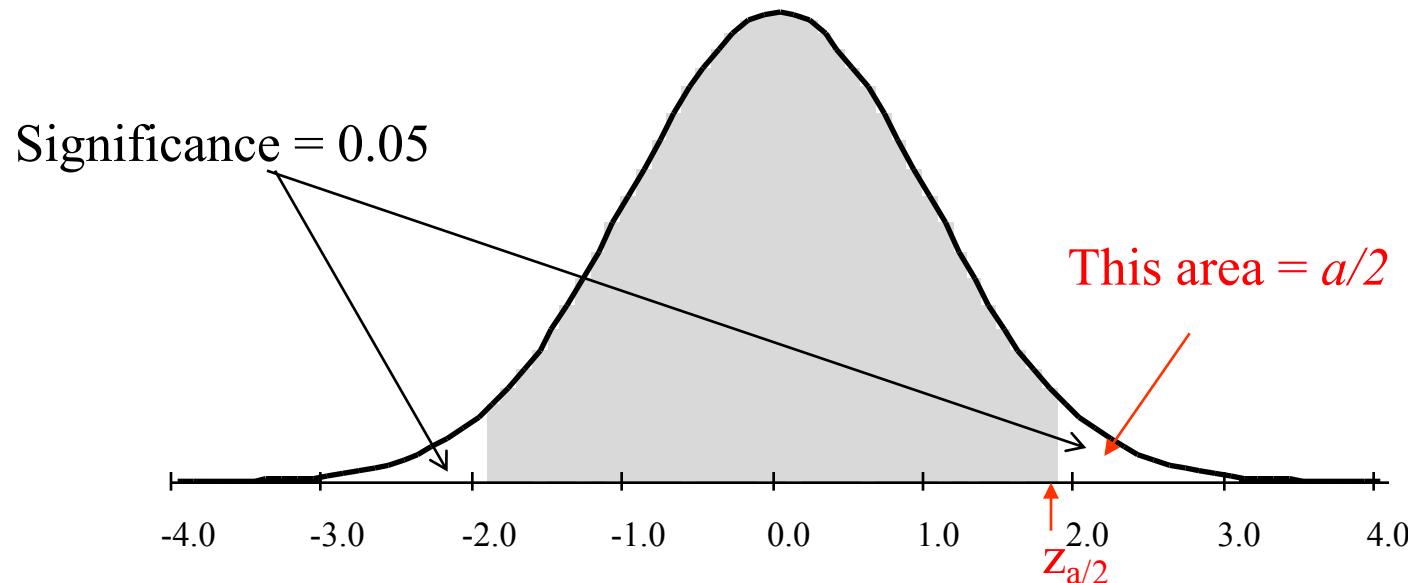
$$\pi = p \pm 1.96\sigma_p \text{ where } \sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$

$$\pi = 0.12 \pm 1.96 \sqrt{\frac{(0.12)(0.88)}{100}} = 0.12 \pm 1.96 \times 0.032$$

Significance

Significance is the probability that the C.I. does not contain the population statistic.

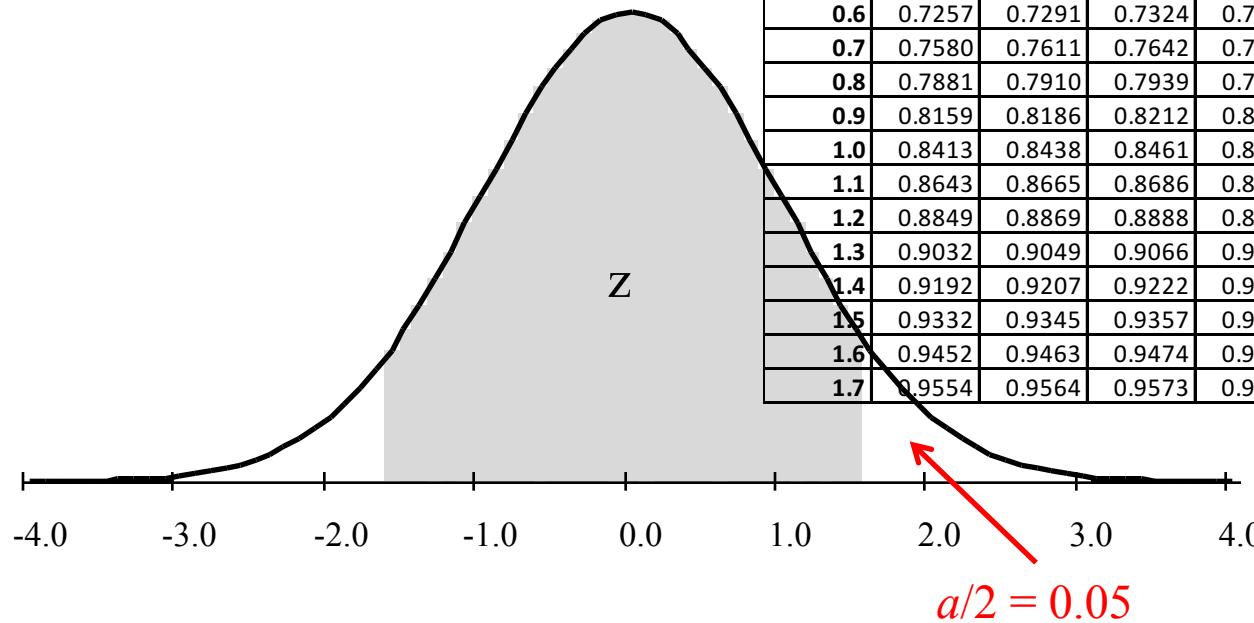
Significance = 1 - Confidence. So a 95% confidence has a significance, $\alpha = 0.05$.



90% Confidence Interval

We want $P(-? < z < ?) = 0.90$.

$$P(-1.645 < z < 1.645) = 0.90.$$



Cumulative Probabilities for the Standard Normal Distribution						
z	Table gives $P(Z < z)$ for $Z \sim N(0,1)$					
	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599

A General C.I. for μ and π

Based on the normal distribution a confidence interval at the α significance is.

$$\mu = \bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}} \text{ where } \sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The sample standard deviation s is used as an estimate of the population standard deviation, σ .

The C.I. for p is created the same way.

$$\pi = p \pm z_{\alpha/2} \sigma_p \text{ where } \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Sums and Differences of Variables

Consider two independent random variables, X and Y:

$$E(X + Y) = E(X) + E(Y)$$

$$E(X - Y) = E(X) - E(Y)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Rule: always add variances.

The Difference of Means

Consider two populations 1 and 2 with means μ_1 and μ_2 .

Let σ_1^2 and σ_2^2 be the population variances.

We take samples of size n_1 and n_2 .

Let \bar{X}_1 and \bar{X}_2 be the sample means, Then :

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

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma_{\bar{X}_1 - \bar{X}_2}^2$$

The Difference of Proportions

Consider two populations 1 and 2 with population proportions π_1 and π_2 .

We take samples of size n_1 and n_2

Let P_1 and P_2 be the sample proportions

Then :

$$E(P_1 - P_2) = \pi_1 - \pi_2$$

$$Var(P_1 - P_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Difference of means and proportions

When finding the confidence interval for the difference of means or proportions use the following to calculate standard error.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Reading (Selvanathan)

Reading: Estimation

7th Ed. Chapter 9 + Sections 10.1, 10.2, 10.3.

Next lecture

Small samples.

The t-distribution – which adjusts the C.I. when σ is estimated from the data by s and corrects for small samples.

Setting the sample size for a required level of accuracy.

FIT1006 Lecture 18

Small samples.

The t-Distribution.

Setting the sample size.

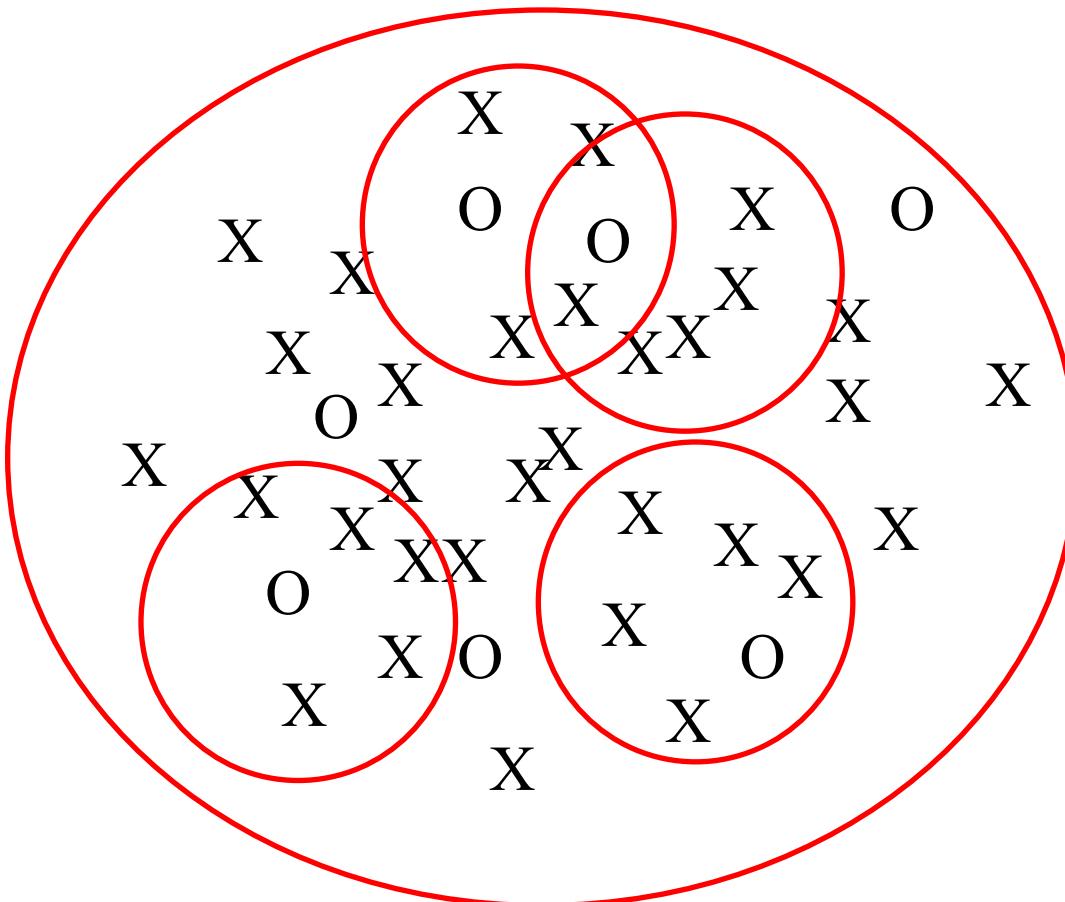
Small Samples

The one of the fundamental assumptions of the Central Limit Theorem is that of large sample sizes are used.

‘Large’ means at least 30 in practice.

When sample sizes are small and the variance of the population unknown, the Normal distribution cannot be used as the basis of a confidence interval. Instead the t-Distribution is used.

Populations and Samples



Student's t-Distribution

The t-Distribution was derived by W. S. Gosset, a scientist working for the Guinness brewery. He published under the pseudonym ‘Student.’ As a consequence the distribution is commonly known as Student’s t distribution.

The t-Distribution has three parameters, μ , σ and ‘degrees of freedom’, v .

The t distribution is (heavy-tailed) for small values of n . As n increases, the shape of the t-Distribution becomes closer to the Normal distribution.

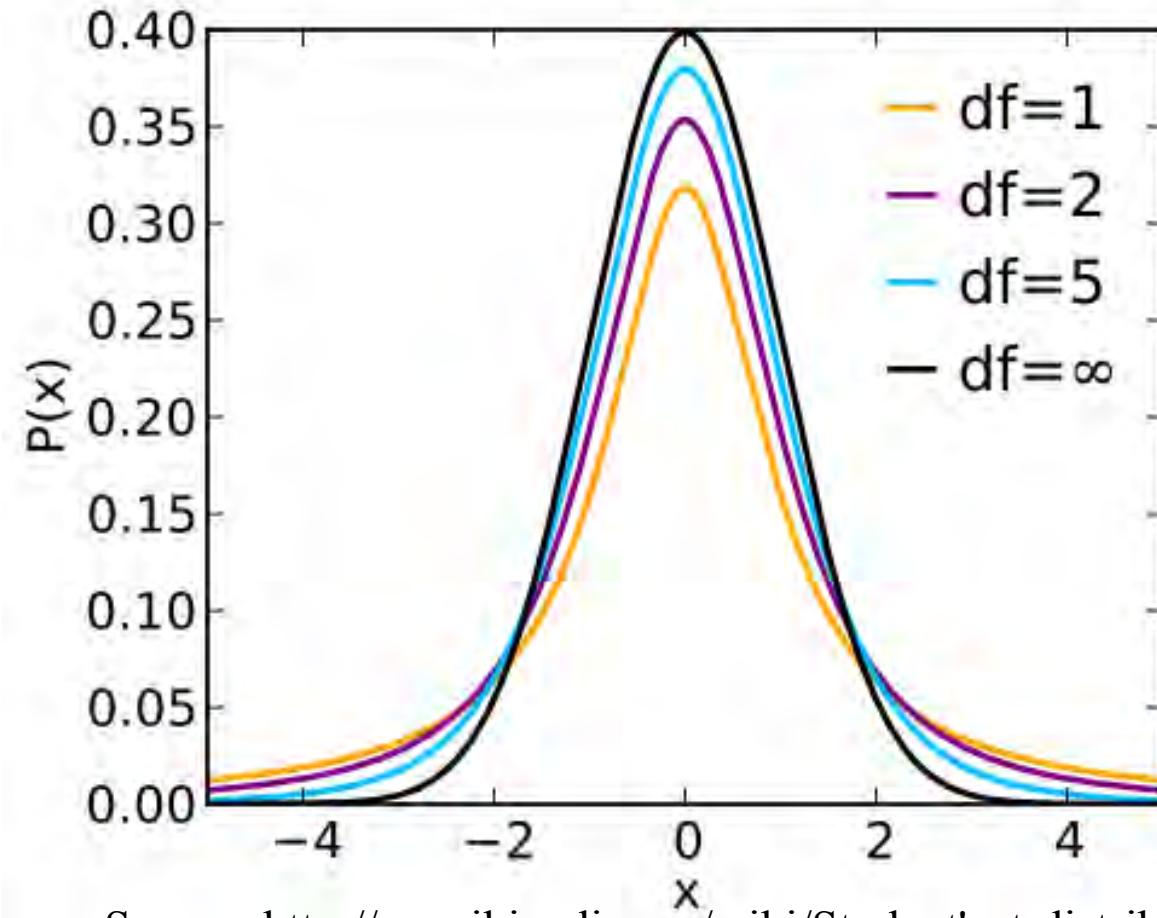
Degrees of Freedom

The number of degrees of freedom or v , refers to the number of observations that are free to vary when determining the variance or standard error of a sample.

The general rule for calculating the number of degrees of freedom is to count the number of observations and subtract 1 for each statistic that is derived from the sample.

In practice, for one-sample problems, v equals the number of observations less 1 (because the *derived* (estimated) sample mean).

Comparison of t and z



Source: http://en.wikipedia.org/wiki/Student's_t-distribution

Tables for the t-Distribution.

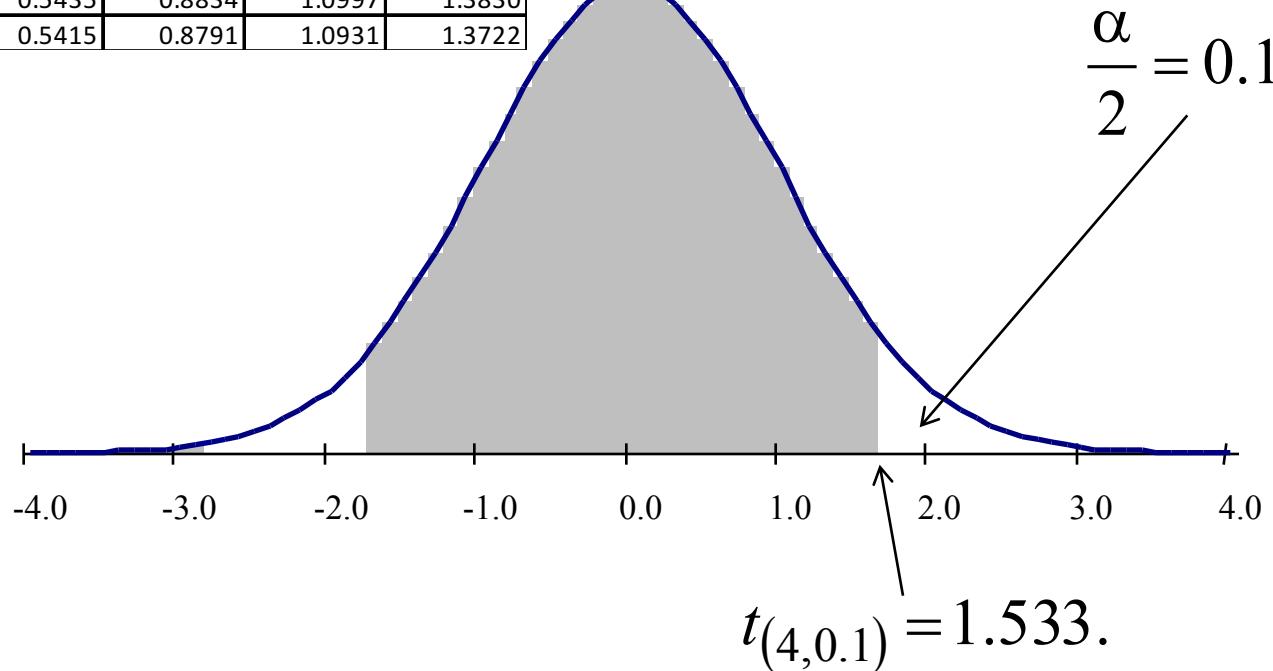
On Excel file PROBDIST.XLS

Critical Values of the t Distribution								
n	<i>a</i>							
	0.300	0.200	0.150	0.100	0.050	0.025	0.010	0.005
1	0.7265	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.6172	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.5844	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.5686	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.5594	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.5534	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.5491	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.5459	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.5435	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.5415	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.5399	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.5386	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.5375	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.5366	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.5357	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467

Upper critical value

n	0.300	0.200	0.150	0.100
1	0.7265	1.3764	1.9626	3.0777
2	0.6172	1.0607	1.3862	1.8856
3	0.5844	0.9785	1.2498	1.6377
4	0.5686	0.9410	1.1896	1.5332
5	0.5594	0.9195	1.1558	1.4759
6	0.5534	0.9057	1.1342	1.4398
7	0.5491	0.8960	1.1192	1.4149
8	0.5459	0.8889	1.1081	1.3968
9	0.5435	0.8834	1.0997	1.3830
10	0.5415	0.8791	1.0931	1.3722

Upper critical value is based on upper region.



Example 1

Five experiments were conducted to determine the amount of silica in water, measured in parts per million (ppm).

Data: 229, 255, 280, 203, 229.

Estimate the mean amount of silica using a 99% confidence interval.

Solution

$\bar{x} = 239$ and $s = 29.3$

$\alpha = (1 - \text{confidence level}) = (1 - 0.99) = 0.01$, Thus $\frac{\alpha}{2} = 0.005$.

The sample size is 5, hence $\text{DOF}(\nu)$ is 4.

From tables of the t - distribution $t_{(4, 0.005)} = 4.604$.

A 99% CI for μ is $\mu = \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$.

Thus a 99% CI is $\mu = 239.2 \pm 4.604 \left(\frac{29.3}{\sqrt{5}} \right)$.

i.e. $\mu = 239.2 \pm 60.3$ ppm at the 99% confidence level.

Confidence Intervals in SYSTAT

The descriptive statistics menu in SYSTAT determines 95% confidence intervals by default, but can be set to any value. Using the data from the previous question.

SILICA_PPM	
N of cases	5
Minimum	203.000
Maximum	280.000
Mean	239.200
95% CI Upper	275.575
95% CI Lower	202.825
Standard Dev	29.295

SILICA_PPM	
N of cases	5
Minimum	203.000
Maximum	280.000
Mean	239.200
99% CI Upper	299.519
99% CI Lower	178.881
Standard Dev	29.295

Example 2

A shop reported the following numbers of shoppers over two weeks. Calculate a 95% confidence interval for the average number of customers.

99 179 126 156 132 31 122 126 123 150 158 160 67 111

Descriptive statistics are:

SHOPPERS	
N of cases	14
Minimum	31.000
Maximum	179.000
Mean	124.286
Standard Dev	39.169

Solution

From the data : $\bar{x} = 124.3$, $\sigma_{\bar{x}} = 10.5$, $t_{0.025(13)} = 2.160$

$$\begin{aligned}95\% C.I. &= 124.3 \pm 2.160 \times 10.5 \\&= (101.7, 146.9)\end{aligned}$$

SHOPPERS	
N of cases	14
95% CI Upper	146.901
95% CI Lower	101.670

Pooled Samples – Diff. of means

The usual way to calculate the standard error

For the difference of means is: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

However, when we have two small samples of similar variance it is possible to calculate the variance of the ‘pooled’ sample which gives a smaller standard error. See following slide.

Pooled Samples – C.I. Calculations

We can determine a confidence interval for the difference of population means for small samples using the variance of the pooled sample.

Suppose we have \bar{x}_1 and \bar{x}_2 , s_1^2 and s_2^2 we wish to find a C.I. for $\mu_1 - \mu_2$.

We assume both populations have the same variance and make an estimate of the population standard deviation with the formula

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \text{ and standard error } s_{\bar{x}_1 - \bar{x}_2} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

We use the t distribution with degrees of freedom $v = n_1 + n_2 - 2$.

Our $(1 - \alpha)$ confidence interval is given by $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$

Pooled Samples – Example

The number of claims processed by two workers is measured over a period of (different) days.

Worker A: 23, 45, 21, 22, 17, 42, 45, 41, 49, 19.

Worker B: 33, 23, 19, 51, 32, 15.

Calculate a 95% C.I. For the difference in the average number of claims (A-B) processed by the workers.

Pooled Samples – Summary Stats

	Worker A	Worker B
	23	33
	45	23
	21	19
	22	51
	17	32
	42	15
	45	
	41	
	49	
	19	
N	10.00	6.00
Mean	32.40	28.83
St Dev	12.92	12.97

Pooled Samples – Finished C.I.

From the data :

$$\bar{x}_1 = 32.40, \quad s_1 = 12.92, \quad n_1 = 10, \quad \bar{x}_2 = 28.83, \quad s_2 = 12.97, \quad n_2 = 6$$

$$s = \sqrt{\frac{9 \times 12.92^2 + 5 \times 12.97^2}{14}} = 12.94$$

$$s_{\bar{x}_1 - \bar{x}_2} = 12.94 \sqrt{\frac{1}{10} + \frac{1}{6}} = 6.68$$

$$t_{0.025(14)} = 2.147$$

$$\begin{aligned} 95\% C.I. &= (32.40 - 28.83) \pm 2.147 \times 6.68 = 3.57 \pm 14.34 \\ &= (-10.78, 17.91) \end{aligned}$$

Pooled Samples – SYSTAT Output

Variable	N	Standard	
		Mean	Deviation
<hr/>			
WORKERA	10.000	32.400	12.920
WORKERB	6.000	28.833	12.968

Separate Variance

Variable	Mean Difference	95.00% Confidence Interval			t	df	p-Value
		Lower Limit	Upper Limit				
<hr/>							
WORKERA	3.567	-11.214	18.348	0.533	10.634	10	0.605
WORKERB							

Pooled Variance

Variable	Mean Difference	95.00% Confidence Interval			t	df	p-Value
		Lower Limit	Upper Limit				
<hr/>							
WORKERA	3.567	-10.762	17.896	0.534	14.000	14	0.602
WORKERB							

Factors Affecting Sample Size

Factors affecting width of confidence interval:

- The degree of confidence required, 99, 95, 90% etc.
- The number of degrees of freedom for small samples.
- The standard error of the estimate.

Degrees of Freedom and Standard Error diminish as sample size increases.

For $n > 30$, the values of the t-Distribution are close enough to the Normal distribution and so we must adjust sample size to further reduce standard error.

Choosing a Sample Size

The confidence level for estimating the population mean is

$$\mu = \bar{x} \pm Z_{\alpha/2} s_{\bar{x}}$$

Thus, $Z_{\alpha/2} s_{\bar{x}}$ is half the width of the confidence interval.

Suppose we want to ensure that the half width is less than a desired value, E . We want $Z_{\alpha/2} s_{\bar{x}} < E$. But $s_{\bar{x}} = s / \sqrt{n}$.

We want a value of n such that $\frac{Z_{\alpha/2} s}{\sqrt{n}} \leq E$, that is, $n \geq \left(\frac{Z_{\alpha/2} s}{E} \right)^2$.

Example 4

A bank is interested in determining the average disposable income of its customers. From a pilot study they estimate the standard deviation of average disposable income to be \$90. How many customers should they sample if they want to obtain an accuracy of \$5 at the 95% level?

Solution

Using a one sided calculation :

$$n \geq \left(\frac{z_{\alpha/2} s}{E} \right)^2$$

$$\geq \left(\frac{1.96 \times 90}{5} \right)^2$$

≥ 1244.6 or 1245

$$\left(\frac{\bar{x}}{\$5 \quad | \quad \$5} \right)$$

What You Should Know

You should have some idea of degrees of freedom and be able to read the table for the t distribution.

You should be able to calculate a confidence interval for the population mean based on a small sample.

You should be able to calculate the required sample size for a given confidence interval.

Reading (Selvanathan)

Reading: Estimation

7th Ed. Sections 10.3, 10.5

FIT1006 Lecture 19

Hypothesis Testing

Motivating Problem

Would Labor have won a Federal Election held February 2018?

The Australian Newspoll had the two-party preferred vote at: Labor 53% Liberal-NP 47% from a sample of approximately 1,160 people chosen at random.

<http://newspoll.com.au/>

Statistical Inference

Statistical inference is concerned with the way we draw conclusions about the population using a sample.

There are two approaches to statistical inference:

- confidence intervals for population parameters,
- hypothesis testing, in which we test an assumption, or point of view about a population parameter.

The two approaches are closely related.

Chance Difference?

When we are hypothesis testing, we are attempting to determine whether the sample statistic could plausibly have come from a population having a certain parameter value, or whether it just occurred by chance.

For example, if we take a sample of 100 student test results and find the sample mean is 37. We would be more accepting of the hypothesis that the population mean is 40 rather than 60, as it is unlikely that a population with mean 60 would yield a sample of 100 having a mean of 37.

But at what value would your acceptance change?

What is an Hypothesis?

An hypothesis is an assumption, or statement about a population parameter. Some possible hypotheses are:

- The population mean = 50.
- The proportion of left handed people in the class is 0.1.

We then test the hypothesis in a systematic way and state a conclusion based on the test.

The Null and Alternate Hypothesis

When we test the hypothesis, we test against an alternative. For example, the hypothesis:

- H_0 : The population mean = 50
Could be tested against the alternative hypothesis:
- H_1 : The population mean > 50.

We use H_0 to denote the null hypothesis – this is the hypothesis that our population parameter has no difference from a particular value. The alternative hypothesis is denoted H_1 .

One and Two Sided Tests

H_0 : The population mean = 50

Could be tested against **one sided** hypothesis:

H_1 : The population mean > 50.

Reject the hypothesis for high values of the sample statistic.

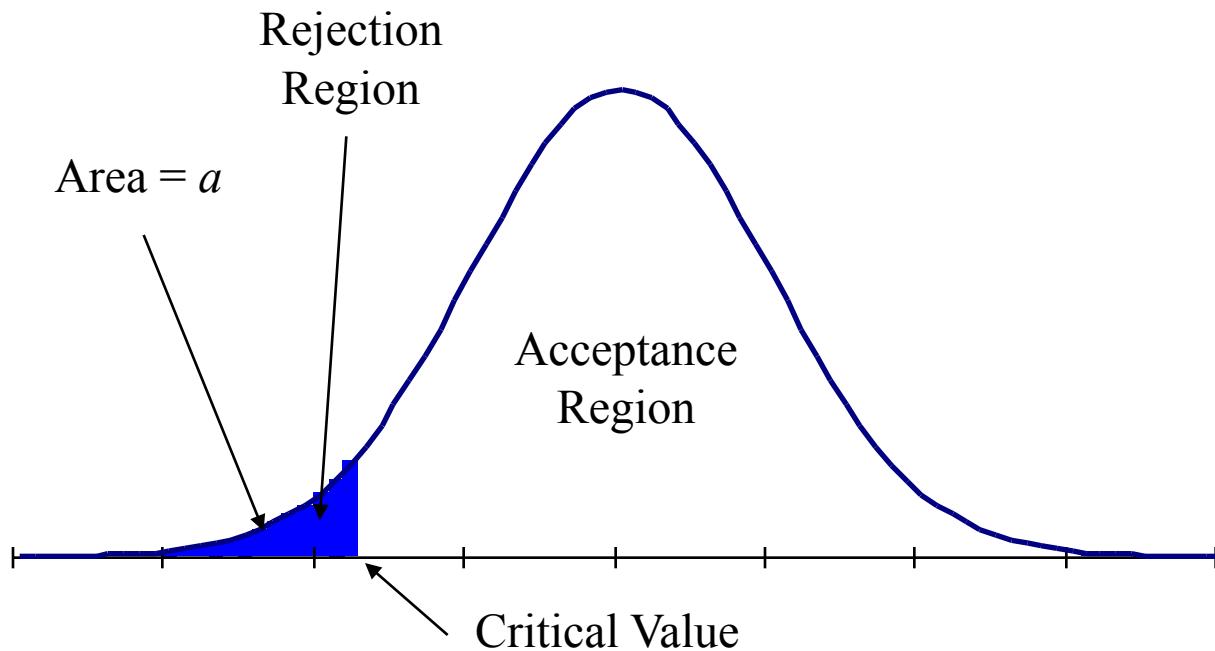
Alternatively H_0 could be tested against the **two sided** hypothesis:

H_1 : The population mean \neq 50.

Reject the hypothesis for high or low values of the sample statistic.

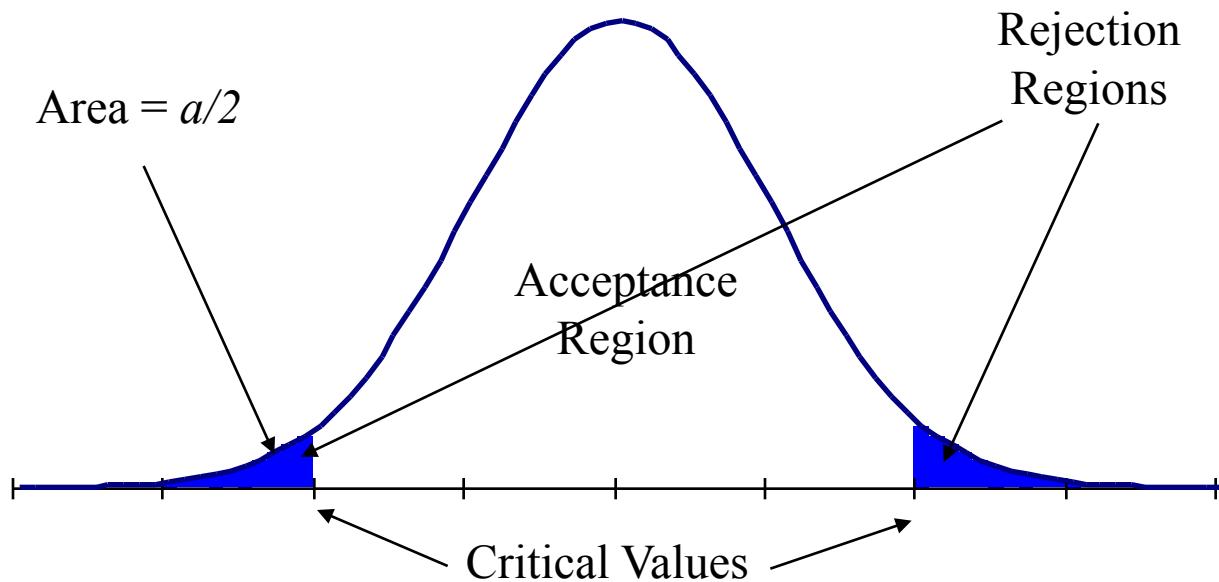
One Sided Test: rejection region

For a one sided test there is an upper or lower rejection region. Reject H_0 if the statistic lies the rejection region as defined by our test.

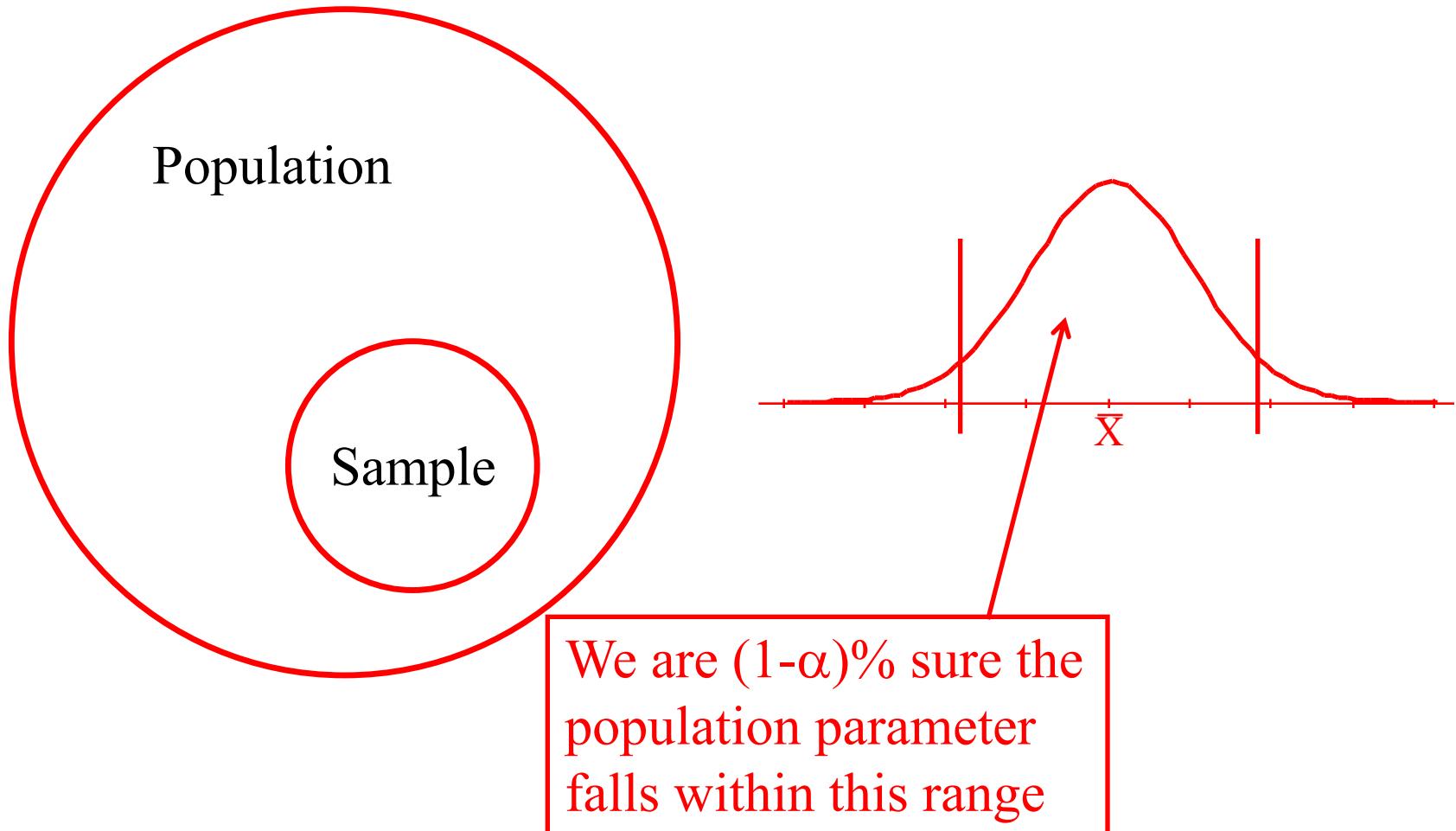


Two Sided Test: rejection region

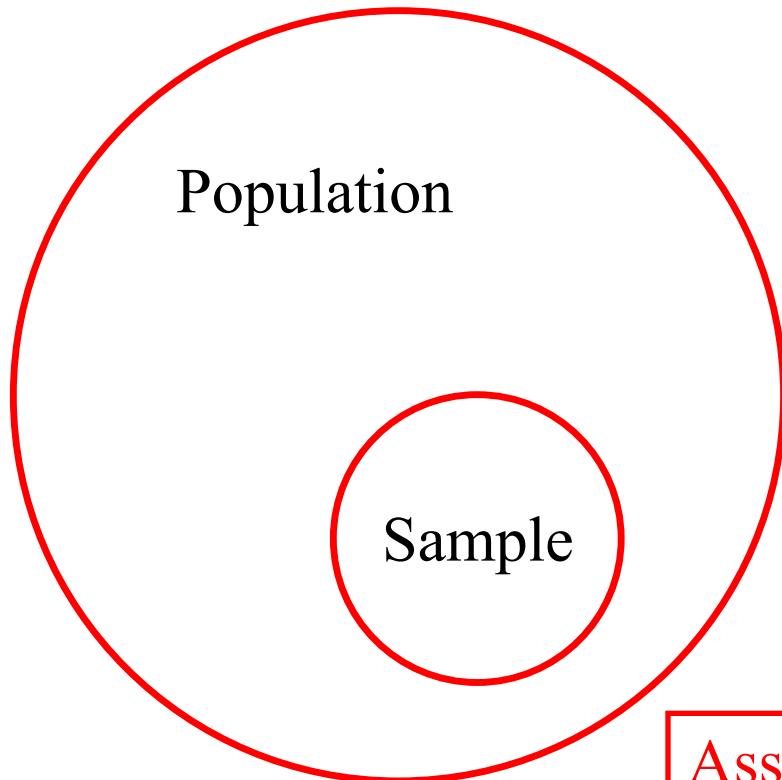
For a two sided test there is an upper and lower rejection region. Reject H_0 if the statistic lies in either region.



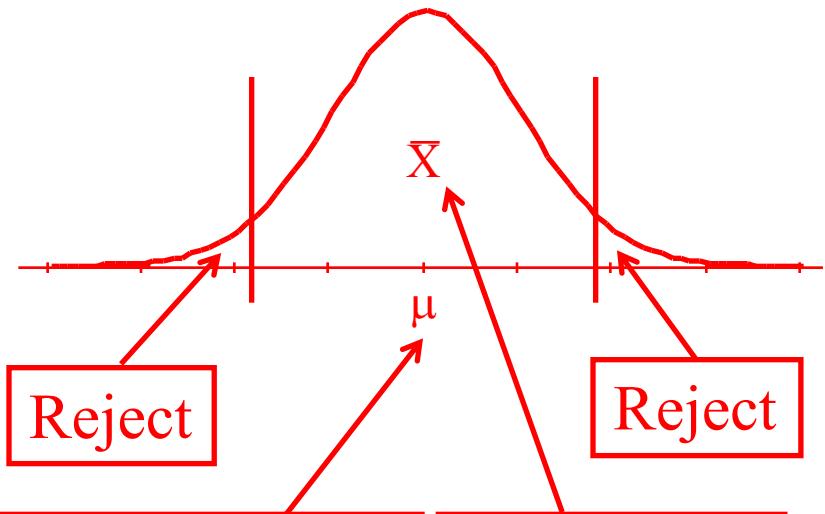
Estimation



Hypothesis testing



Set critical value(s) to be $(1-\alpha)\%$ certain about μ .



Assume population parameter is μ .

Accept assumption

The Steps of Hypothesis Testing

- 1 Decide on a null hypothesis H_0 .
- 2 Decide on an alternative hypothesis H_1 .
- 3 Decide on a significance level.
- 4 Calculate the appropriate test statistic.
- 5 Find from tables the corresponding tabulated test statistic.
- 6 Compare calculated and tabulated test statistics and decide whether to accept or reject the null hypothesis.
- 7 State the conclusion and assumptions of the test.

Source Rees, D.G. Essential Statistics, Chapman and Hall 1995.

Example 1

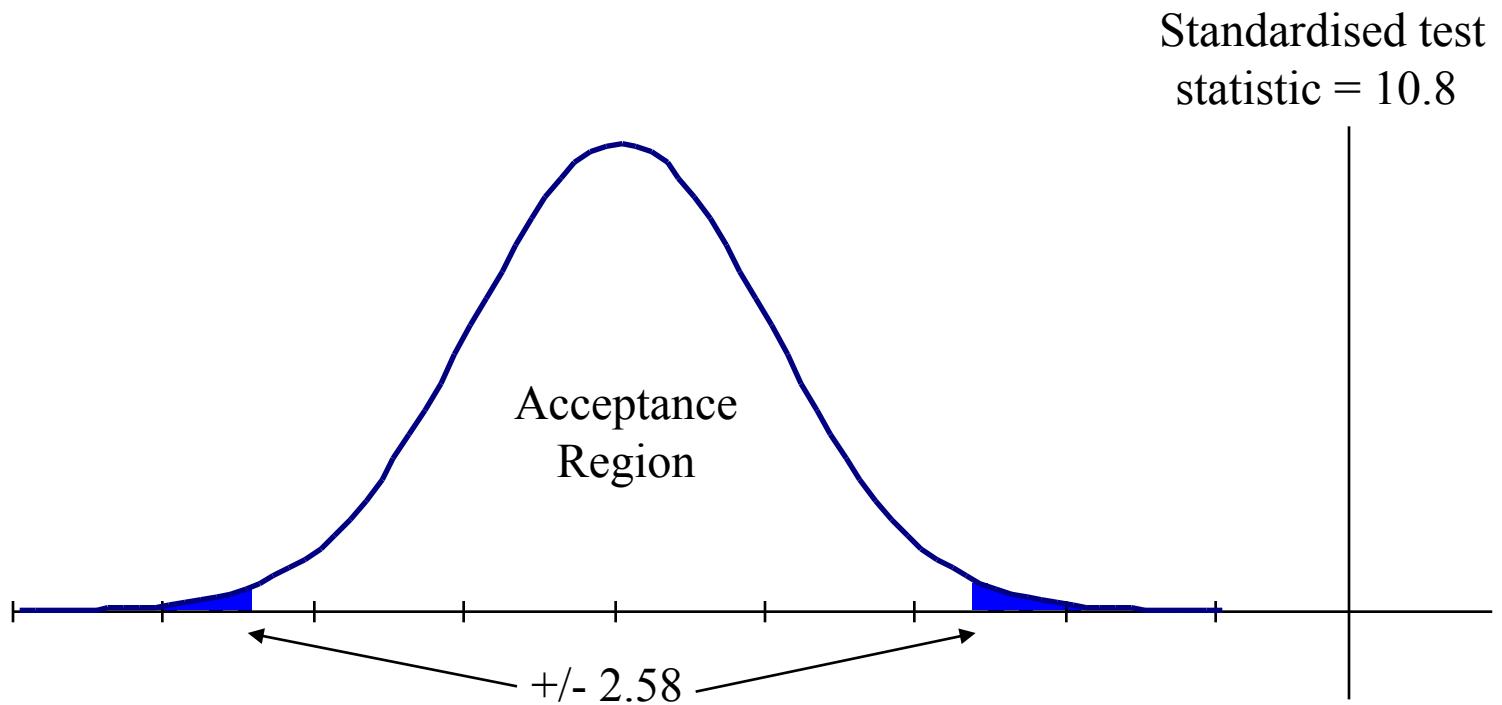
A hypothesis test for the population mean when the population variance is known. (Note: unrealistic situation that lets us use the Normal distribution)

- The Axle manufacturing company has been making axles for a long time and kept records for every axle produced.
- Population parameters are $\mu = 90\text{mm}$ and $\sigma = 2.5\text{mm}$.
- A sample of 100 axles from new machine has mean = 92.7.
- Is the new machine making parts with same average length required (90mm)?
- Assume a 1% significance.

Solution

- 1 $H_0, \mu = 90\text{mm}$
- 2 $H_1, \mu \neq 90\text{mm}$ (a two sided experiment)
- 3 Significance = 0.01.
- 4 The test statistic, $\bar{x} = 92.7$. We calculate Z_x .
(standardising) $Z = (92.7 - 90)/(2.5/\sqrt{100}) = 10.8$
- 5 From tables the calculated critical values are ± 2.58
- 6 We see that $10.8 > 2.58$ and thus we reject H_0 .
- 7 Thus we conclude that the axles produced by the new machine have a mean significantly different from 90mm (1% level), assuming that axle length is normally distributed.

• • •



Example 2

A hypothesis test for the population mean when the population variance is not known.

It is claimed that Melbourne families are spending more than \$150 per week on food and grocery items on average. A sample of 15 families was surveyed and the amount spent each week was recorded. Do these results support this thesis?

Weekly food and grocery expenditure (\$):

156, 234, 199, 78, 256, 189, 221, 49, 220, 178, 120, 290, 97, 177, 231.

Summary Statistics from Excel

Mean	179.7
Standard Error	17.7
Median	189.0
Mode	#N/A
Standard Deviation	68.5
Sample Variance	4697.0
Kurtosis	-0.5
Skewness	-0.5
Range	241.0
Minimum	49.0
Maximum	290.0
Sum	2695.0
Count	15.0

Solution

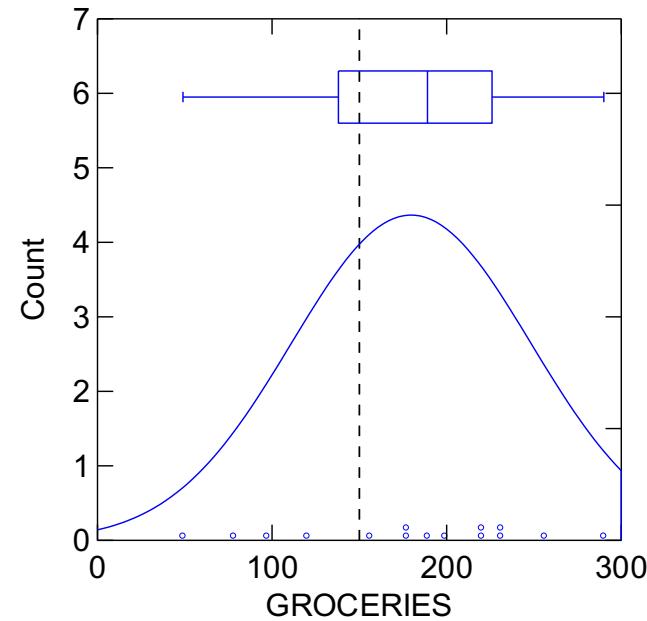
- 1 $H_0, \mu = \$150$
- 2 $H_1, \mu > \$150$ (a one tailed experiment)
- 3 Significance = 0.01.
- 4 The test statistic, $\bar{x} = 179.7$. We calculate T_x . (*standardising*)
$$T = (179.7 - 150) / (68.5/\sqrt{15}) = 1.67$$
- 5 From tables our calculated critical value $T_{(0.01)(v=14)}$ is 2.625
- 6 We see that $1.67 < 2.625$ and thus we do not reject H_0 .
- 7 Thus we conclude that the mean expenditure is not significantly greater than \$150, assuming that expenditure is normally distributed.

SYSTAT Solution

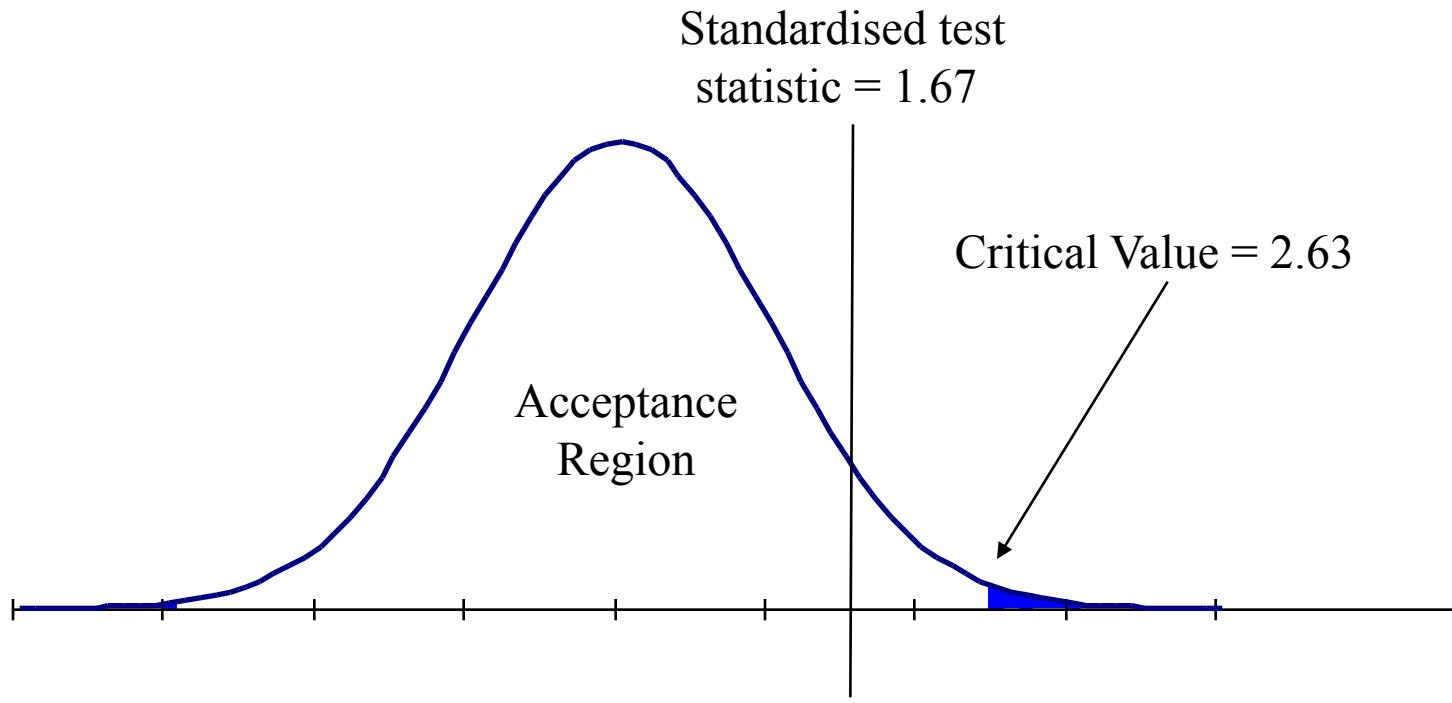
One-sample t-test of GROCERIES with 15 cases

Ho: Mean = 150.000 against Alternative = 'greater than'

Mean	= 179.667
SD	= 68.534
t	= 1.677
df	= 14
p-value	= 0.058



• • •



Example 3

A hypothesis test for a population proportion.

In a contest of two political parties, a party will win if it gets more than 50% of the vote. In a contest between the Liberal Party and the Australian Labor Party for a particular electorate as survey of 237 voters, 180 indicated that they would vote for the ALP in the next election. Test the hypothesis that the ALP will win the next election.

Example 3 continued

We assume that the rules of binomial probabilities hold (that is, n independent trials with probability π of success, $np > 5$ and $n(1-p) > 5$) and test the hypothesis that the population parameter $\pi > 0.5$.

The standard error of the sample is determined using the assumed proportion in the hypothetical distribution, $H_0(\pi)$.

$$p = \frac{180}{237} = 0.7594$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.5(1-0.5)}{237}} = 0.0325$$

Solution

- 1 $H_0, \pi = 0.5$
- 2 $H_1, \pi > 0.5$ (a one tailed experiment)
- 3 Significance = 0.01.
- 4 The test statistic, $p = 0.7594$. We calculate the Z_x .
(standardising) $Z = (0.7594 - 0.5) / 0.0325 = 7.98$
- 5 From tables our calculated critical value $Z_{(0.01)}$ is 2.33
- 6 We see that $7.98 > 2.33$ and thus we reject H_0 .
- 7 Thus we conclude that proportion of voters intending to vote ALP is greater than 50% at the 1% level. Assuming the rules for a binomial probability hold.

Motivating Problem

Would Labor have won a Federal Election held February 2018?

The Australian Newspoll had the two-party preferred vote at: Labor 53% Liberal-NP 47% from a sample of approximately 1,160 people chosen at random.

<http://newspoll.com.au/>

Motivating Problem in Groups

- (a) Is Labor going to win the next election based on these data? (Assume 1% Significance)

$$p = 0.53$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1160}} = \underline{0.0147}$$

$$Z_{Calc} = \frac{\underline{0.53} - \underline{0.50}}{\underline{0.0147}} = 2.041$$

$$Z_{Tables} = 2.33$$

Difference of Proportions

Note that we can also test the hypothesis that the difference of two proportions exceeds a certain value by calculating the test statistic as

$$\hat{\theta} = p_1 - p_2$$

with standard error given by

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Note that this approach assumes that the sample size is large enough to use the sample proportion as an estimate of the population proportion.

Critical values are determined from the Normal distribution.

Reading/Questions (Selvanathan)

Reading: Hypothesis Testing

7th Ed. Sections 12.1, 12.2, 12.3, 12.4, 12.6.

FIT1006 Lecture 20

Hypothesis Testing continued...

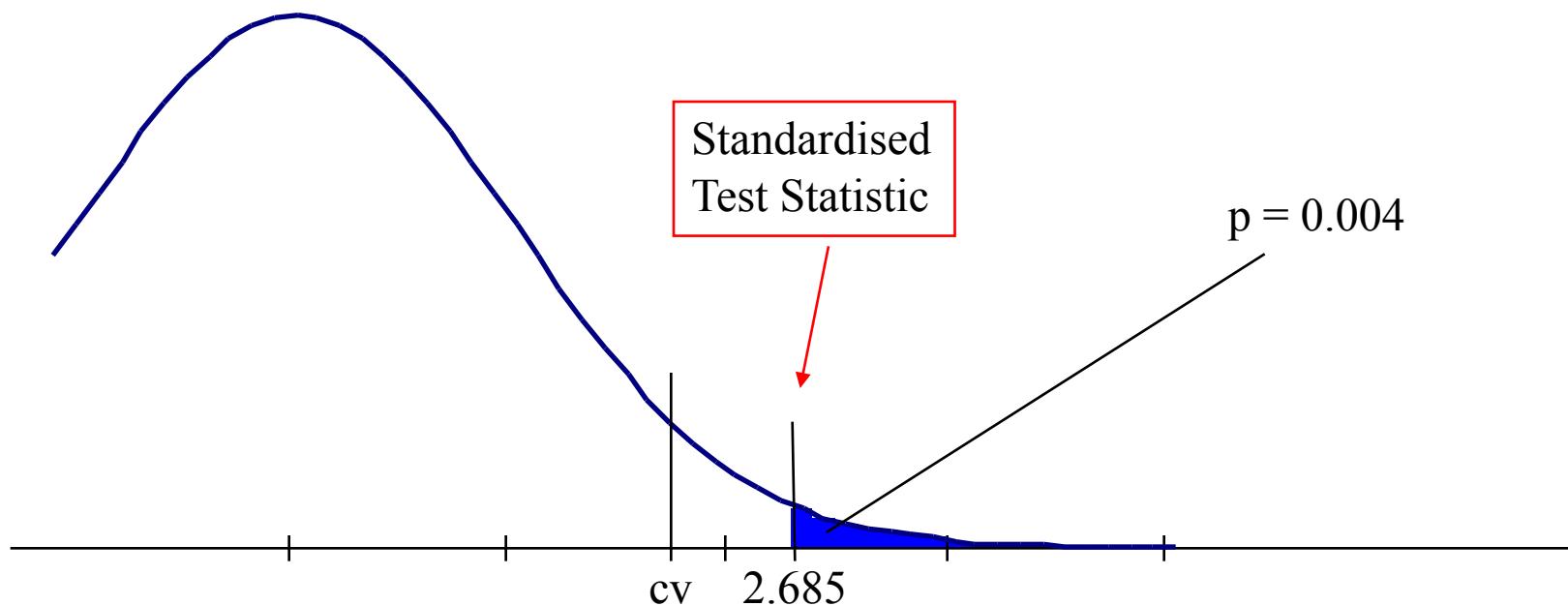
The Steps of Hypothesis Testing

- 1 Decide on a null hypothesis H_0 .
- 2 Decide on an alternative hypothesis H_1 .
- 3 Decide on a significance level.
- 4 Calculate the appropriate test statistic.
- 5 Find from tables the corresponding tabulated test statistic.
- 6 Compare calculated and tabulated test statistics and decide whether to accept or reject the null hypothesis.
- 7 State the conclusion and assumptions of the test.

Source Rees, D.G. Essential Statistics, Chapman and Hall 1995.

p-value

The p-value of a test is the smallest value of the critical region that leads to the rejection of H_0 . For example, if a test had the following test statistic:



Calculating the p-value

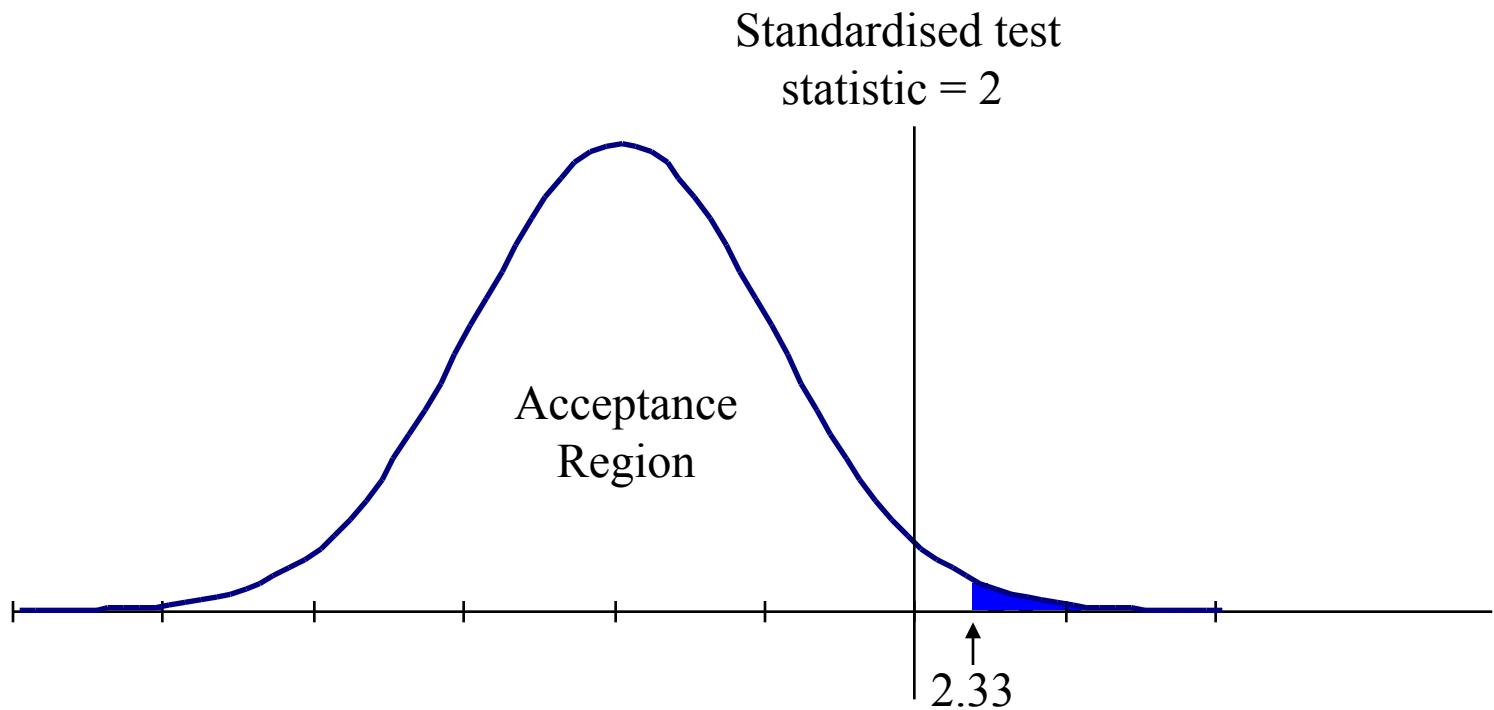
Modified Question from last lecture:

- A hypothesis test for the population mean when the population variance is known. (Note: unrealistic situation)
- The Axle manufacturing company has been making axles for a long time and kept records for every axle produced.
- Population parameters are $\mu = 90\text{mm}$ and $\sigma = 2.5\text{mm}$.
- A sample of 100 axles from new machine has **mean = 90.5**.
- Is the new machine making parts with same average length required (90mm) – **or are they longer than this?**
- **What is the p-value of the test?**

Solution

- 1 $H_0, \mu = 90\text{mm}$
- 2 $H_1, \mu > 90\text{mm}$ (a one-sided experiment)
- 3 Significance = 0.01.
- 4 The test statistic, $\bar{x} = 90.5$. We calculate Z_x .
(standardising) $Z = (90.5 - 90)/(2.5/\sqrt{100}) = 2$.
- 5 From tables the calculated critical value is 2.33.
- 6 We see that $2 < 2.33$ and thus do not reject H_0 .
- 7 Conclude ...
- 8 *p-value* of the test is 0.0228 – this is the highest level of significance at which H_0 will be rejected.

• • •



Errors in Hypothesis Testing

We can make two errors when testing hypotheses

A Type I error occurs when the null hypothesis is correct, but is rejected.

A Type II error occurs when the null hypothesis is incorrect but is not rejected.

		H_0 is Not Rejected	H_0 is Rejected
H_0 is Correct	H_0 is Correct	Correct Decision $(1 - \alpha)$	Type I Error α
	H_1 is Correct	Type II Error β	Correct Decision $(1 - \beta)$

• • •

The distinction between type I and II errors is easier to understand using a criminal law analogy.

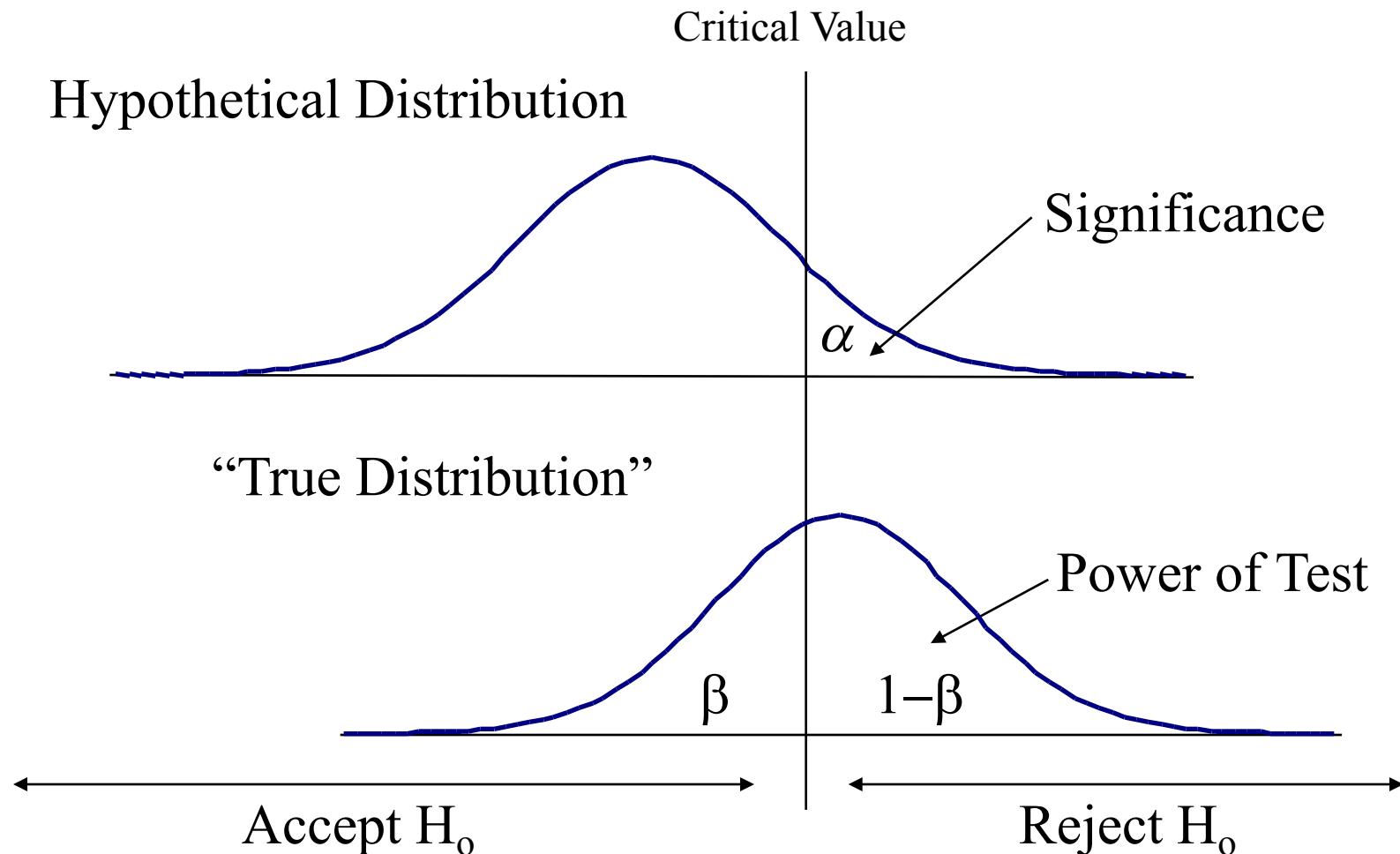
	Person Acquitted	Person Convicted
Person is Innocent	Correct Decision	Type I Error
Person is Guilty	Type II Error	Correct Decision

When we attempt to minimize the type I error, the consequence is an increase in the type II error.

The Power of the Test

- The probability of making a Type II error is β . If the null hypothesis is incorrect, the probability of avoiding a Type II error is $(1-\beta)$, this is the power of the test.
- α and β are interrelated. Increasing α reduces β and *vice versa*. To reduce both, sample size needs to be increased.
- Only α or β (usually α) can be chosen. The true value of β also depends on the actual value of the population parameter – which is often unknown.

The Power of the Test ctd..



Calculating β

Last lecture's example

- The Axle manufacturing company has been making axles for a long time and kept records for every axle produced.
- Population parameters are $\mu = 90\text{mm}$ and $\sigma = 2.5\text{mm}$.
- A sample of 100 axles from new machine has mean = 92.7.
- Is the new machine making parts with same average length required (90mm)?
- Assume a 1% significance.
- What is the power of the test if it eventuates that the new machines produces axles with a mean length of 91mm?

Solution

From our previous calculations

$H_0, \mu = 90\text{mm}$

$H_1, \mu \neq 90\text{mm}$ (a two tailed experiment)

Significance = 0.01.

From tables the critical values are $\pm 2.58\sigma_x$

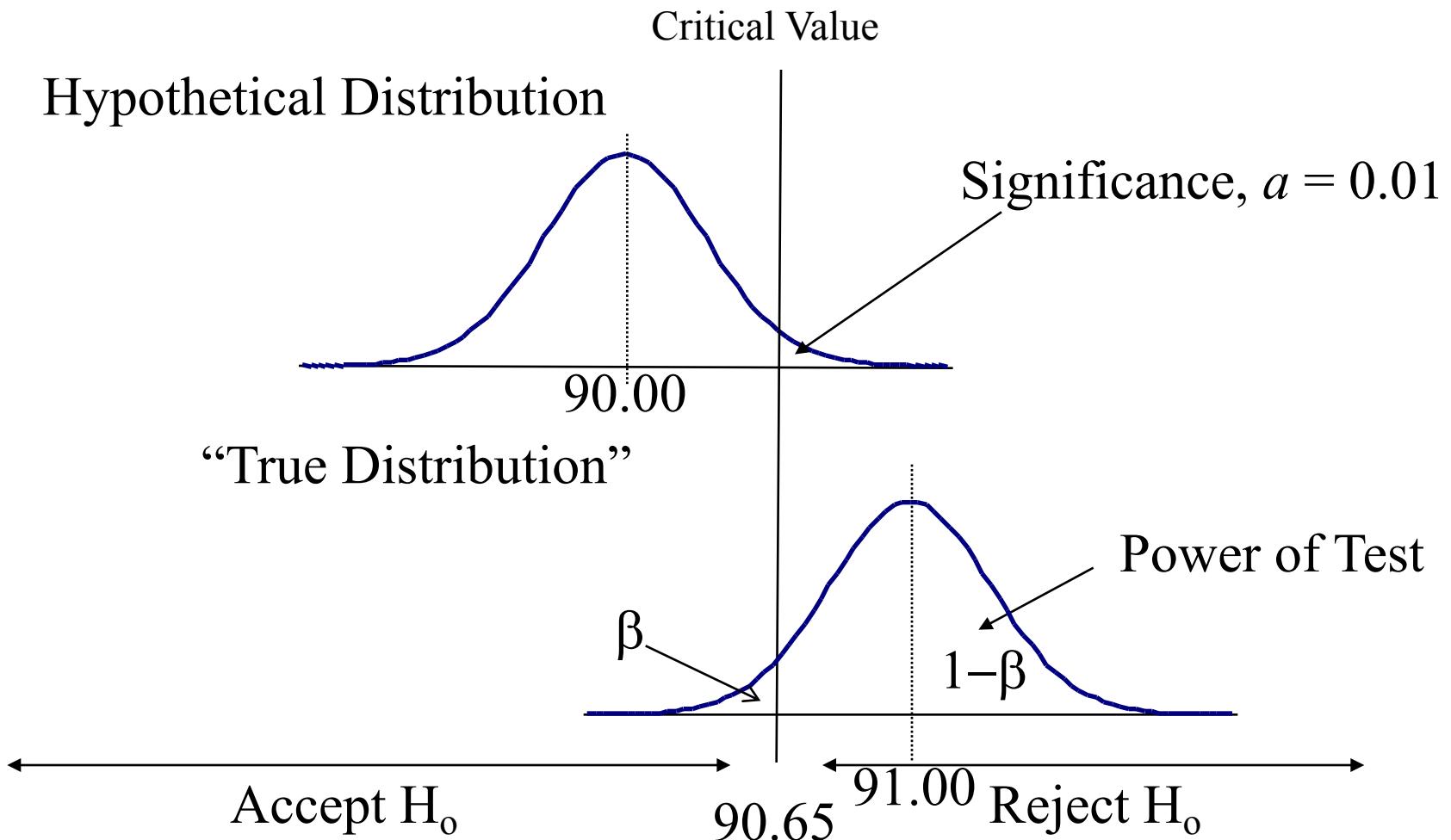
Use upper critical value since alternative is $> H_0$

Calculate upper critical value = $90 + 2.58(0.25) = 90.65$

$P(x < 90.65)$ when $X \sim N(91, 0.25^2)$ is $P(z < -1.42) = 0.08$

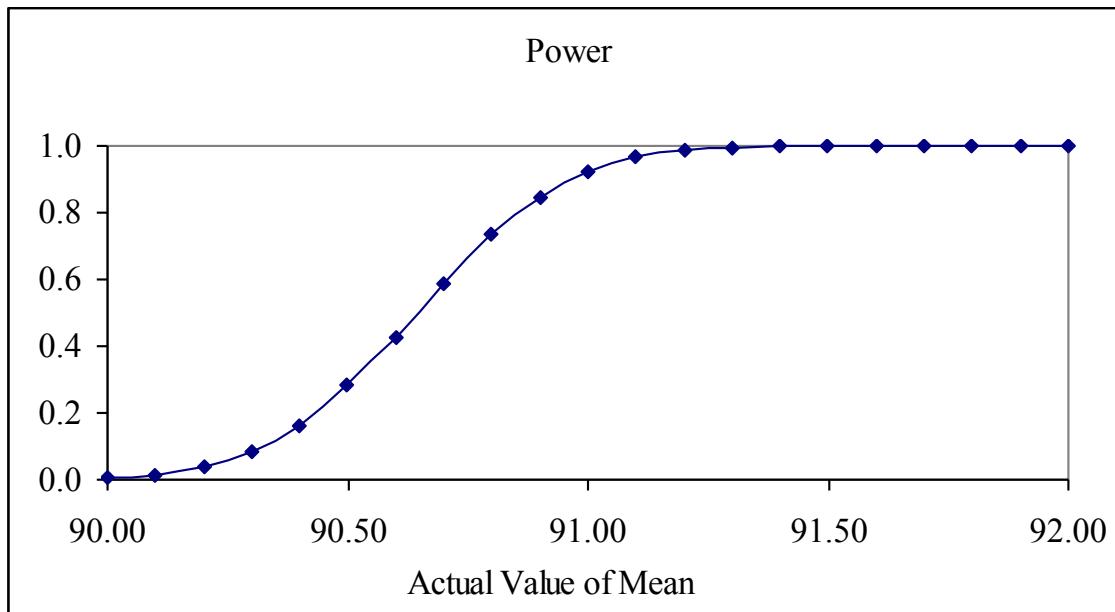
This is the probability that we fail to reject the null hypothesis, when the population mean for the new lathe is 91mm. The power of the test = 0.92

The Power of the Test ctd..



Power Curves

To see how the power of a test changes as the value of the alternative hypothesis changes, we can calculate the power for a range of values and view the resulting curve.



Example

From last lecture...

It is claimed that Melbourne families are spending more than \$150 per week on food and grocery items on average. A sample of 15 families was conducted and the amount spent each week was recorded. Do these results support this thesis? (assume a 1% significance)

Weekly food and grocery expenditure (\$)

156, 234, 199, 78, 256, 189, 221, 49, 220, 178, 120, 290, 97, 177, 231.

What is the power of the test if the average cost of groceries is \$160?

Summary Statistics from Excel

Mean	179.7
Standard Error	17.7
Median	189.0
Mode	#N/A
Standard Deviation	68.5
Sample Variance	4697.0
Kurtosis	-0.5
Skewness	-0.5
Range	241.0
Minimum	49.0
Maximum	290.0
Sum	2695.0
Count	15.0

Discussion in groups:

The power of the test is the probability that H_0 is rejected when it should be rejected...

• • •

$H_0, \mu = \$150$

$H_1, \mu > \$150$ (a one tailed experiment)

Significance = 0.01.

From tables the critical value $T_{(0.01)(v=14)}$ is $2.624\sigma_x$

Calculate upper critical value: $150 + 2.624(17.7) = 196.5$

Assume alternative distribution Normal, new μ , original SE.

$P(x < 196.5)$ when $X \sim N(160, 17.7^2) = P(z < 2.061) = 0.98$

This is β , the probability that we fail to reject the null hypothesis, when the population mean for expenditure is \$160.

The power of the test is 0.02.

Other Hypothesis Tests

We have looked at hypothesis tests for the population mean, a population proportion, and for the difference of means.

There are a range of other hypothesis tests which are regularly used, these include tests for:

- Variance
- Distribution Shape
- Independence.

Reading/Questions (Selvanathan)

Reading: Hypothesis Testing

7th Ed. Sections 12.3, 12.5, 12.6.

FIT1006 Lecture 21

Time Series Analysis and Forecasting

Lecture 21/22 motivating problem

Given the value of building work (quarterly) from Mar 2008 – Dec 2015 create a forecast for 2016 & 2017.

Data source: ABS

<http://www.abs.gov.au>

Season/Year	Value of Building Work (all sectors) \$B
Mar-2008	17.50
Jun-2008	20.24
Sep-2008	21.36
Dec-2008	21.17
Mar-2009	18.00
Jun-2009	18.85
Sep-2009	19.62
Dec-2009	20.71
Mar-2010	19.67
Jun-2010	23.17
Sep-2010	23.64
Dec-2010	22.90
Mar-2011	19.49
Jun-2011	21.16
Sep-2011	21.96
Dec-2011	21.47
Mar-2012	19.67

Motivating problem cont...

If, after building the model, you find out the actual value of building work in 2016 & 2017, calculate the error of the forecast.

Dec-2015	27.01
Mar-2016	25.35
Jun-2016	28.79
Sep-2016	28.47
Dec-2016	29.25
Mar-2017	26.20
Jun-2017	29.22
Sep-2017	30.21
Dec-2017	30.35

1345.0 - Key Economic Indicators, 2018

The screenshot shows the homepage of the Australian Bureau of Statistics (ABS) website. At the top left is the ABS logo. To its right is a search bar with a magnifying glass icon and the word "Search". Below the search bar is a horizontal menu with four items: "Statistics", "Census", "Complete your survey", and "About us". Underneath this menu, a breadcrumb navigation shows the path: "> By Catalogue Number". The main title "1345.0 - Key Economic Indicators, 2018" is displayed prominently. Below it, a sub-header "LATEST ISSUE Released at 11:30 AM (CANBERRA TIME) -1/-1/-1 Released Today" is shown. A green horizontal bar below the title contains five links: "Summary", "Downloads", "Explanatory Notes", "Related Information" (which is highlighted in green), and "Past & Future Releases".

RELATED INFORMATION

[Australian National Accounts: National Income, Expenditure and Product - 5206.0 - Dec 2017](#)

[International Trade in Goods and Services, Australia - 5368.0 - Mar 2018](#)

[Balance of Payments and International Investment Position, Australia - 5302.0 - Dec 2017](#)

[Retail Trade, Australia - 8501.0 - Mar 2018](#)

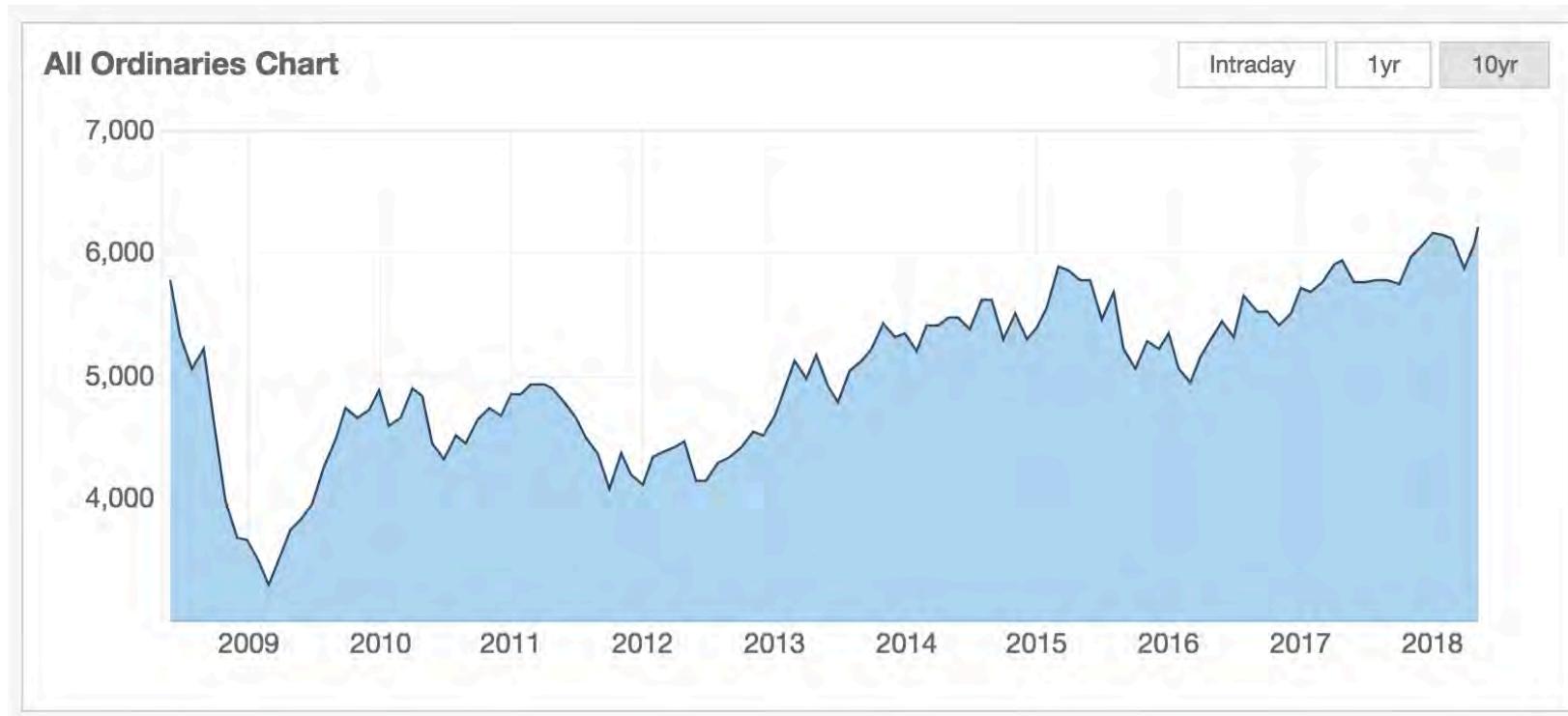
Time Series Data

A Time Series describes a set of observations made over a period of time. Daily maximum temperatures, hourly share prices, annual population counts, weekly sales figures are all examples of time series.

It is usual, but not strictly necessary, that the observations are recorded at equal intervals.

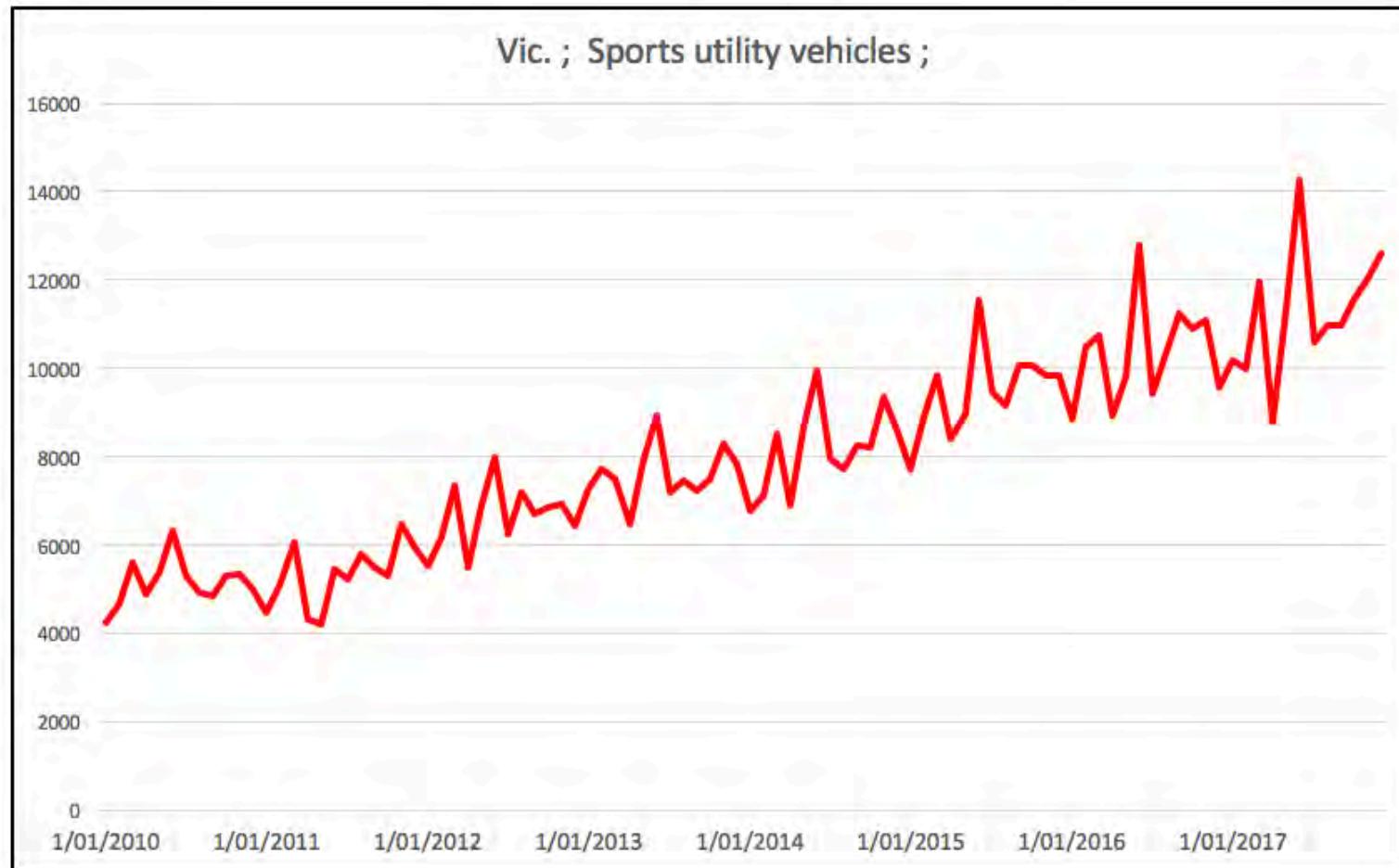
Some examples of time series follow:

Australian All Ordinaries

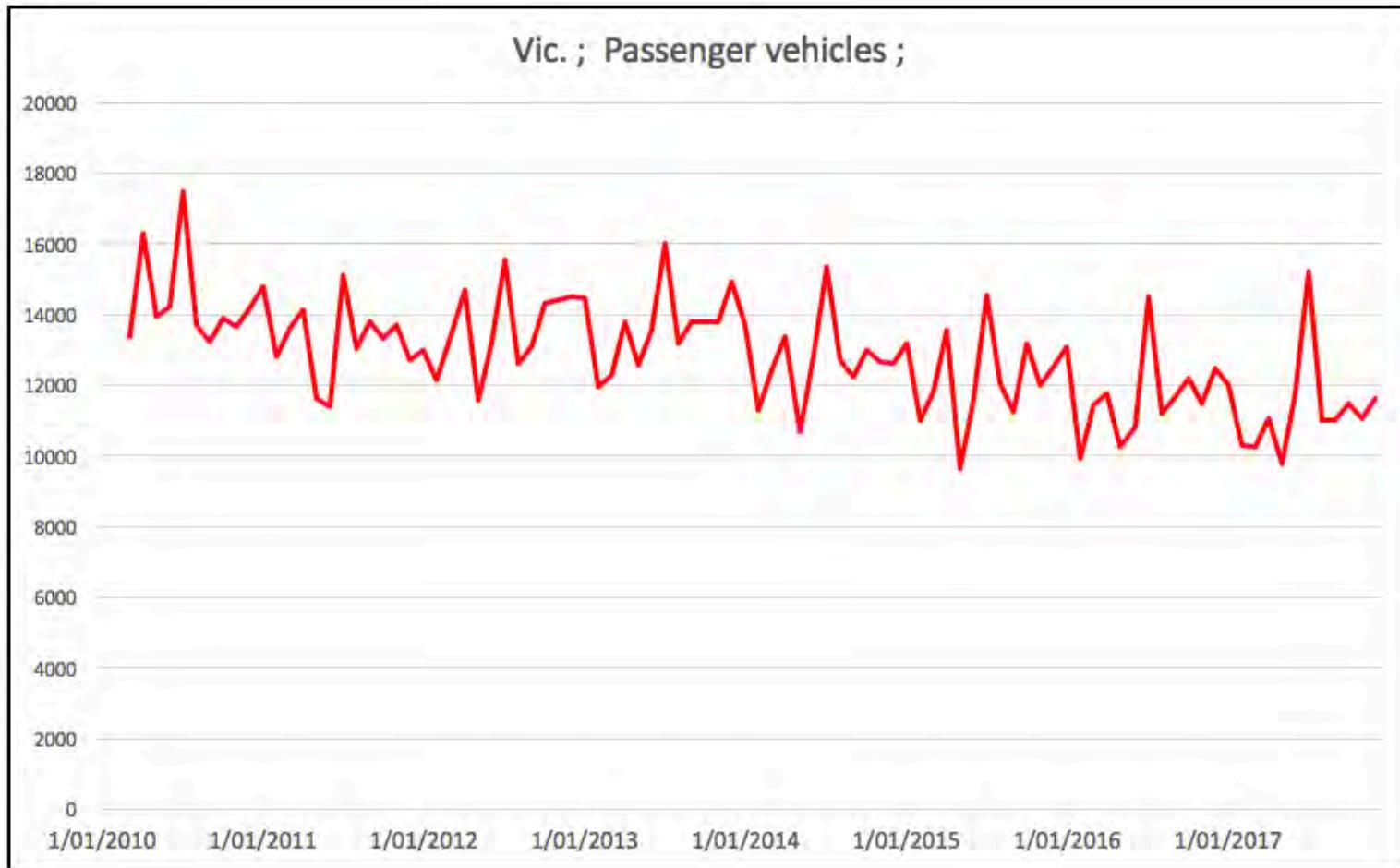


Source: <https://www.marketindex.com.au/all-ordinaries>

SUV sales, Monthly



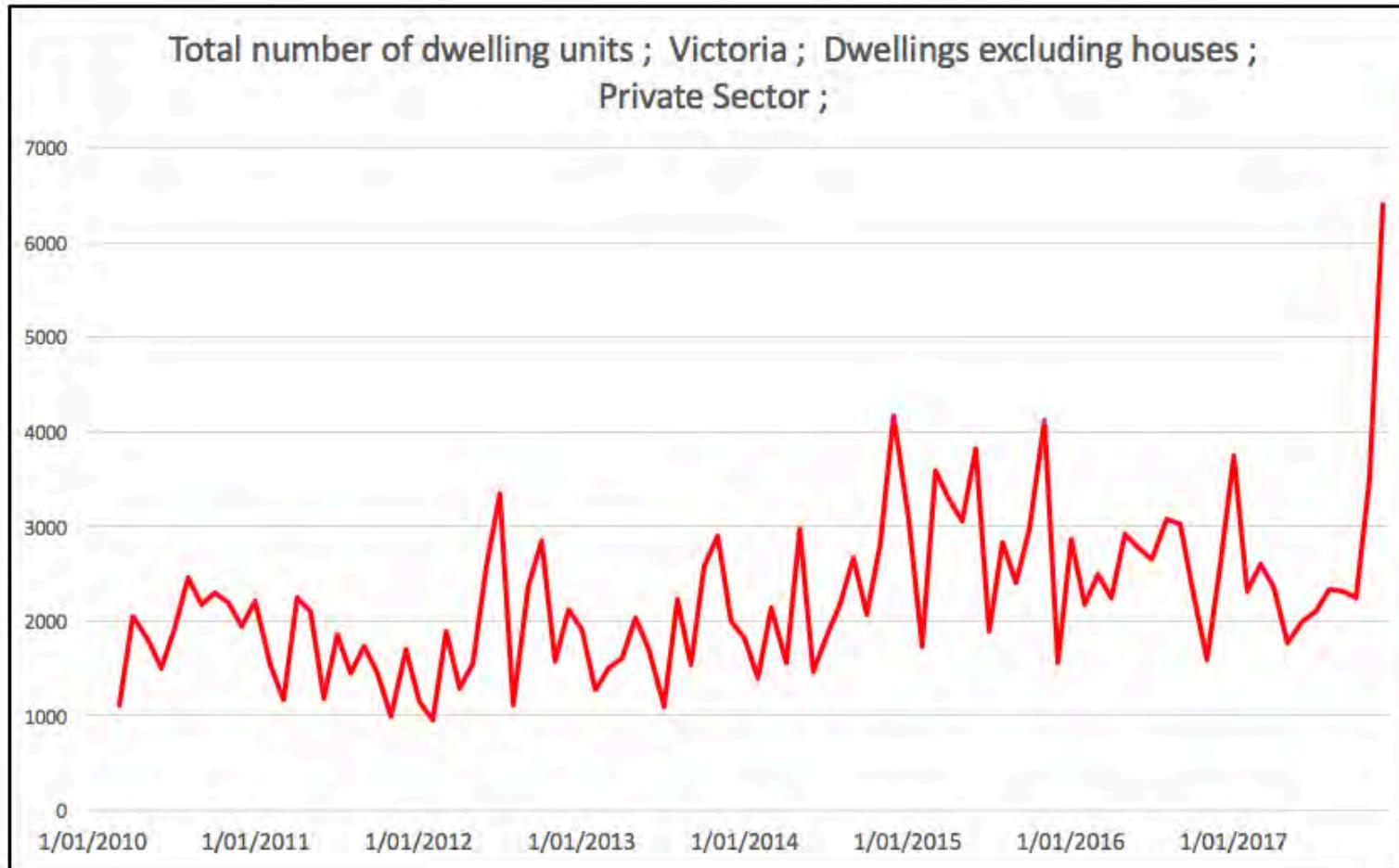
Passenger vehicle sales, Monthly



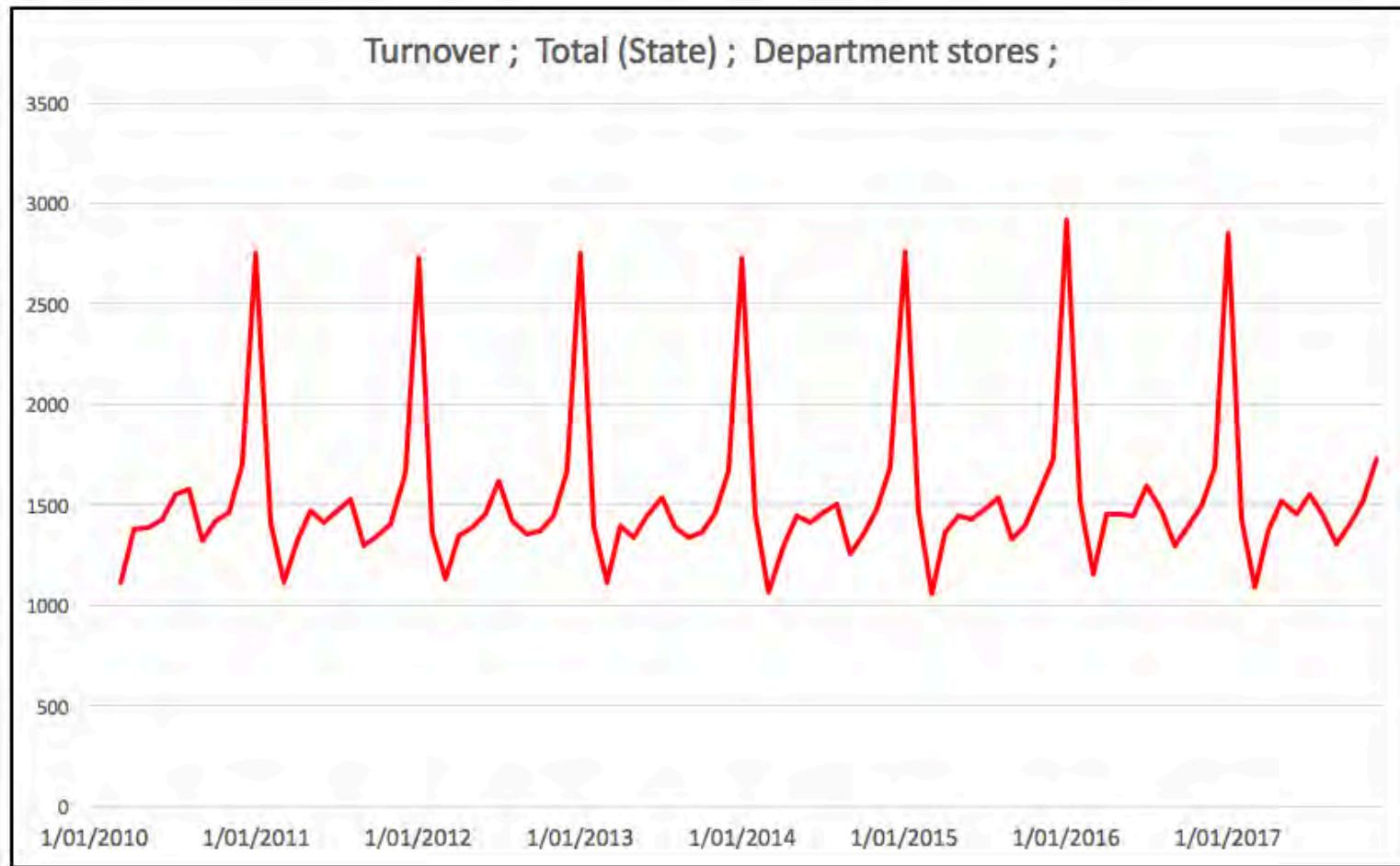
New houses commenced



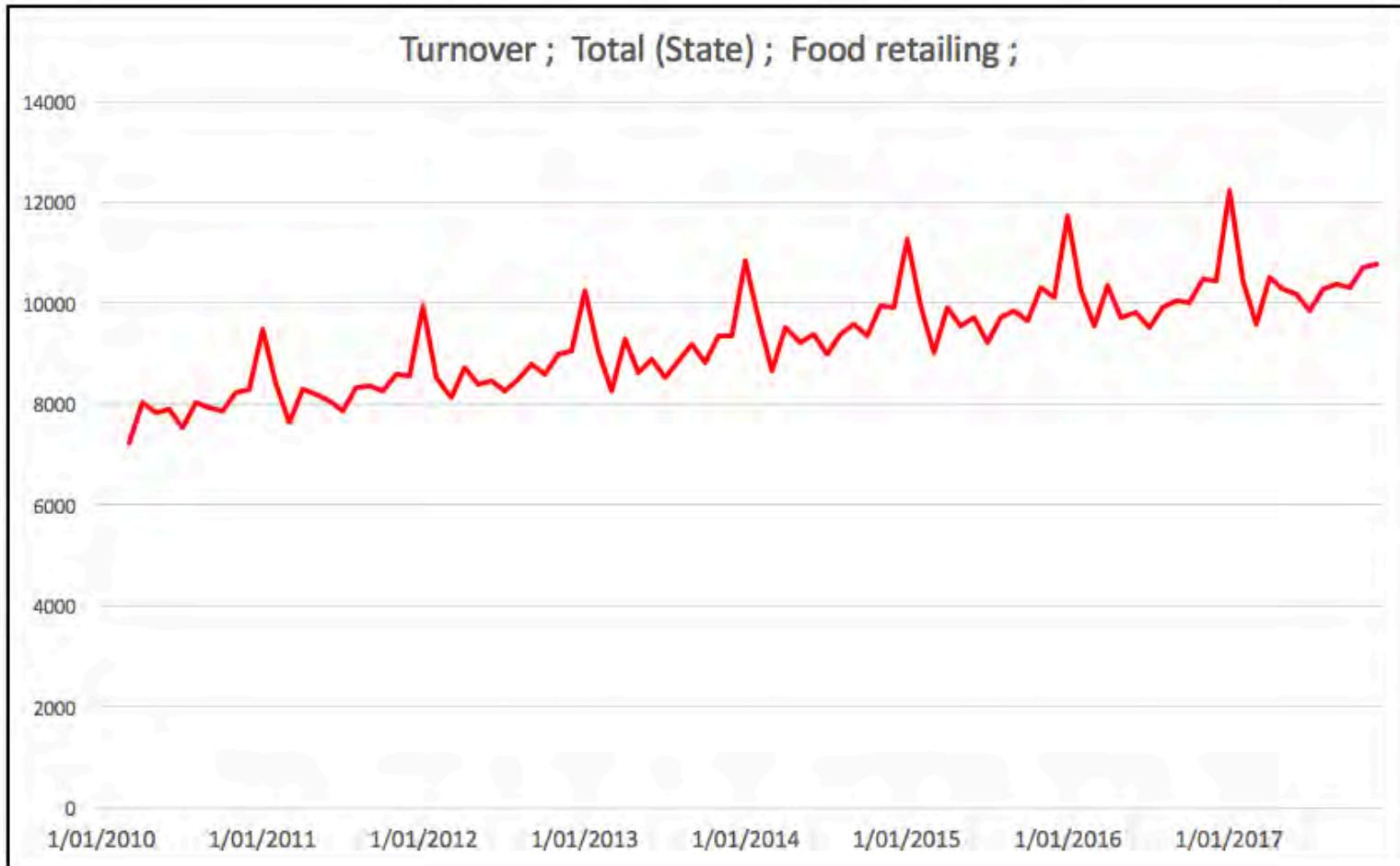
Other dwellings commenced



Retail turnover, department stores



Food retailing, monthly



TSA vs forecasting

A Time Series describes a set of observations made over a period of time. Daily maximum temperatures, hourly share prices, annual population counts, weekly sales figures are all examples of time series.

It is usual, but not strictly necessary, that the observations are recorded at equal intervals.

Some examples of time series follow:

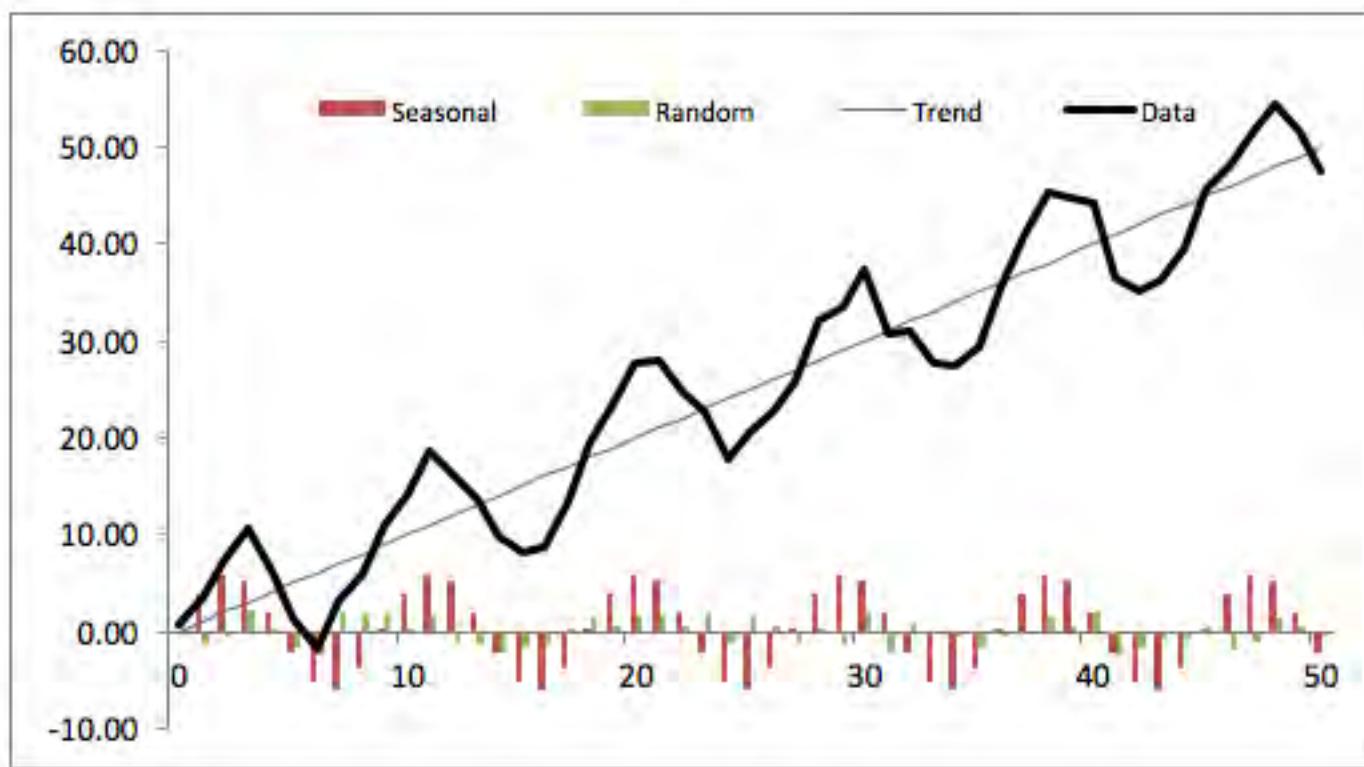
Components of a time series

Time series can be thought of as being composed of three (or four) elements:

- *Trend*, (absence of trend =‘stationary’),
- *Seasonal* and/or *cyclic* element, and a
- *Random* component.

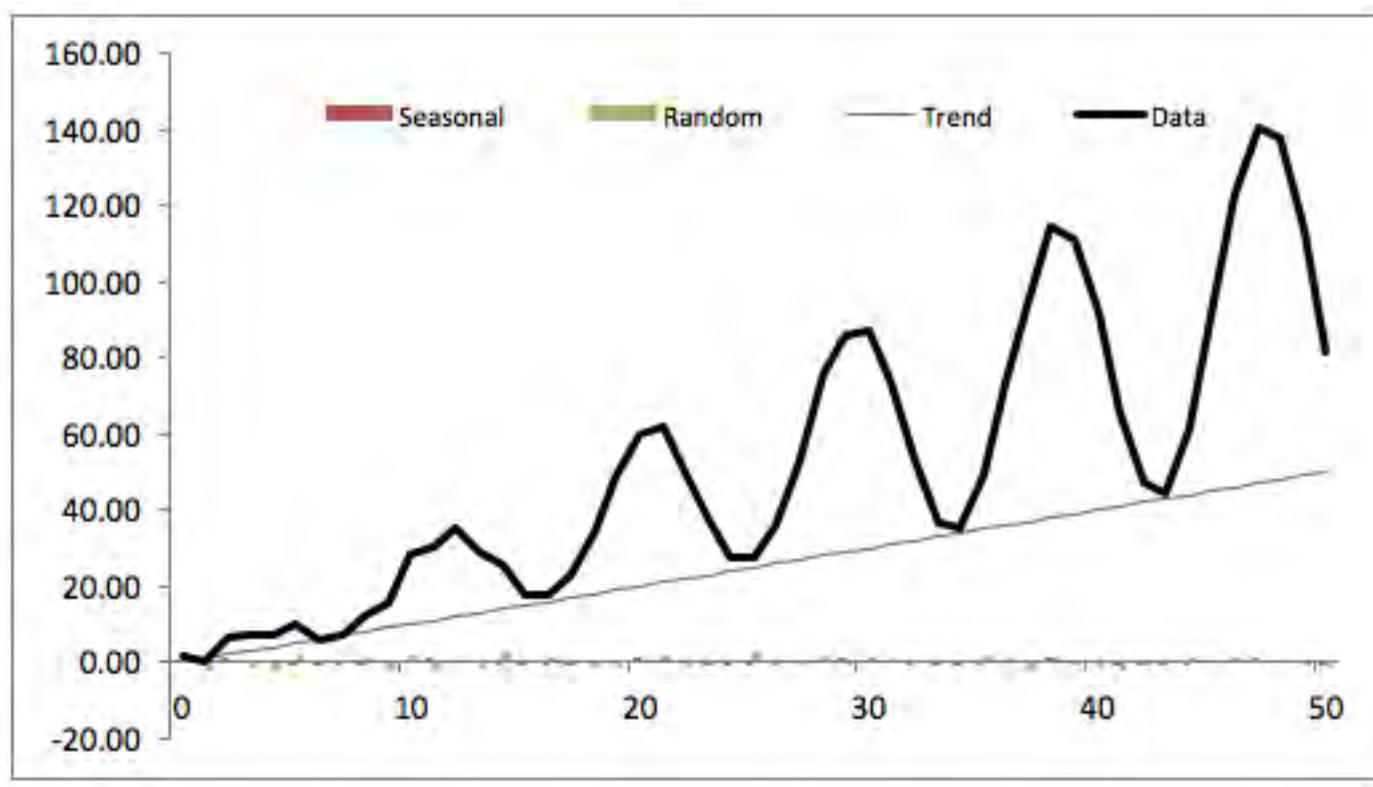
Additive model

$Data = Trend + Seasonal Variation + Random Variation$



Multiplicative model

$Data = Trend * Seasonal Variation * Random Variation$



Moving averages

One of the first tasks in the analysis of additive time series is to smooth the data using a moving average.

As the name suggests, a moving average works by successively taking observations over a number of periods and averaging. The average of the time indexes locates the moving average in time.

Odd numbers of data are preferred for MA's because the data remains centered (time index is an integer), 3, 5, 7 being usual lengths. For quarterly data, a centered 4 period average is used. Medians are also used for robust smoothing.

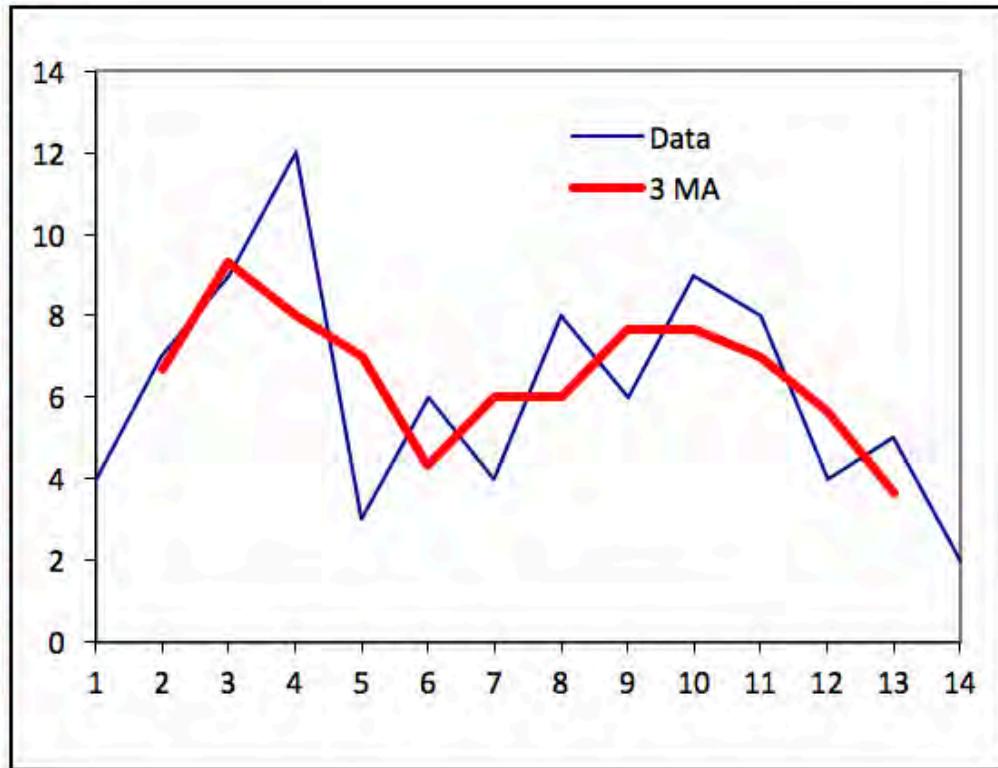
Moving averages

Data *and* Period are averaged to create the moving average.
For a 3 period MA:

Period	Data
1	4
2	7
3	9
4	12
5	3
c	c

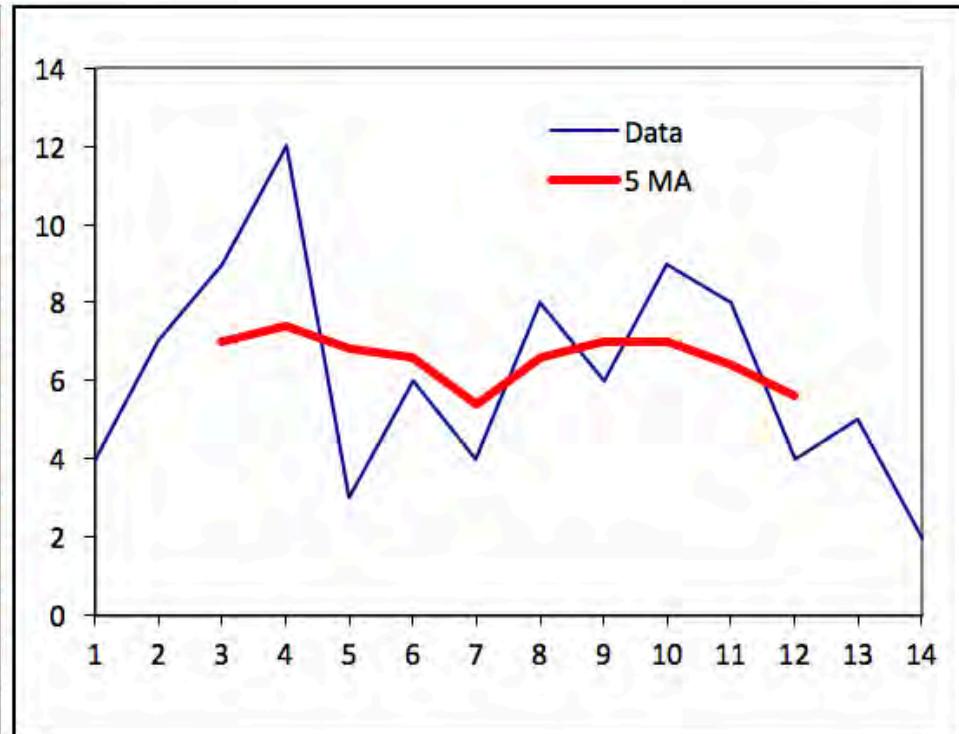
3 period moving average

Period	Data	3 MA	3 MA
1	4		
2	7	$=(4+7+9)/3$	6.67
3	9	$=(7+9+12)/3$	9.33
4	12	$=(9+12+3)/3$	8.00
5	3	$=(12+3+6)/3$	7.00
6	6	...	4.33
7	4	...	6.00
8	8	...	6.00
9	6	...	7.67
10	9	...	7.67
11	8	...	7.00
12	4	...	5.67
13	5	...	3.67
14	2		



5 period moving average

Period	Data	5 MA	5 MA
1	4		
2	7		
3	9 $= (4+7+9+12+3)/5$	7.00	
4	12 $= (7+9+12+3+6)/5$	7.40	
5	3 $= (9+12+3+6+4)/5$	6.80	
6	6	...	6.60
7	4	...	5.40
8	8	...	6.60
9	6	...	7.00
10	9	...	7.00
11	8	...	6.40
12	4	...	5.60
13	5		
14	2		

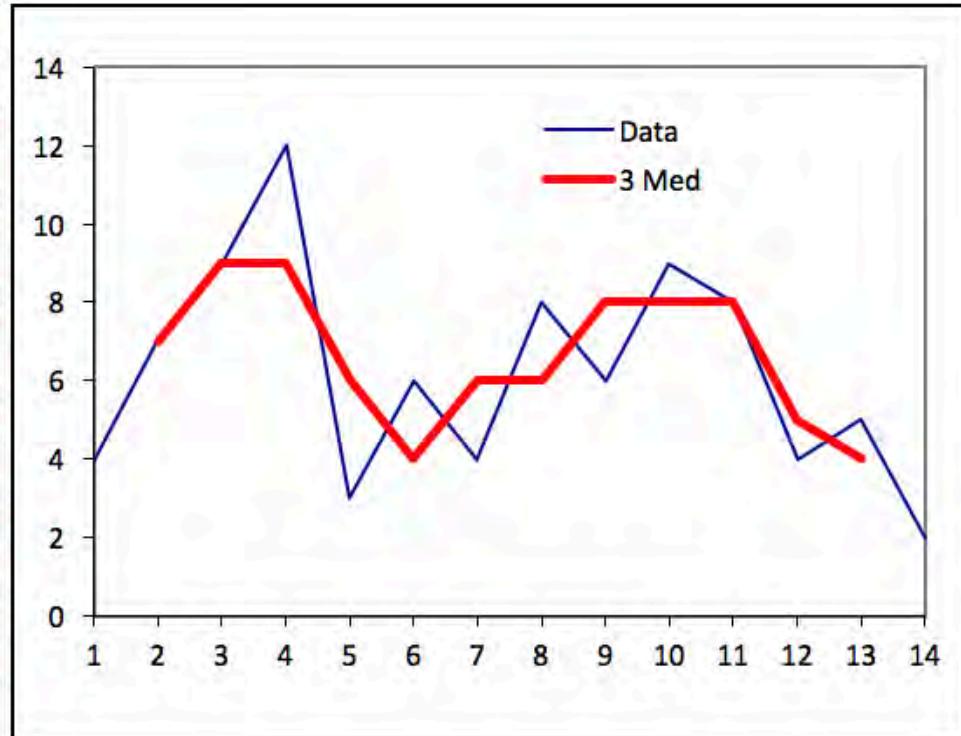


Smoothing with medians

Medians can also be used to smooth data. They are robust to outliers, although not as ‘smooth’ as means.

3 period median smoothing

Period	Data	3 Median	3 Med
1	4		
2	7	=Median(4, 7, 9)	7.00
3	9	=Median(7, 9, 12)	9.00
4	12	=Median(9, 12, 3)	9.00
5	3	...	6.00
6	6	...	4.00
7	4	...	6.00
8	8	...	6.00
9	6	...	8.00
10	9	...	8.00
11	8	...	8.00
12	4	...	5.00
13	5	...	4.00
14	2		



Centered 4 period moving average

For quarterly data, or other data with cycles of 4 periods, a centered 4 period moving average is often used.

The reasoning for this method is as follows:

The moving average contains 4 observations, which comprise a single cycle (Summer Autumn Winter Spring).

For observations in periods 1, 2, 3 and 4, the time index of the average is at period 2.5, i.e., between observations 2 and 3.

We thus take the average of pairs of off-centred observations to re-centre them.

This method can be adapted for other even numbered cycles.

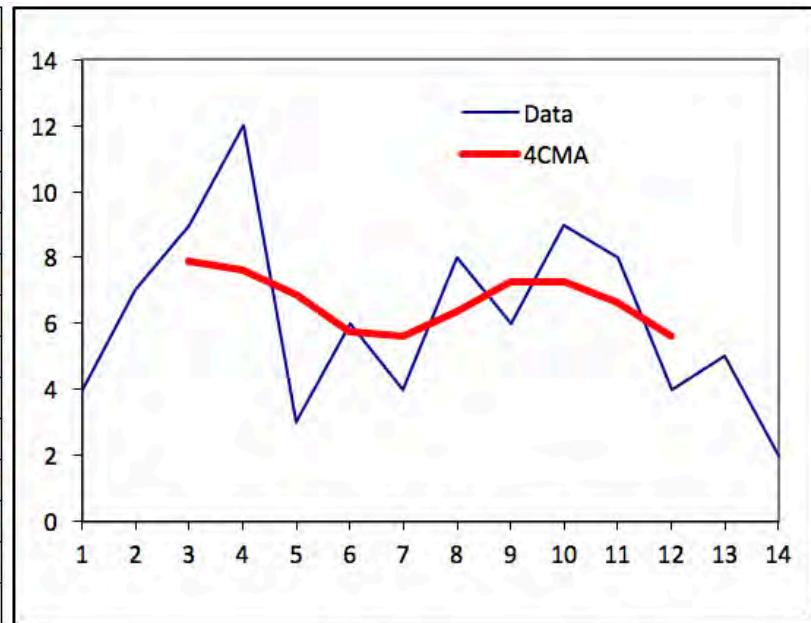
Centered moving averages

For a 4 period Centered MA:

Period	Data
1	4
2	7
3	9
4	12
5	3
c	c

Centered 4 period moving average

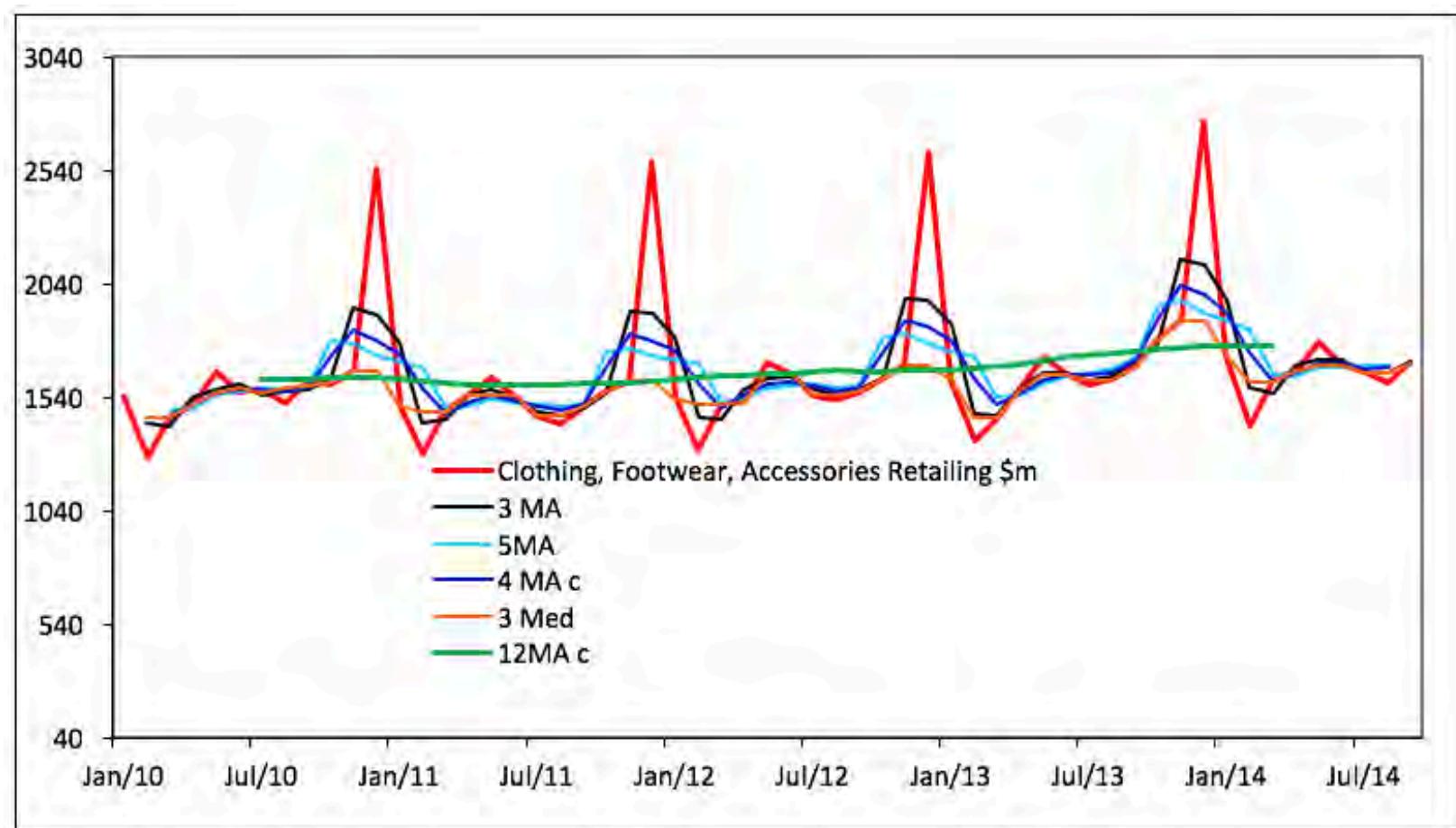
Period	Data	4 MA	4MA	4CMA	4CMA
1	4				
2	7	$(4+7+9+12)/4$	8.00		
3	9	$(7+9+12+3)/4$	7.75	$(8.00+7.75)/2$	7.88
4	12	$(9+12+3+6)/4$	7.50	$(7.75+7.50)/2$	7.63
5	3	$(12+3+6+4)/4$	6.25	$(7.50+6.25)/2$	6.88
6	6	...	5.25	...	5.75
7	4	...	6.00	...	5.63
8	8	...	6.75	...	6.38
9	6	...	7.75	...	7.25
10	9	...	6.75	...	7.25
11	8	...	6.50	...	6.63
12	4	...	4.75	...	5.63
13	5	
14	2				



Methods compared

Month Year	Clothing, Footwear, Accessories Retailing \$m	3 MA	SMA	4 MA c	3 Med	12MA c
1/01/2010	1545.9					
1/02/2010	1273.6	1423.9			1452.3	
1/03/2010	1452.3	1411.1	1486.6	1458.3	1452.3	
1/04/2010	1507.5	1537.8	1489.1	1507.4	1507.5	
1/05/2010	1653.6	1573.2	1550.2	1558.8	1558.6	
1/06/2010	1558.6	1597.1	1561.9	1575.1	1579.0	
1/07/2010	1579	1549.5	1582.5	1570.1	1558.6	1622.2
1/08/2010	1511	1566.7	1570.7	1569.2	1579.0	1621.0
1/09/2010	1610.1	1571.9	1590.2	1583.3	1594.6	1622.8
1/10/2010	1594.6	1620.3	1783.8	1722.5	1610.1	1625.6
1/11/2010	1656.3	1932.6	1781.4	1838.1	1656.3	1626.3
1/12/2010	2546.8	1900.7	1717.7	1786.3	1656.3	1624.6
1/01/2011	1499	1779.2	1694.4	1726.2	1499.0	1618.9
1/02/2011	1291.8	1423.0	1673.1	1579.3	1478.2	1610.1
1/03/2011	1478.2	1440.0	1489.0	1470.6	1478.2	1601.7
1/04/2011	1549.9	1551.4	1498.4	1518.3	1549.9	1596.0
1/05/2011	1626.1	1574.0	1531.3	1547.3	1549.9	1592.9

Methods compared



Analysis of share prices

Some share analysts use simple moving average forecasts to determine when a share is trending up or down.



Source: <https://shareinvesting.anz.com/>

Exponential smoothing

Exponential smoothing is a way of forecasting one, two, three ... periods ahead, using historical data.

This method of smoothing is an adaptive technique: the forecast for the next period is based on the actual value of the data in the current observation less a proportion of the error made in the current forecast...

Thus, this method uses the most recent information to correct (update) the forecast.

Exponential smoothing

New forecast = previous forecast + α (previous actual - previous forecast), where α is between 0 and 1.

$$\hat{y}_{t+1} = \hat{y}_t + a(y_t - \hat{y}_t)$$

$$\hat{y}_{t+1} = \hat{y}_t + a(error)$$

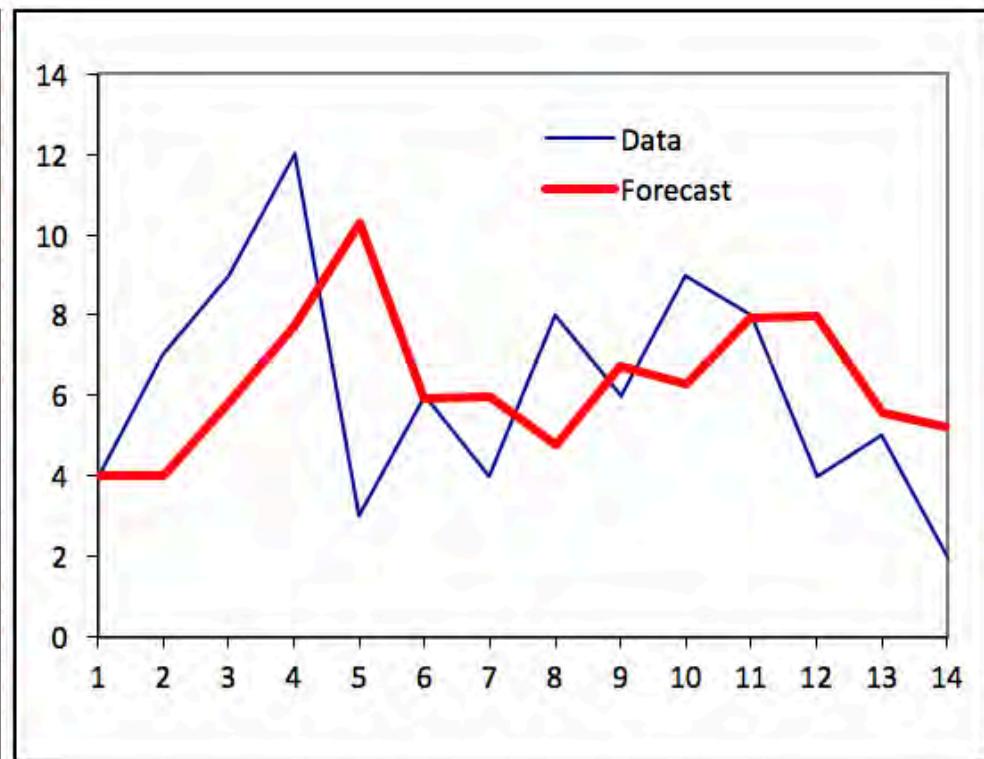
\hat{y}_{t+1} = forecast for next period

\hat{y}_t = forecast at current period

y_t = actual value at current period

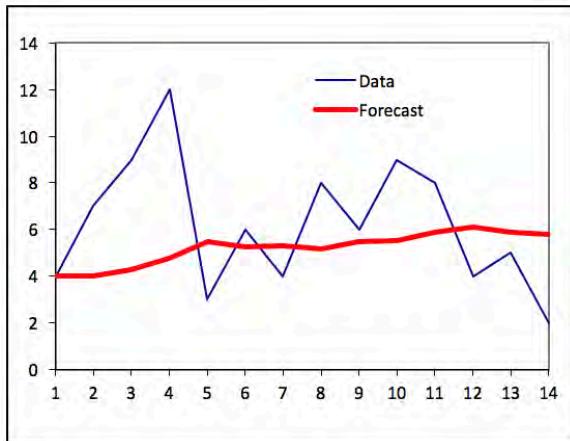
Alpha = 0.6

Period	Data	Forecast	Error
1	4	4	0.00
2	7	4.00	3.00
3	9	5.80	3.20
4	12	7.72	4.28
5	3	10.29	-7.29
6	6	5.92	0.08
7	4	5.97	-1.97
8	8	4.79	3.21
9	6	6.71	-0.71
10	9	6.29	2.71
11	8	7.91	0.09
12	4	7.97	-3.97
13	5	5.59	-0.59
14	2	5.23	-3.23

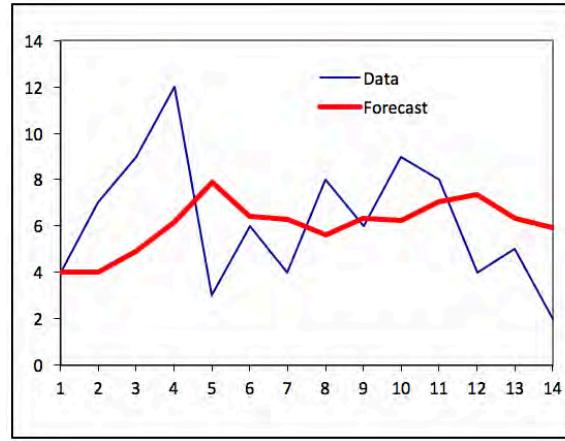


Changing the value of α

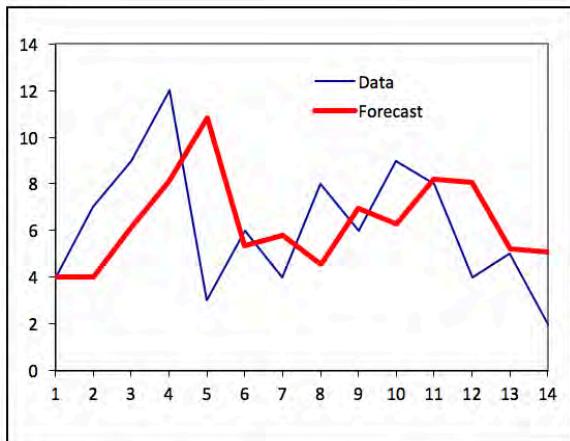
$\alpha = 0.1$



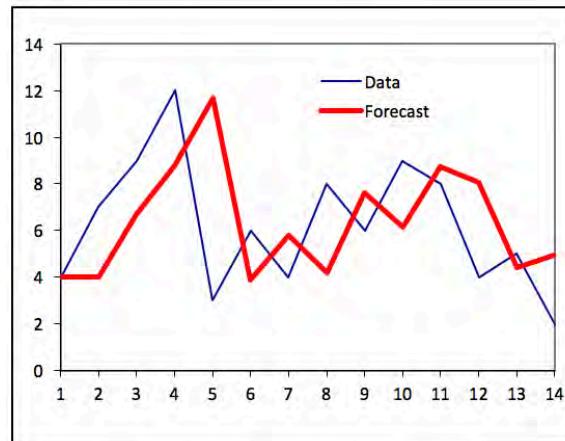
$\alpha = 0.3$



$\alpha = 0.7$



$\alpha = 0.9$



Forecast Accuracy

One measure of forecast accuracy is Mean Absolute Percent Error (MAPE), evaluated over a series of forecasts.

$$MAPE = \frac{1}{n} \times \sum_{i=1}^n \frac{|\hat{Y}_i - y_i|}{y_i}$$

\hat{y}_i = forecast at period i

y_i = actual value period i

n = number of terms evaluated

MAPE example

Alpha = 0.7

Period	Data	Forecast	Error	APE
1	4	4	0.00	
2	7	4.00	3.00	0.43
3	9	6.10	2.90	0.32
4	12	8.13	3.87	0.32
5	3	10.84	-7.84	2.61
6	6	5.35	0.65	0.11
7	4	5.81	-1.81	0.45
8	8	4.54	3.46	0.43
9	6	6.96	-0.96	0.16
10	9	6.29	2.71	0.30
11	8	8.19	-0.19	0.02
12	4	8.06	-4.06	1.01
13	5	5.22	-0.22	0.04
14	2	5.07	-3.07	1.53

First term is not a true forecast for exponential smoothing so don't use.

$$APE = \frac{|\hat{Y}_i - y_i|}{y_i}$$

MAPE 0.60

$$MAPE = \frac{1}{n} \times \sum_{i=1}^n \frac{|\hat{Y}_i - y_i|}{y_i}$$

Summary

You should be able to:

Plot a time series graph.

Recognise the components of a time series:

Trend;

Seasonal or cyclic component;

Random fluctuations (or noise).

Construct a moving average.

Make a one period forecast using exponential smoothing.

Know the effect of different values of α .

Calculate the accuracy of a forecast using MAPE.

Reading/Questions (Selvanathan)

Reading: Time Series

7th Ed. Sections 17.1, 17.2, 17.7.

FIT1006 Lecture 22

Time Series Analysis and Forecasting cont.

Lecture 21/22 motivating problem

Given the value of building work (quarterly) from Mar 2008 – Dec 2015 create a forecast for 2016 & 2017.

Data source: ABS

<http://www.abs.gov.au>

Season/Year	Value of Building Work (all sectors) \$B
Mar-2008	17.50
Jun-2008	20.24
Sep-2008	21.36
Dec-2008	21.17
Mar-2009	18.00
Jun-2009	18.85
Sep-2009	19.62
Dec-2009	20.71
Mar-2010	19.67
Jun-2010	23.17
Sep-2010	23.64
Dec-2010	22.90
Mar-2011	19.49
Jun-2011	21.16
Sep-2011	21.96
Dec-2011	21.47
Mar-2012	19.67

Motivating problem cont...

If, after building the model, you find out the actual value of building work in 2016 & 2017, calculate the error of the forecast.

Dec-2015	27.01
Mar-2016	25.35
Jun-2016	28.79
Sep-2016	28.47
Dec-2016	29.25
Mar-2017	26.20
Jun-2017	29.22
Sep-2017	30.21
Dec-2017	30.35

Regression-based forecasting

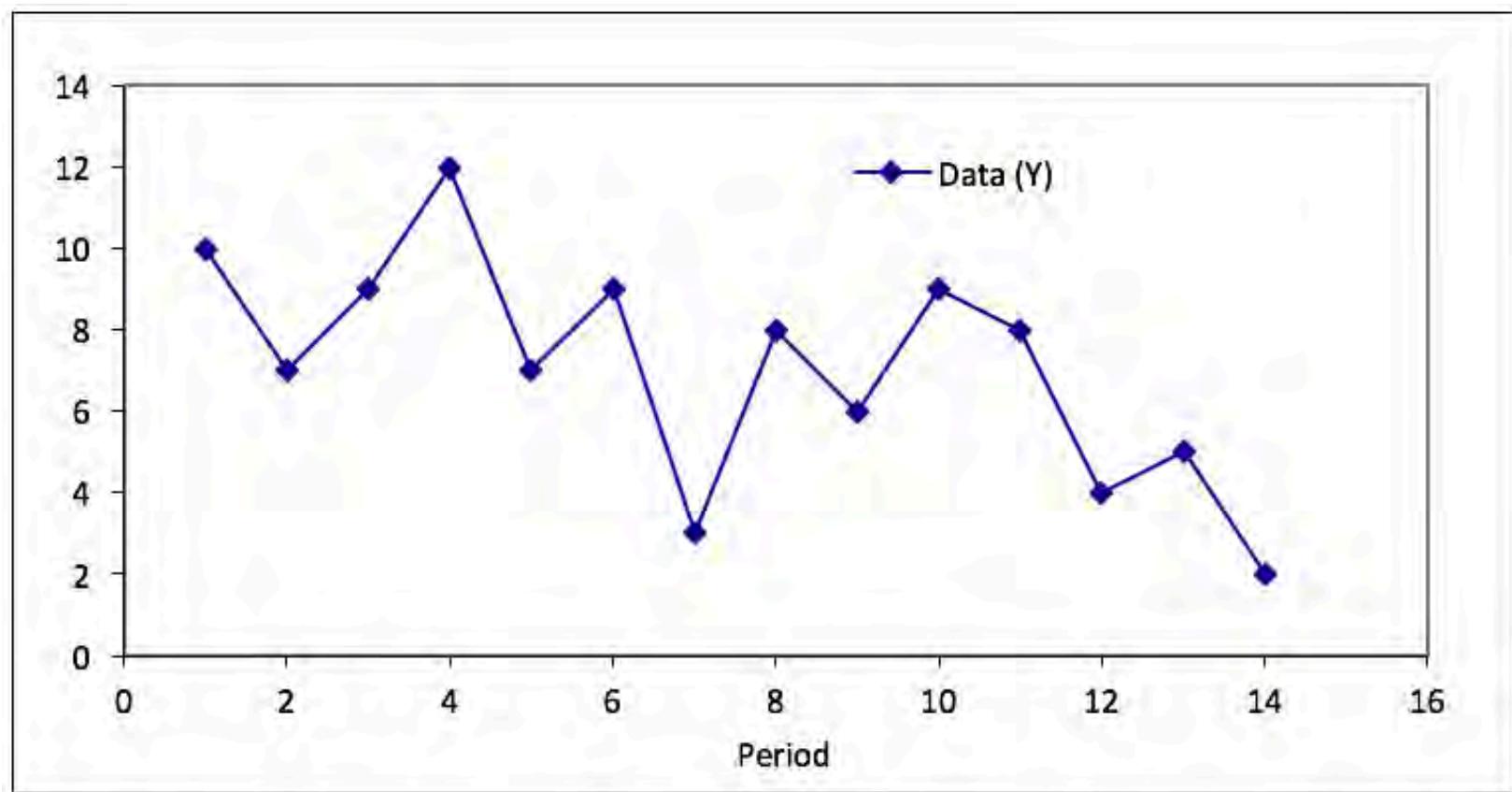
When data has an underlying linear trend, a linear model (equation) using least squares regression can be fitted.

This approach enables a longer term forecast to be made (in contrast to the one or two step forecasts using exponential smoothing).

Simple linear models can be extended to include additive and multiplicative seasonality.

Regression-based forecasting

Model the following by a linear model.



Regression-based forecasting

The first step in the regression is to code the successive observations with an index.

Typically use numbering such as, 1, 2, 3 ..., or 0, 1, 2, ... for this task (assuming equal time intervals).

Eg, for an example time series, we code:

Observation: 10, 7, 9, 12, ...

Period: 1, 2, 3, 4, ...

• • •

Then make the regression calculations.

Period (X)	Data (Y)	XX	YY	XY
1	10	1	100	10
2	7	4	49	14
3	9	9	81	27
4	12	16	144	48
5	7	25	49	35
6	9	36	81	54
7	3	49	9	21
8	8	64	64	64
9	6	81	36	54
10	9	100	81	90
11	8	121	64	88
12	4	144	16	48
13	5	169	25	65
14	2	196	4	28
S	105	99	1015	646

Slope

-0.42

Intercept

10.25

• • •

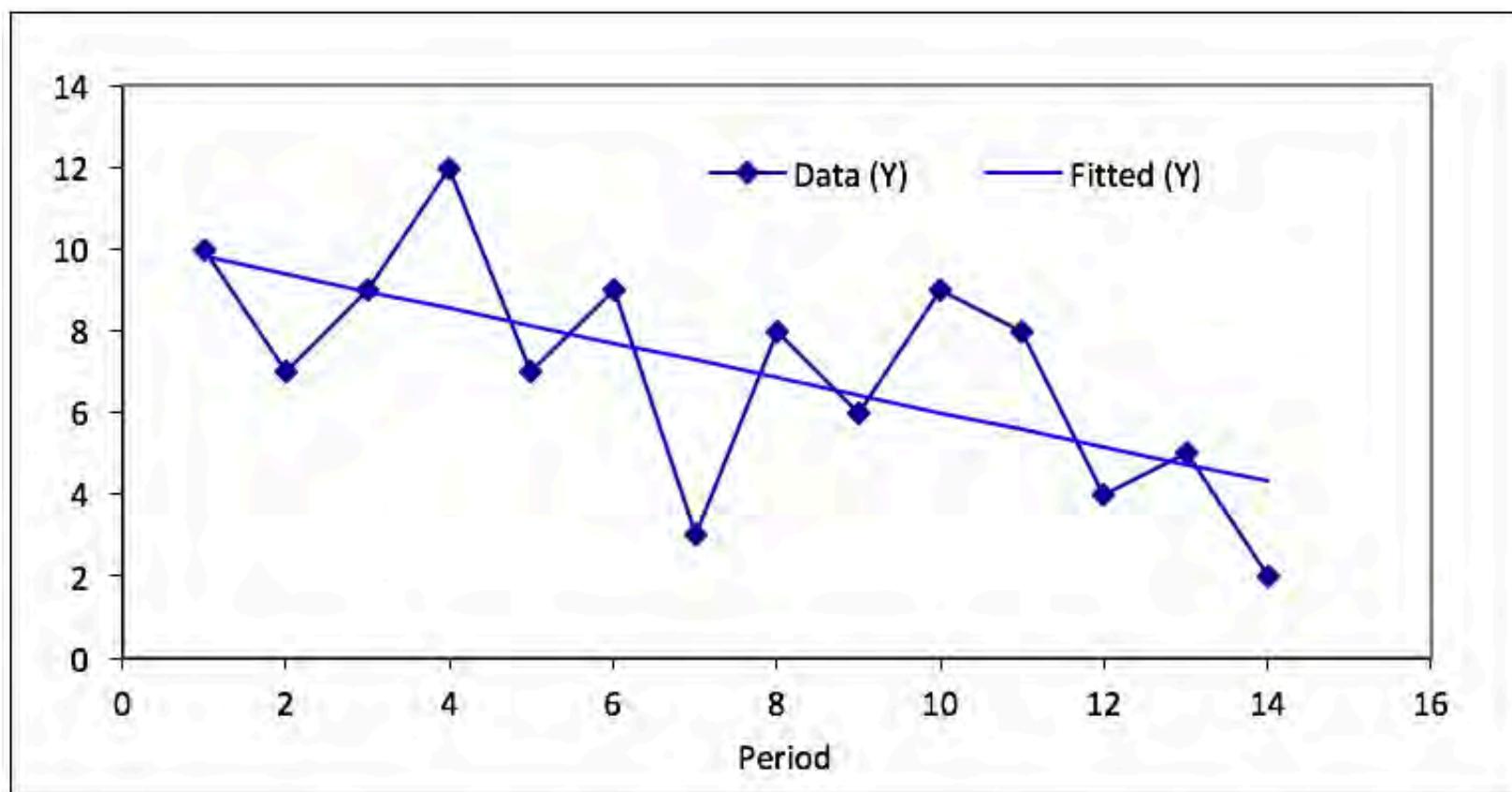
The equation of the line is $Y = 10.25 - 0.42X$.

From which a table of fitted values can be made.

Period (X)	Data (Y)	Fitted (Y)	
1	10	9.83	$25 - 0.42*1$
2	7	9.40	$25 - 0.42*2$
3	9	8.98	$25 - 0.42*3$
4	12	8.56	$25 - 0.42*4$
5	7	8.13	...
6	9	7.71	
7	3	7.28	
8	8	6.86	
9	6	6.44	
10	9	6.01	
11	8	5.59	
12	4	5.16	
13	5	4.74	
14	2	4.31	

• • •

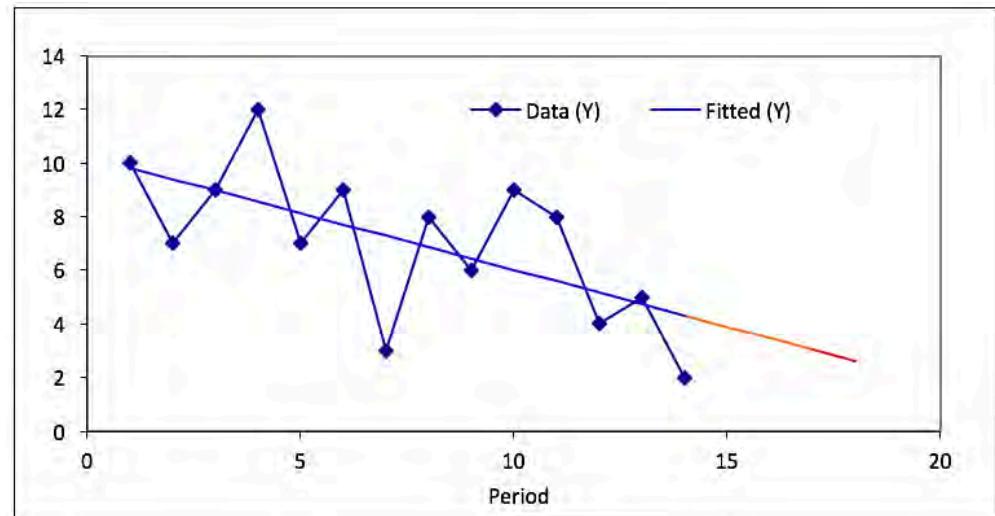
Plotting observed vs fitted values.



Forecasting

To forecast, extend the model beyond the observed data.
The forecast for periods 15, 16, 17 and 18 is:

Period (X)	Data (Y)	Fitted (Y)
3	9	8.98
4	12	8.56
5	7	8.13
...
13	3	4.74
14	8	4.31
15	4	3.89
16	6	3.47
17	5	3.04
18	2	2.62

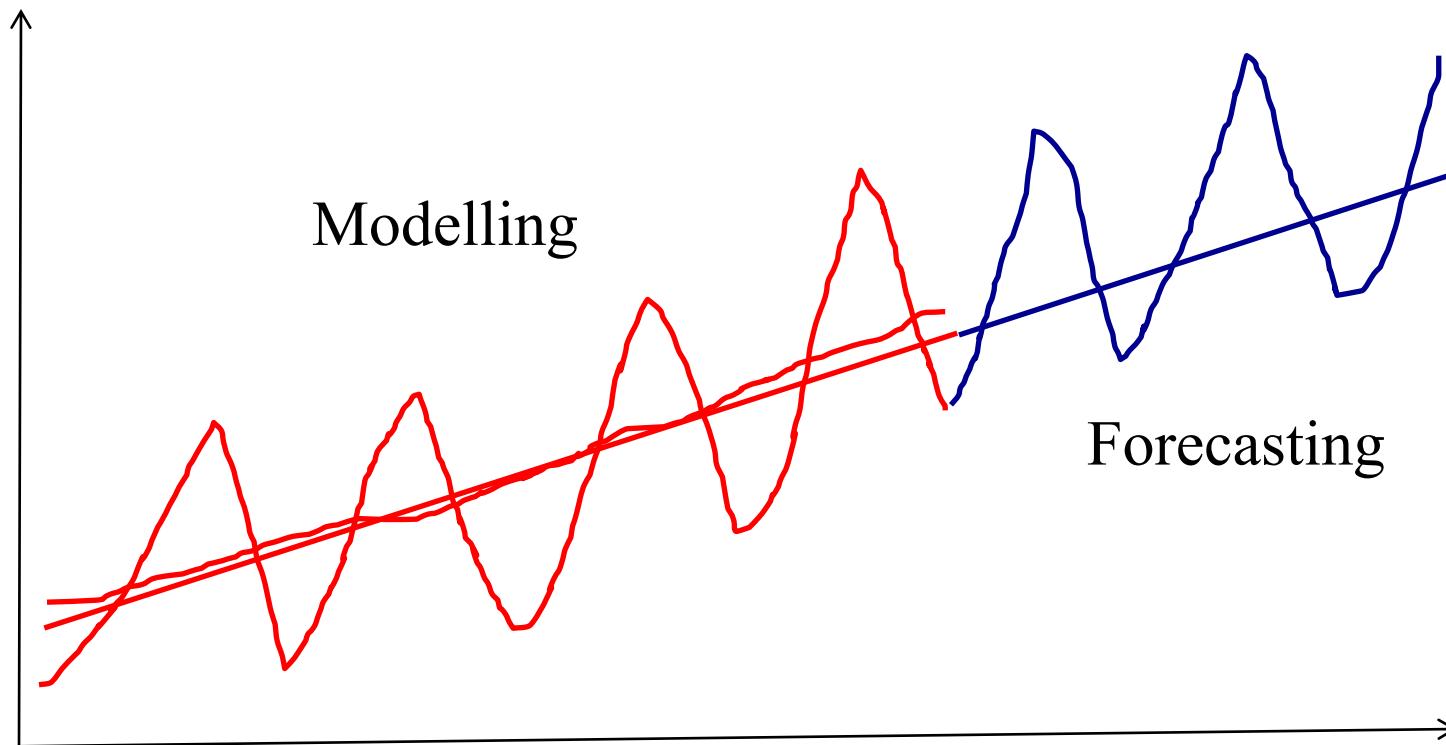


Accuracy of the Forecast

If the observed values at periods 15 to 18 are subsequently found to be 4, 6, 5, 2, then MAPE is:

Period (X)	Data (Y)	Fitted (Y)	APE
3	9	8.98	
4	12	8.56	
5	7	8.13	
...	
13	3	4.74	
14	8	4.31	
15	4	3.89	0.03
16	6	3.47	0.42
17	5	3.04	0.39
18	2	2.62	0.31
		MAPE	0.29

Forecasting: general process



Forecasting Seasonal Data

When forecasting seasonal data we need to observe whether the model follows an additive or multiplicative model.

If the underlying model is additive then multiple regression is the usual approach to modelling the time series. (*not covered in this course*)

If the underlying model is multiplicative, then seasonal indices can be determined and the deseasonalised series can be forecast.

Calculating Seasonal Indices

Multiplicative model:

Ratios to moving average method.

- The time series is smoothed. (Use 4MA C for quarterly data).
- Divide each observation by its corresponding moving average.
- Calculate the average ratio for each season.
- Normalise ratios (to have an average of 1)
- Method can be adapted for periods of any length.

Example

Calculate the seasonal indices for the following data:

Quarter	Sales	Centred 4 Period MA	Ratio Obs/MA
1	362		
2	385		
3	432	382.50	1.13
4	341	388.00	0.88
1	382	399.25	0.96
2	409	413.25	0.99
3	498	430.38	1.16
4	387	454.75	0.85
1	473	478.25	0.99
2	513	499.63	1.03
3	582	519.38	1.12
4	474	536.88	0.88
1	544	557.88	0.98
2	582	580.63	1.00
3	681	601.50	
4	557	627.63	

The observed value is 113% of what the trend predicts it to be.
ie $432/382.5 = 1.13$.

Quarter	1	2	3	4		
			1.13	0.88		
	0.96	0.99	1.16	0.85		
	0.99	1.03	1.12	0.88		
	0.98	1.00				
Average	0.97	1.01	1.14	0.87	sum =	3.99
Indices	0.98	1.01	1.14	0.87	sum =	4

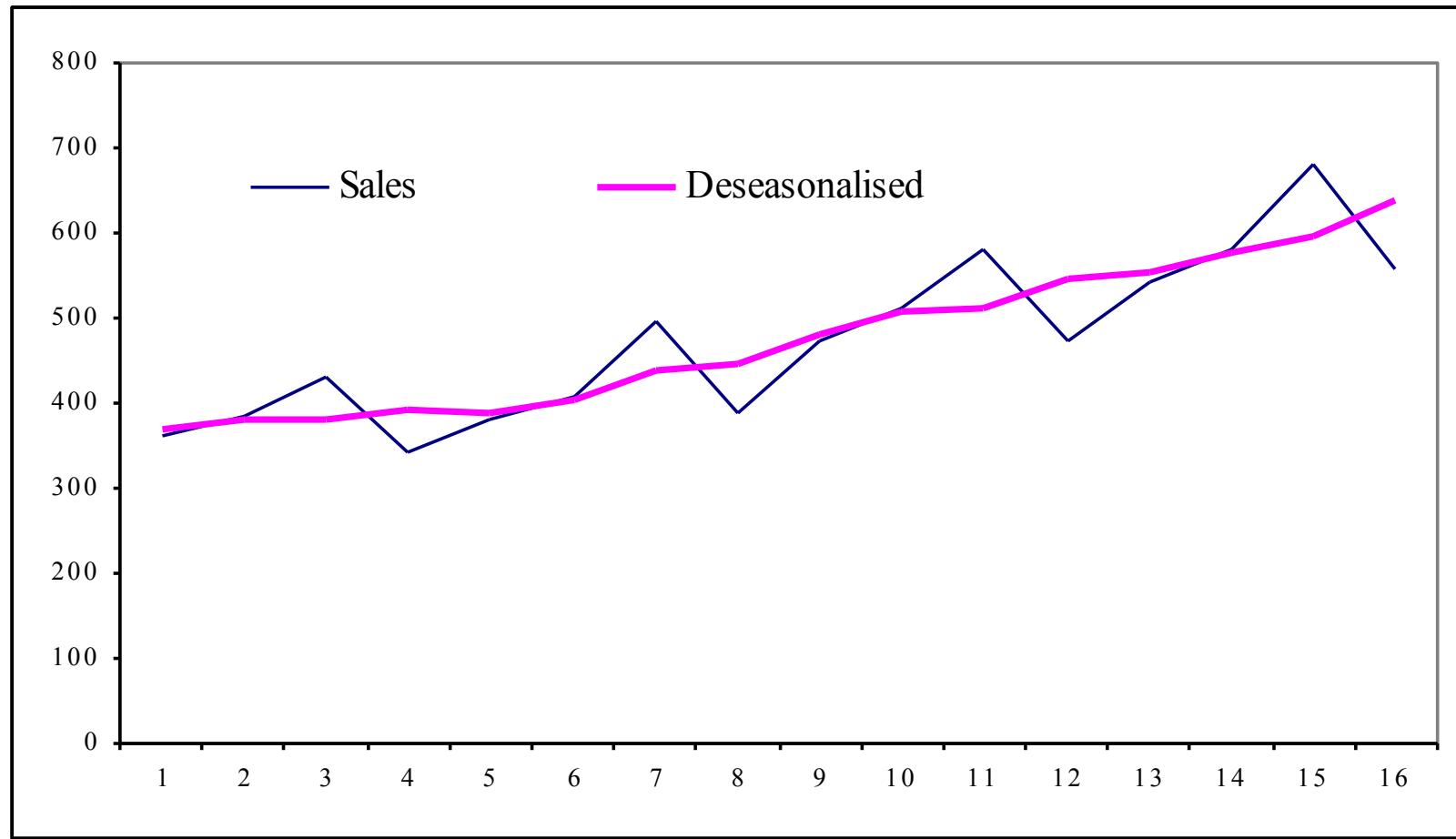
Each average is multiplied by $4/3.99$ to get calculate the index.

De-seasonalising Data

De-seasonalise the time series by dividing each observation by its seasonal factor.

Period	Quarter	Sales	Index	Deseasonalised	Trend and Error
1	1	362	0.98	369.39	
2	2	385	1.01	381.19	
3	3	432	1.14	378.95	
4	4	341	0.87	391.95	
5	1	382	0.98	389.80	
6	2	409	1.01	404.95	←
7	3	498	1.14	436.84	
8	4	387	0.87	444.83	
9	1	473	0.98	482.65	
10	2	513	1.01	507.92	
11	3	582	1.14	510.53	
12	4	474	0.87	544.83	
13	1	544	0.98	555.10	
14	2	582	1.01	576.24	
15	3	681	1.14	597.37	
16	4	557	0.87	640.23	

De-Seasonalised Time Series



Seasonal Forecasting

Having de-seasonalised the data, we can fit a least squares line of best fit, this will create a non-seasonal forecast or trend equation.

We can use this equation to create a trend for future periods.

We then re-seasonalise the trend by multiplying by the seasonal indices.

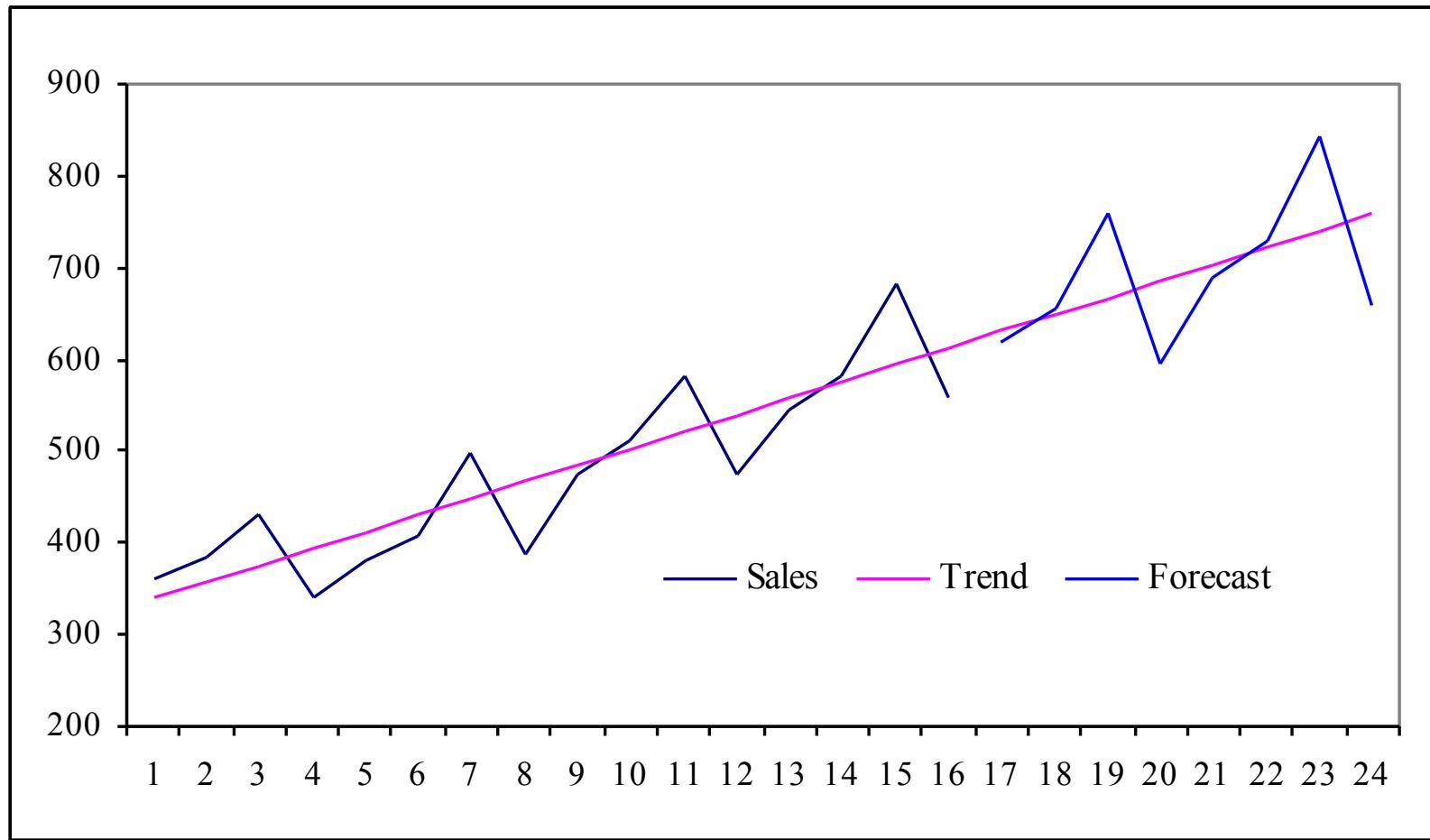
The next slide shows all steps.

Period	Quarter	Sales	Index	Deseasonalised	Trend	Forecast	Slope	18.20
							Intercept	321.10
1	1	362	0.98	369.39	339.30			
2	2	385	1.01	381.19	357.50			
3	3	432	1.14	378.95	375.70			
4	4	341	0.87	391.95	393.90			
5	1	382	0.98	389.80	412.10			
6	2	409	1.01	404.95	430.30			
7	3	498	1.14	436.84	448.50			
8	4	387	0.87	444.83	466.70			
9	1	473	0.98	482.65	484.90			
10	2	513	1.01	507.92	503.10			
11	3	582	1.14	510.53	521.30			
12	4	474	0.87	544.83	539.50			
13	1	544	0.98	555.10	557.70			
14	2	582	1.01	576.24	575.89			
15	3	681	1.14	597.37	594.09			
16	4	557	0.87	640.23	612.29			
17	1		0.98		630.49	617.88		
18	2		1.01		648.69	655.18		
19	3		1.14		666.89	760.26		
20	4		0.87		685.09	596.03		
21	1		0.98		703.29	689.23		
22	2		1.01		721.49	728.71		
23	3		1.14		739.69	843.25		
24	4		0.87		757.89	659.36		



↑
Forecast
Values
↓

Plot of Data, Trend and Forecast



Summary

You should be able to:

Calculate the least squares regression model for a linear time series.

De-seasonalise data using the ratio to moving average method.

Make a de-seasonalised and seasonal forecast using regression.

Calculate the accuracy of your forecast using MAPE.

Lecture 21/22 motivating problem

Given the value of building work (quarterly) from Mar 2008 – Dec 2015 create a forecast for 2016 & 2017.

Data source: ABS

<http://www.abs.gov.au>

Season/Year	Value of Building Work (all sectors) \$B
Mar-2008	17.50
Jun-2008	20.24
Sep-2008	21.36
Dec-2008	21.17
Mar-2009	18.00
Jun-2009	18.85
Sep-2009	19.62
Dec-2009	20.71
Mar-2010	19.67
Jun-2010	23.17
Sep-2010	23.64
Dec-2010	22.90
Mar-2011	19.49
Jun-2011	21.16
Sep-2011	21.96
Dec-2011	21.47
Mar-2012	19.67

Motivating problem cont...

If, after building the model, you find out the actual value of building work in 2016 & 2017, calculate the error of the forecast.

Dec-2015	27.01
Mar-2016	25.35
Jun-2016	28.79
Sep-2016	28.47
Dec-2016	29.25
Mar-2017	26.20
Jun-2017	29.22
Sep-2017	30.21
Dec-2017	30.35

Reading/Questions (Selvanathan)

Reading: Time Series

7th Ed. Sections 17.3, 17.5, 17.6, 17.8